The New Encyclopædia Britannica

Volume 22

MACROPÆDIA

Knowledge in Depth

FOUNDED 1768
15 TH EDITION

Encyclopædia Britannica, Inc.
Robert P. Gwinn, Chairman, Board of Directors
Peter B. Norton, President
Philip W. Goetz, Editor in Chief

Chicago
Auckland/Geneva/London/Madrid/Manila/Paris
Rome/Seoul/Sydney/Tokyo/Toronto

THE UNIVERSITY OF CHICAGO

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

The *Encyclopædia Britannica* is published with the editorial advice of the faculties of the University of Chicago.

Additional advice is given by committees of members drawn from the faculties of the Australian National University, the universities of British Columbia (Can.), Cambridge (Eng.), Copenhagen (Den.), Edinburgh (Scot.), Florence (Italy), Leiden (Neth.), London (Eng.), Marburg (Ger.), Montreal (Can.), Oxford (Eng.), the Ruhr (Ger.), Sussex (Eng.), Toronto (Can.), Victoria (Can.), and Waterloo (Can.); the Complutensian University of Madrid (Spain); the Max Planck Institute for Biophysical Chemistry (Ger.); the New University of Lisbon (Port.); the School of Higher Studies in Social Sciences (Fr.); Simon Fraser University (Can.); and York University (Can.).

First Edition	1768–1771
Second Edition	1777–1784
Third Edition	1788–1797
Supplement	1801
Fourth Edition	1801–1809
Fifth Edition	1815
Sixth Edition	1820–1823
Supplement	1815–1824
Seventh Edition	1830–1842
Eighth Edition	1852–1860
Ninth Edition	1875–1889
Tenth Edition	1902–1903

Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

Fourteenth Edition
© 1929, 1930, 1932, 1933, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943,
1944, 1945, 1946, 1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954,
1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964,
1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973
By Encyclopædia Britannica, Inc.

Fifteenth Edition
© 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985,
1986, 1987, 1988, 1989, 1990, 1991
By Encyclopædia Britannica, Inc.

Printed in U.S.A.

Library of Congress Catalog Card Number: 89-81675
International Standard Book Number: 0-85229-529-4

CONTENTS

Muḥammad and the Religion of Islām

Islām is a major world religion belonging to the Semitic family; it was promulgated by the Prophet Muḥammad in Arabia in the 7th century AD. The Arabic term islām, literally "surrender," illuminates the fundamental religious idea of Islām—that the believer (called a Muslim, from the active particle of *islām*) accepts "surrender to the will of Allāh (Arabic: God)." Allāh is viewed as the sole God—creator, sustainer, and restorer of the world. The will of Allāh, to which man must submit, is made known through the sacred scriptures, the Qur'ān (Koran), which Allāh revealed to his messenger, Muḥammad. In Islām Muḥammad is considered the last of a series of prophets (including Adam, Noah, Jesus, and others), and his message simultaneously consummates and abrogates the "revelations" attributed to earlier prophets.

Retaining its emphasis on an uncompromisng monotheism and a strict adherence to certain essential religious practices, the religion taught by Muḥammad to a small group of followers spread rapidly through the Middle East to Africa, Europe, the Indian subcontinent, the Malay Peninsula, and China. Although many sectarian movements have arisen within Islām, all Muslims are bound by a common faith and a sense of belonging to a single community.

This article deals with the founding of Islām by Muḥammad, the fundamental beliefs and practices of the religion, and the connection of religion and society in the Islāmic world. The history of the various peoples who embraced Islām is covered in the article ISLĀMIC WORLD.

This article is divided into the following major sections:

THE FOUNDATIONS OF ISLĀM

Muḥammad: the Prophet and his message

LIFE AND WORKS

Muḥammad (in full, Abū al-Qāsim Muḥammad ibn 'Abd Allāh ibn 'Abd al-Muṭṭalib ibn Hāshim) was born in Mecca *c.* 570 after the death of his father, 'Abd Allāh. Muḥammad was at first under the care of his paternal grandfather, 'Abd al-Muṭṭalib. Because the climate of Mecca was considered to be unhealthful, he was given as an infant to a wet nurse from a nomadic tribe and spent some time in the desert. At six he lost his mother, Āminah of the clan of Zuhra, and at eight his grandfather. Though his grandfather had been head of the prestigious Hāshem (Hāshim) clan and was prominent in Mecca politics, he was probably not the leading man in Mecca, as some sources suggest. Muḥammad came under the care of the new head of the clan, his uncle Abū Ṭālib, and is reputed to have accompanied him on trading journeys to Syria. About 595, on such a journey, he was in charge of the merchandise of a rich woman, Khadījah of the clan of Asad, and so impressed her that she offered marriage. She is said to have been about 40, but she bore Muḥammad at least two sons, who died young, and four daughters, of whom the best known was Fāṭimah, the wife of Muḥammad's cousin 'Alī, who is regarded as Muḥammad's divinely ordained successor by the Shī'ah branch of Islām. Until Khadījah's death in 619, Muḥammad took no other wife. The marriage was a turning point in Muḥammad's life. By Arab custom, minors did not inherit, and therefore Muḥammad had no share in the property of his father or grandfather; but by his marriage he obtained sufficient capital to engage in mercantile activity on a scale commensurate with his abilities.

Prophetic call and early religious activity. Muḥammad appears to have been of a reflective turn of mind and is said to have adopted the habit of occasionally spending nights in a hill cave near Mecca. The poverty and misfortunes of his early life doubtless made him aware of tensions in Meccan society. Mecca, inhabited by the tribe of Quraysh (Koreish), to which the Hāshim clan belonged, was a mercantile centre formed around a sanctuary, the Kaʿbah (Kaaba), which assured the safety of those who came to trade at the fairs. In the later 6th century there was extensive trade by camel caravan between the Yemen and the Mediterranean region (Gaza and Damascus), bringing goods from India and Ethiopia to the Mediterranean; and the great merchants of Mecca had obtained monopoly control of this trade. Mecca was thus prosperous, but most of the wealth was in a few hands. Tribal solidarity was breaking up; merchants pursued individual interests and disregarded their traditional duties to the unfortunate. About 610, as he reflected on such matters, Muḥammad had a vision of a majestic being (later identified with the angel Gabriel) and heard a voice saying to him, "You are the Messenger of God." This marked the beginning of his career as messenger (or apostle) of God (*rasūl Allāh*), or Prophet (*nabī*). From this time, at frequent intervals until his death, he received "revelations"—that is, verbal messages that he believed came directly from God. Sometimes these were kept in memory by Muḥammad and his followers, and sometimes they were written down. About 650 they were collected and written in the Qurʾān (or Koran, the sacred scriptures of Islām), in the form that has endured. Muslims believe the Qurʾān is divine revelation, written in the words of God himself.

Muḥammad's vision

Muḥammad is said to have been perturbed after the vision and first revelation but to have been reassured by his wife, Khadījah. In his later experiences of receiving messages there was normally no vision. (Occasionally there were physical concomitants, such as perspiring on a cold day, and these gave rise to the suggestion, now agreed to be unwarranted, that he was an epileptic.) Sometimes he heard a noise like a bell but apparently never a voice. The essence of such an experience was that he found a verbal message in his heart—that is, in his conscious mind. With the help of Khadījah's Christian cousin Waraqah, he came to interpret these messages as in general identical with those sent by God through other prophets or messengers to Jews, Christians, and others and to believe that by the first great vision and by the receipt of the messages he was commissioned to communicate them to his fellow citizens and other Arabs. In addition to proclaiming the messages he received, Muḥammad must have offered explanations and expositions of them in his own words, as is evident in the large body of prophetic traditions that the community has preserved.

Soon he gathered some sympathetic friends who accepted his claim to be a prophet and joined him in common worship and prayers. These culminated in an act of prostration in which they touched the ground with their foreheads in acknowledgment of God's majesty—still a cardinal act in Islāmic worship. In about 613 Muḥammad began preaching publicly, and he and his followers spent their days together in the house of a young man named al-Arqam. It is probable that they sometimes worshipped together in the Kaʿbah, a sanctuary of the Arab pagans.

Pagan religious milieu

The people of Mecca at the time nominally worshipped many gods, but few believed that man was dependent on supernatural powers. The merchants thought most things could be accomplished by wealth and by human planning. Some men regarded Allāh as a "high god" who stood above lesser deities. (Allāh, the Arabic word for God, is used by Christian Arabs as well as by Muslims.) The earliest passages of the Qurʾān revealed to Muḥammad emphasize the goodness and power of God as seen in nature and in the prosperity of the Meccans and call on the latter to be grateful and to worship "the Lord of the Kaʿbah," who is thus identified with God. Gratitude is to be expressed in generosity with one's wealth and avoidance of niggardliness. As a sanction, men are warned that they will appear before God on the Last Day to be judged according to their deeds and assigned to heaven or hell. (The doctrines of the Qurʾān are examined later in this article.)

By proclaiming this message publicly, Muḥammad gained followers—39, it is said—before he entered the house of al-Arqam. The names of 70 followers are known prior to the appearance of opposition to the new religion, and there were probably more. Most were young men under 30 when they joined Muḥammad. They included sons and brothers of the richest men in Mecca, though they might be described as persons excluded from the most lucrative forms of commerce. A handful of Muḥammad's early followers are spoken of as "weak," which merely means that they were not of the tribe of Quraysh and so not effectively protected by any clan. The new religion was eventually called Islām—*i.e.,* "surrender [to the will of God]"—and its adherents were called Muslims—*i.e.,* "those who have surrendered"—though the Qurʾān speaks of them primarily as "the believers."

Opposition at Mecca. Although Muḥammad's preaching was basically religious, there was implicit in it a critique of the conduct and attitudes of the rich merchants of Mecca. Attempts were made to get him to soften his criticism by offering him a fuller share in trade and a marriage alliance with one of the wealthiest families, but he decisively rejected such offers. About 615 more active opposition appeared. Points in the message of the Qurʾān were questioned, such as the assertion that men would be resurrected before the Judgment. Commercial pressure was brought to bear on Muḥammad's supporters, and in some families there was mild persecution of junior members who followed him. It is sometimes suggested that the main reason for opposition was the merchants' fear that the new religion would destroy the recognition of the Kaʿbah as a sanctuary, but this is unlikely. Certainly attacks on idols appeared in the Qurʾān, and Islām began to be characterized by the insistence that "there is no god but God" (Allāh), but no attack was made on the Kaʿbah, and the idols mentioned had their chief shrines elsewhere.

Commercial pressure, boycott, and persecution

A leader of the opposition arose in the person of Abū Jahl, a contemporary of Muḥammad, who probably felt that the latter, despite his claim to be "only a warner" (of Judgment to come), was building a position of authority that might one day make him politically supreme in Mecca, because Arabs deeply respected the kind of wisdom or knowledge that Muḥammad clearly had. In about 616 Abū Jahl organized a boycott of the clan of Hāshim by the chief clans of Mecca, allegedly because the clan continued to protect Muḥammad and did not curb his preaching; but, since few of the clan were Muslims, other questions may have been involved. After three years the boycott lost momentum, perhaps because some of the participants found they were harming their own economic interests.

Both Muḥammad's wife, Khadījah, and his uncle Abū Ṭālib died in about 619, and another uncle, Abū Lahab, succeeded as head of the clan of Hāshim. He was closer to the richest merchants, and at their instigation he withdrew the protection of the clan from Muḥammad. This meant that Muḥammad could easily be attacked and therefore could no longer propagate his religion in Mecca. He left for the neighbouring town of aṭ-Ṭāʾif, but the inhabitants were insufficiently prepared to receive his message, and he failed to find support. Having secured the protection of the head of another clan, he returned to Mecca. In 620 Muḥammad began negotiations with clans in Medina, leading to his emigration, or *hijrah,* there in 622.

It is difficult to assess the nature and extent of the persecution of the Muslims in Mecca. There was little physical violence, and that almost always within the family. Muḥammad suffered from minor annoyances, such as having filth deposited outside his door. The persecution is said to have led to the emigration of some of the Muslims to Ethiopia about 615, but they may have been seeking opportunities for trade or military support for Muḥammad. Some remained until 628, long after Muḥammad was established in Medina. Whatever the nature of the persecution, the Muslims were very bitter about it.

The emigration from Mecca to Medina. In the summer of 621, 12 men from Medina, visiting Mecca for the annual pilgrimage to the Kaʿbah (still a pagan shrine), secretly professed themselves Muslims to Muḥammad and went back to make propaganda for him at Medina. At the pilgrimage in June 622 a representative party of 75 persons from Medina, including two women, not merely professed Islām but also took an oath to defend Muḥammad as they would their own kin. These are known as the two Pledges of al-ʿAqaba. Muḥammad now encouraged his faithful Meccan followers to make their way to Medina in small groups, and about 70 emigrated thus. The Meccans are said to have plotted to kill Muḥammad before he could leave. With his chief lieutenant he slipped away unperceived, used unfrequented paths, and reached Medina safely on September 24, 622. This is the celebrated *hijrah* (Latin Hegira), which may be rendered "emigration," though the basic meaning is the severing of kinship ties. It is the traditional starting point of Islāmic history. The Islāmic Era (AH or Anno Hegirae) begins on the first day of the Arabic year in which the *hijrah* took place—July 16, 622, in the Western calendar.

Medina was different from Mecca. It was an oasis in which date palms flourished and cereals could be grown. Agriculture had been developed by several Jewish clans, who had settled among the original Arabs, and they still had the best lands. Later Arab immigrants belonging to the tribes of al-Aws and al-Khazraj, however, were in a stronger position. The effective units among the Arabs were eight or more clans, but nearly all of these had become involved in serious feuds. Much blood had been shed in a battle in about 618, and peace was not fully restored. In inviting Muḥammad to Medina, many of the Arabs there probably hoped that he would act as an arbiter among the opposing parties. Their contact with the Jews may have prepared them for a messianic religious leader, who would deliver them from oppression and establish a kingdom in which justice prevailed.

The Constitution of Medina

A document has been preserved known as the Constitution of Medina. In its present form it is a combination of at least two earlier documents and is probably later than 627, but its main provisions are almost certainly those originally agreed upon between Muḥammad and the Muslims of Medina. In form the document creates a confederation on traditional Arab lines among nine groups—eight Arab clans and the emigrants from Mecca. Muḥammad is given no special position of authority, except that the preamble speaks of the agreement as made between "Muḥammad the prophet" and the Muslims now resident in Medina, and it is stated that serious disputes are to be referred to him. The Jewish groups had refused to acknowledge Muḥammad as prophet and in the document appear in a secondary character as attached to various Arab clans. For at least five years, Muḥammad had no direct authority over members of other clans, but, in the closing years of his life, the prestige of his military successes gave him almost autocratic power. The revelations he received at Medina frequently contained legal rules for the community of Muslims, but they dealt with political questions only rarely.

The first five years at Medina. The first 18 months at Medina were spent in settling down. Muḥammad was given a piece of land and had a house built, which eventually held apartments grouped around a central courtyard for each of his wives. The Muslims often joined Muḥammad at prayers in his home, which, after his death, became the mosque of Medina. The emigrants (*muhājirūn,* the men from Mecca) were at first guests of brother Muslims in Medina, but Muḥammad cannot have contemplated this situation continuing indefinitely. A few emigrants carried on trade in the local market run by a Jewish clan. Others, with the approval of Muḥammad, set out in normal Arab fashion on razzias (*ghazawāt,* "raids") in the hope of intercepting Meccan caravans passing near Medina on their way to Syria. Muḥammad himself led three such razzias in 623. They all failed, probably because traitors betrayed the Muslim movements to the enemy. At last, in January 624, a small band of men was sent eastward with sealed orders telling them to proceed to Nakhlah, near Mecca, and attack a caravan from Yemen. This they did successfully, and in doing so they violated pagan ideas of sanctity—thereby making the Meccans aware of the seriousness of the threat from Muḥammad.

The Battle of Badr and its consequences

About the same time there was a change in Muḥammad's general policy in important respects. One aspect was the "break with Jews"; instead of making concessions to the Jews in the hope of gaining recognition of his prophethood, he asserted the specifically Arabian character of the Islāmic religion. Hitherto the Muslims had faced Jerusalem in prayer, but a revelation now bade them face Mecca. Perhaps because of this change some Muslims of Medina were readier to support Muḥammad. In March 624 he was able to lead about 315 men on a razzia to attack a wealthy Meccan caravan returning from Syria. The caravan, led by Abū Sufyān, the head of the Umayyah clan, eluded the Muslims by devious routes and forced marches. Abū Jahl, the head of the Makhzūm clan, however, leading a supporting force of perhaps 800 men, wanted to teach Muḥammad a lesson and did not withdraw. On March 15, 624, near a place called Badr, the two forces found themselves in a situation, perhaps contrived by Muḥammad, from which neither could withdraw without disgrace. In the ensuing battle at least 45 Meccans were killed, including Abū Jahl and other leading men, and nearly 70 taken prisoner, while only 14 Muslims died. To Muḥammad this appeared to be a divine vindication of his prophethood, and he and all the Muslims were greatly elated.

In the flush of victory some persons in Medina who had satirized Muḥammad in verse were assassinated, perhaps with his connivance. He also made a minor disturbance an excuse for expelling the Jewish clan, which ran the market. This weakened his most serious opponent there, the "hypocrite" (*munāfiq*), or nominal Muslim, ʿAbd Allāh ibn Ubayy, who was allied with the local Jews. The remaining waverers among the Arabs probably became Muslims about this time. Thus the victory of Badr greatly strengthened Muḥammad. At the same time he was using marriage relationships to bring greater cohesion to the emigrants. Of his daughters, Fāṭimah was married to ʿAlī (later fourth caliph, or leader of the Islāmic community) and Umm Kulthūm to ʿUthmān (third caliph). He himself was already married to ʿĀʾishah, daughter of Abū Bakr (first caliph), and was now espoused also to Ḥafṣah, daughter of ʿUmar (second caliph), whose previous husband was one of the Muslims killed at Badr.

The Battle of Uḥud

In the same year Muḥammad led larger Muslim forces on razzias against hostile nomadic tribes and had some success. Presumably, he realized that the Meccans were bound to try to avenge their defeat. Indeed, Abū Sufyān was energetically mobilizing Meccan power. On March 21, 625, he entered the oasis of Medina with 3,000 men. One of the features of Medina was a large number of small forts that were impregnable to Arab weapons and tactics. Muḥammad would have preferred the Muslims to retire to these; but those whose cereal crops were being laid waste persuaded him to go out to fight. By a night march with 1,000 men, he reached the hill of Uḥud on the further side of the Meccan camp. On the morning of March 23 the Meccan infantry attacked and was repulsed with considerable loss. As the Muslims pursued, the Meccan cavalry launched a flank attack after the archers guarding the Muslim left had abandoned their position. The Muslims were thrown into confusion. Some made for a fort and were cut down, but Muḥammad and the bulk of his force managed to gain the lower slopes of Uḥud, where they were safe from the cavalry. The Meccans, because of their losses, were unable to press home their advantages and without delay set out for home, while Muḥammad the next day made a show of pursuing. The battle produced neither a clear victor nor loser. In Badr and Uḥud together, the Meccans had killed about as many men as they had lost; but they had boasted that they would make the Muslims pay several times over, and they had not shown the degree of superiority appropriate to their leading position in Arabia. Muḥammad, though he had lost above 70 men, realized that this was a military reverse, not a defeat; but the confidence of the Muslims and perhaps his own had been struck a serious blow. If the victory of Badr was a sign of God's support, did Uḥud indicate that he had abandoned the Muslims? Muḥammad's faith soon overcame any momentary doubts, and he was gradually able to restore the confidence of his followers.

The siege of Medina

For two years after Uḥud, both sides prepared for a decisive encounter. In the razzias Muḥammad led or sanctioned, he seems to have aimed at extending his own alliances and at preventing others from joining the Meccans. In at least two cases a small party of Muslims was tricked or ambushed, and most of their lives were lost. And another Jewish clan was expelled from Medina. At length, in April 627 Abū Sufyān led a great confederacy of 10,000 men against Medina. On this occasion Muḥammad had ordered the crops to be harvested and a trench to be dug to defend the main part of the oasis from the Meccan cavalry. For a fortnight the confederates besieged the Muslims. Attempts to cross the trench failed, and fodder for the horses was scarce, while Muḥammad's agents among the attackers fomented potential dissensions. Then, after a night of wind and rain the great army melted away. The Meccans had exerted their utmost might and had failed to dislodge Muḥammad, whose position was now greatly strengthened.

For more than two years now there had been opposition to Muḥammad in Medina, chiefly from ʿAbd Allāh ibn Ubayy and other so-called hypocrites (*munāfiquin*) who had abandoned Muḥammad at Uḥud and who together had fostered disaffection. Shortly before the siege Muḥammad had a showdown with ʿAbd Allāh ibn Ubayy, who had joined in spreading slanders about Muḥammad's wife ʿĀ'ishah. This confrontation revealed that ʿAbd Allāh had little support in Medina, and he became reconciled to Muḥammad. After the siege of Medina, Muḥammad attacked the Jewish clan of Qurayẓah, which had probably been intriguing against him. When they surrendered, the men were all executed and the women and children sold as slaves.

The winning of the Meccans. Muḥammad's farsightedness as a statesman is manifest in the policies he next adopted. He might have proceeded to crush the Meccans, and he indeed put economic pressure on them; but his main aim was to gain their willing adherence to Islām. He had already realized that, insofar as the Arabs became Muslims, it would be necessary to direct outward the energies expended on razzias against one another. There could be no question of Muslims raiding Muslims. It is noteworthy that his largest razzias, apart from the expeditions against the Meccans, were along the route to Syria followed by the Arab armies after his death (see ISLAMIC WORLD). He doubtless realized that the administrative skill of the Meccan merchants would be required for any expansion of his embryonic state.

The Treaty of al-Ḥudaybiyah

In a dream Muḥammad saw himself performing the annual pilgrimage to Mecca, and in March 628 he set out to do so, driving sacrificial animals; but he was disappointed because no more than 1,600 men would accompany him. The Meccans were determined to prevent the Muslims from entering their town, so Muḥammad halted at al-Ḥudaybiyah, on the edge of the sacred territory of Mecca. After some critical days the Meccans made a treaty with Muḥammad. Hostilities were to cease, and the Muslims were to be allowed to make the pilgrimage to Mecca in 629. The orderly withdrawal showed how completely Muḥammad controlled his followers. Partly to reward this orderly conduct, Muḥammad two months later led the same force against the Jewish oasis of Khaybar, north of Medina. After a siege it submitted, but the Jews were allowed to remain on condition of sending half of the date harvest to Medina. Thus throughout 628 and 629 Muḥammad's power was growing, since success led more men to become Muslims, for the religious attraction of Islām was apparently supplemented by material motives.

Meanwhile Mecca was in decline. Several leading men had emigrated to Medina and become Muslims. New leaders had taken over from Abū Sufyān but had accomplished little, although the treaty with Muḥammad had removed his pressure on their caravans. Shortly after the treaty, Muḥammad had married Umm Ḥabībah, a daughter of Abū Sufyān and a widow whose Muslim husband had died in Ethiopia. This led to an understanding with Abū Sufyān, who began to work for the peaceful surrender of Mecca. It was probably when he was in Mecca for the pilgrimage in March 629 that Muḥammad became reconciled with another uncle, al-ʿAbbās, and married his uncle's sister-in-law Maymūnah.

The victorious entrance into Mecca

An attack by Meccan allies in about November 629 upon allies of Muḥammad led to the latter's denunciation of the treaty of al-Ḥudaybiyah. After secret preparations he marched on Mecca in January 630 with 10,000 men. Abū Sufyān and other leading Meccans went out to meet him and formally submitted, and Muḥammad promised a general amnesty. When he entered Mecca there was virtually no resistance. Two Muslims and 28 of the enemy were killed. A score of persons were specifically excluded from the amnesty, but some were later pardoned. Thus Muḥammad, who had left Mecca as a persecuted prophet, not merely entered it again in triumph but also gained the allegiance of most of the Meccans. Though he did not insist on their becoming Muslims, many soon did so.

Muḥammad spent 15 to 20 days in Mecca settling various matters of administration. Idols were destroyed in the Kaʿbah and in some small shrines in the neighbourhood. To relieve the poorest among his followers, he demanded loans from some of the wealthy Meccans. When he marched east to meet a new threat, 2,000 Meccans went with him.

The closing years: the unification of Arabia. Ever since the *hijrah,* Muḥammad had been forming alliances with nomadic tribes. At first these were probably nonaggression pacts, but, when he was strong enough to offer protection, he made it a condition of alliance that the tribe should become Muslim. While in Mecca Muḥammad had word of a large concentration of hostile nomads, and he set out to confront them. A battle took place at Ḥunayn in which part of Muḥammad's army was put to flight, but he himself and some older Muslims stood firm. The enemy was finally routed, and their dependents and possessions were all captured. They were allowed to ransom wives and children, but their livestock was divided as booty.

Muḥammad was now militarily the strongest man in Arabia. Most tribes sent deputations to Medina seeking alliance. It is difficult to say how much of Arabia joined his alliance, for the inner politics of each tribe were complex, and in some cases the deputation might represent only a small section. Muḥammad benefitted from the defeat of the Persian Empire by the Byzantine (Christian) Empire (627–628), for, in the Yemen and in places on the Persian Gulf, minorities that had relied on Persian support against Byzantium now turned to Muḥammad instead.

March to the Syrian border. The greatest of all of Muḥammad's razzias occurred at the end of 630, when he took 30,000 men on a month's journey to the Syrian border. In this campaign he pioneered the invasion of Syria and made agreements that became models for treaty arrangements with captured peoples. Some of the tribes near Syria were Christian and adhered to the Byzantines; chiefly as a result of this, Muḥammad's earlier friendship for the Christians, notably those of Ethiopia, changed to hostility. Before his death, armed opposition to him appeared in one or two parts of Arabia, but the Islāmic state was strong enough to deal with this. Thus he left most of Arabia united and poised for expansion into Syria and Iraq.

Muḥammad personally led the pilgrimage to Mecca, in March 632, in a form according with Islāmic belief. Although he had been in poor health for some time, no arrangement had been made for the succession. Thus his death at Medina in June 632 provoked a major crisis among his followers. The dispute over the leadership of the Muslim community eventually resulted in the most important schism in the history of Islām. (This development is discussed later in this article; see below *Theology and sectarianism.*)

CHARACTER AND ACHIEVEMENTS

Although greatly maligned by medieval European scholars—whose opinions still retain some influence—Muḥammad came to be viewed more objectively in the 19th century. Some of the evidence against him, such as his connivance at assassinations and his approval of the execution of the men of a Jewish clan, are historical matters that cannot be denied.

By his contemporaries, however, Muḥammad was admired for his courage, resoluteness, and impartiality, and for a firmness that was tempered by generosity. He won men's hearts by his personal charm. He was gentle, especially with children. Though he was sometimes silent in thought, for the most part he was engaged in purposeful activity. He walked vigorously and spoke rapidly. He became for later Muslims an exemplar of virtuous character, and stories presented him as realizing the Islāmic ideal of human life.

Muḥammad's chief significance is as founder of a state and of a religion. In his lifetime he created a federation of Arab tribes, which, in less than 20 years after his death, defeated the Byzantine and Persian empires, occupied a vast territory from Libya to Persia, and then developed into the Arab, or Islāmic, Empire. He made the religion of Islām the basis of Arab unity. Islāmic doctrine maintains that God is the founder of the religion, not Muḥammad, but the latter played an obviously important part in fostering the nascent religion. His concern with ultimate questions, his mystical outlook, and his moral seriousness were important adjuncts to the preaching of the Qurʾānic message. (W.M.W./Ed.)

THE LEGACY OF MUḤAMMAD

From the very beginning of Islām, Muḥammad had inculcated a sense of brotherhood and a bond of faith among his followers, both of which helped to develop among them a feeling of close relationship that was accentuated by their experiences of persecution as a nascent community in Mecca. The conspicuous socioeconomic content of Islāmic religious practices cemented this bond of faith. In AD 622, when the Prophet fled to Medina, his preaching was soon accepted, and the community-state of Islām emerged. During this early period, Islām acquired its characteristic ethos as a religion uniting in itself both the spiritual and temporal aspects of life and seeking to regulate not only the individual's relationship to God (through his conscience) but human relationships in a social setting as well. Thus, there is not only an Islāmic religious institution but also an Islāmic law, state, and other institutions governing society. Not until the 20th century were the religious (private) and the secular (public) distinguished by some Muslim thinkers and separated formally, as in Turkey.

This dual religious and social character of Islām, expressing itself in one way as a religious community commissioned by God to bring its own value system to the world through the *jihād* ("holy war" or "holy struggle"), explains the astonishing success of the early generations of Muslims. Within a century after the Prophet's death in AD 632, they had brought a large part of the globe—from Spain across Central Asia to India—under a new Arab Muslim empire.

Relationship to other religions

The period of Islāmic conquests and empire building marks the first phase of the expansion of Islām as a religion. Islām's essential egalitarianism within the community of the faithful and its official discrimination against the followers of other religions won rapid converts. Jews and Christians were assigned a special status as communities possessing scriptures and called the "people of the Book" (*ahl al-kitāb*) and, therefore, were allowed religious autonomy. They were, however, required to pay a per capita tax called *jizyah*, as opposed to pagans, who were required to either accept Islām or die. The same status of the "people of the Book" was later extended to Zoroastrians and Hindus, but many "people of the Book" joined Islām in order to escape the disability of the *jizyah*. A much more massive expansion of Islām after the 12th century was inaugurated by the Ṣūfīs (Muslim mystics), who were mainly responsible for the spread of Islām in India, Central Asia, Turkey, and sub-Saharan Africa (see below).

Besides the *jihād* and Ṣūfī missionary activity, another factor in the spread of Islām was the far-ranging influence of Muslim traders, who not only introduced Islām quite early to the Indian east coast and South India but who proved as well to be the main catalytic agents (besides the Ṣūfīs) in converting people to Islām in Indonesia, Malaya, and China. Islām was introduced to Indonesia in the 14th century, hardly having time to consolidate itself there politically before coming under Dutch colonial domination.

The vast variety of races and cultures embraced by Islām (estimated to total from 600,000,000 to 700,000,000 persons worldwide) has produced important internal differences. All segments of Muslim society, however, are bound by a common faith and a sense of belonging to a single community. With the loss of political power during the period of Western colonialism in the 19th and 20th centuries, the concept of the Islāmic community (*ummah*), instead of weakening, became stronger. The faith of Islām helped various Muslim peoples in their struggle to gain political freedom in the mid-20th century and the unity of Islām contributed to later political solidarity.

Sources of Islāmic doctrinal and social views

Islāmic doctrine, law, and thinking in general are based upon four sources, or fundamental principles (*uṣūl*): (1) the Qurʾān, (2) the *sunnah* ("traditions"), (3) *ijmāʿ* ("consensus"), and (4) *ijtihād* ("individual thought").

The Qurʾān (literally, Reading, or Recitation) is regarded as the Word, or Speech, of God delivered to Muḥammad by the angel Gabriel. Divided into 114 *sūrah*s (chapters) of unequal length, it is the fundamental source of Islāmic teaching. The *sūrah*s revealed at Mecca during the earliest part of Muḥammad's career are concerned with ethical and spiritual teachings and the Day of Judgment. The *sūrah*s revealed at Medina at a later period in the career of the Prophet are concerned with social legislation and the politico-moral principles for constituting and ordering the community. *Sunnah* ("a well-trodden path") was used by pre-Islāmic Arabs to denote their tribal or common law; in Islām it came to mean the example of the Prophet; *i.e.,* his words and deeds as recorded in compilations known as Ḥadīth.

Ḥadīth, or collection of the Prophet's sayings

Ḥadīth (a Report, or collection of sayings attributed to the Prophet) provide the written documentation of the Prophet's word and deeds. Six of these collections, compiled in the 3rd century AH (9th century AD) came to be regarded as especially authoritative by the largest group in Islām, the Sunnah. Another large group, the Shīʿah, has its own Ḥadīth.

The doctrine of *ijmāʿ*, or consensus, was introduced in the 2nd century AH (8th century AD) in order to standardize legal theory and practice and to overcome individual and regional differences of opinion. Though conceived as a "consensus of scholars," in actual practice *ijmāʿ* was a more fundamental operative factor. From the 3rd century AH *ijmāʿ* has amounted to a principle of rigidity in thinking; points on which consensus was reached in practice were considered closed and further substantial questioning of them prohibited. Accepted interpretations of the Qurʾān and the actual content of the *sunnah* (*i.e.,* Ḥadīth and theology) all rest finally on the *ijmāʿ*.

Ijtihād, meaning "to endeavour" or "to exert effort," was required to find the legal or doctrinal solution to a new problem. In the early period of Islām, because *ijtihād* took the form of individual opinion (*raʾy*), there was a wealth of conflicting and chaotic opinions. In the 2nd century AH *ijtihād* was replaced by *qiyās* (reasoning by strict analogy), a formal procedure of deduction based on the texts of the Qurʾān and the Ḥadīth. The transformation of *ijmāʿ* into a conservative mechanism and the acceptance of a definitive body of Ḥadīth virtually closed the "gate of *ijtihād*." Nevertheless, certain outstanding Muslim thinkers (*e.g.,* al-Ghazālī, died AD 1111) continued to claim the right of new *ijtihād* for themselves, and reformers of the 18th and 19th centuries, because of modern influences, have caused this principle to once more receive wider acceptance.

The Qurʾān and Ḥadīth are treated in the following sections. The significance of *ijmāʿ* and *ijtihād* are discussed below in the contexts of Islāmic theology, philosophy, and law. (F.R./Ed.)

Islāmic scripture: the Qurʾān

The Qurʾān (Arabic: Reading or Recitation; often spelled Koran), the holy book of Islām, is regarded by believers as the true word of God as revealed to the Prophet Muḥammad. In its written form it is accepted as the earthly reproduction of an uncreated and eternal heavenly original, according to the general view referred to in the Qur’ān itself as "the well-preserved tablet" (*al-lawḥ al-maḥfūẓ*; Qur’ān 75:22). The word *qur’ān* is derived from the verb *qara’a* "to read," "to recite," but there is probably also some connection with Syriac *qeryānā,* "reading," used for the scriptural lessons in the Syrian Church. In the Qur’ān itself the word is not used with reference to the book as a whole but only as a term for separate revelations or for the divine revelation in general. The Qur’ān is held in high esteem as the ultimate authority in all matters legal and religious and is generally regarded as infallible in all respects. Its Arabic language is thought to be unsurpassed in purity and beauty and to represent the highest ideal of style. To imitate the style of the Qur’ān is a sacrilege.

FORM

Division into chapters and verses

In length the Qur’ān is approximately comparable with the New Testament. For purposes of recitation during the holy month of Ramaḍān it is divided into 30 "portions" (*juz’,* plural *ajzā’*), one for each day of the month. Its main division, however, is into 114 chapters, called *sūrah*s, of very unequal length. With the exception of the first *sūrah,* the so-called *fātiḥah* ("opening" of the book), which is a short prayer, the *sūrah*s are arranged roughly according to length, *sūrah* 2 being the longest and the last two or three the shortest. Because the longest *sūrah*s generally derive from the latter part of Muḥammad's activity, the consequence of this arrangement is that the oldest *sūrah*s are generally to be found toward the end of the book and the youngest generally appear at its beginning.

In the accepted version of the Qur’ān now in use, each *sūrah* has a heading containing the following elements: (1) a title, which is usually derived from some conspicuous word in the *sūrah,* such as "The Cow," "The Bee," "The Poets," but is usually not an indication of the contents of the whole chapter; (2) the *basmalah; i.e.,* the formula-prayer "In the name of God, the Merciful, the Compassionate"; (3) an indication of whether the *sūrah* was revealed at Mecca or at Medina and of the number of its verses; and finally (4) in some cases one or more *fawātiḥ,* or detached letters (*e.g., ṭā’ sīn, ṭā’ sīn mīm*), or *alif lām mīm,* the meaning of which has not been satisfactorily explained, though it is thought that they might stand for abbreviated words, indicate certain collections of *sūrah*s, or have an esoteric significance.

The verses in the Qur’ān are called *āyah* (plural *āyāt,* literally "signs") and vary considerably in length. The shortest verses generally occur in the earliest *sūrah*s, in which the style of Muḥammad's revelation comes very close to the rhymed prose (*saj‘*) used by the *kāhin*s, or soothsayers, of his time. As the verses get progressively longer and more circumstantial, the rhymes come farther and farther apart. There is also a change of linguistic style: the earlier *sūrah*s are characterized by short sentences, vivid expressions, and poetic force; and the later ones become more and more detailed, complicated and, at times, rather prosaic in outlook and language. As a result, it is sometimes difficult to decide whether or not a rhyme is intended to indicate the end of a verse; and consequently, there are variations in the numbering of verses (*e.g.,* between the European editions long used by Western scholars and the official Egyptian edition that has now replaced them in most scholarly works).

The Qur’ān generally appears as the speech of God, who mostly speaks in the first person plural ("we"). When the prophet Muḥammad is speaking to his compatriots, his words are introduced by the command, "Say," thus emphasizing that he is also speaking on divine injunction only. At times the form is also dramatic, bringing in objections by Muḥammad's opponents and answering them by counter-arguments. Narrative passages are mostly brief. Stories of prophets and biblical persons are often alluded to as though they are known to the audience. The stress is not on the narrative but on its didactic uses.

On closer analysis very few of the *sūrah*s turn out to be uniform in style or content. The longest text dealing with one subject is *sūrah* 12, which tells the story of Joseph, differing from the biblical account in a great many details, most of which seem to outside historians to have been drawn from Jewish sources. Otherwise the longer *sūrah*s are composed of several brief sections dealing with a variety of topics. Thus the Qur’ān does not give the appearance of a planned, organized, or systematic treatise, an impression that is further heightened by the fact that certain favourite phrases such as "but God is forgiving, compassionate," "God is knowing, wise," "most of them know nothing" often have little or no apparent connection with the immediate context. In fact, some skeptics claim that these additions served only to produce a needed rhyme.

Heterogeneous style

It is often emphasized that Muḥammad brought to his people "an Arabic Qur’ān"; *i.e.,* a book or set of recitations in the Arabs' own language comparable to those of Judaism and Christianity. Also the vocabulary of the Qur’ān is overwhelmingly of Arabic origin, but there are, nevertheless, borrowed words, mostly from Hebrew and Syriac, bearing witness to Muḥammad's debt to Judaism and Christianity. These loan words are primarily technical terms such as *injīl,* "gospel" (Greek *evangelion*); *taurāt,* "the law, or Torah" of Judaism; *Iblīs,* "the Devil" (Greek *diabolos*); or translations or adaptations of theological terms such as *āmana,* "to believe" (Hebrew or Aramaic); *ṣalāt,* "prayer" (probably Syriac). Such explanations are usually regarded with suspicion by Muslims, since orthodox doctrine holds that the language of the Qur’ān is the purest Arabic. (H.R./Ed.)

DOCTRINES OF THE QUR’ĀN

God. The doctrine about God in the Qur’ān is rigorously monotheistic: God is one and unique; he has no partner and no equal. Trinitarianism, the Christian belief that God is three persons in one substance, is vigorously repudiated. Muslims believe that there are no intermediaries between God and the creation that he brought into being by his sheer command: "Be." Although his presence is believed to be everywhere, he does not inhere in anything. He is the sole Creator and sustainer of the universe, wherein every creature bears witness to his unity and lordship. But he is also just and merciful: his justice ensures order in his creation, in which nothing is believed to be out of place, and his mercy is unbounded and encompasses everything. His creating and ordering the universe is viewed as the act of prime mercy for which all things sing his glories. The God of the Qur’ān, described as majestic and sovereign, is also a personal God; he is viewed as being nearer to man than man's jugular vein, and, whenever a person in need or distress calls him, he responds. Above all, he is the God of guidance and shows everything, particularly man, the right way, "the straight path."

The God of the Qur’ān

This picture of God—wherein the attributes of power, justice, and mercy interpenetrate—is related to the Judeo-Christian tradition, whence it is derived with certain modifications, and also to the concepts of pagan Arabia, to which it provided an effective answer. The pagan Arabs believed in a blind and inexorable fate over which man had no control. For this powerful but insensible fate the Qur’ān substituted a powerful but provident and merciful God. The Qur’ān carried through its uncompromising monotheism by rejecting all forms of idolatry and eliminating all gods and divinities that the Arabs worshipped in their sanctuaries (ḥarams), the most prominent of which was Ka‘bah sanctuary in Mecca itself.

The universe. In order to prove the unity of God, the Qur’ān lays frequent stress on the design and order in the universe. There are no gaps or dislocations in nature. Order is explained by the fact that every created thing is endowed with a definite and defined nature whereby it falls into a pattern. This nature, though it allows every created thing to function in a whole, sets limits; and this idea of the limitedness of everything is one of the most fixed points in both the cosmology and theology of the Qur’ān. The universe is viewed, therefore, as autonomous, in the sense that everything has its own inherent laws of behaviour, but not as autocratic, because the patterns of behaviour have been endowed by God and are strictly limited. "Everything has been created by us according to a measure." Though every creature is thus limited and "measured out" and hence depends upon God, God alone, who reigns unchallenged in the heavens and the earth, is unlimited, independent, and self-sufficient.

Man. According to the Qurʾān, God created two apparently parallel species of creatures, man and *jinn,* the one from clay and the other from fire. About the *jinn,* however, the Qurʾān says little, although it is implied that the *jinn* are endowed with reason and responsibility but are more prone to evil than man. It is with man that the Qurʾān, which describes itself as a guide for the human race, is centrally concerned. The Judeo-Christian story of the Fall of Adam (the first man) is accepted, but the Qurʾān states that God forgave Adam his act of disobedience, which is not viewed in the Qurʾān (in contradistinction to its understanding in the Christian doctrine) as original sin.

In the story of man's creation, angels, who protested to God against the creation of man, who "would sow mischief on earth," lost in a competition of knowledge against Adam. The Qurʾān, therefore, declares man to be the noblest of all creation, the created being who bore the trust (of responsibility) that the rest of the creation refused to accept. The Qurʾān thus reiterates that all nature has been made subservient to man: nothing in all creation has been made without a purpose, and man himself has not been created "in sport," his purpose being service and obedience to God's will.

Despite this lofty station, however, the Qurʾān describes human nature as frail and faltering. Whereas everything in the universe has a limited nature, and every creature recognizes its limitation and insufficiency, man is viewed as rebellious and full of pride, arrogating to himself the attributes of self-sufficiency. Pride, thus, is viewed as the cardinal sin of man, because by not recognizing in himself his essential creaturely limitations he becomes guilty of ascribing to himself partnership with God (*shirk:* associating a creature with the Creator) and of violating the unity of God. True faith (*īmān*), thus, consists of belief in the immaculate Divine Unity and Islām in one's submission to the Divine Will.

Pride, the cardinal sin

Satan, sin, and repentance. In order to communicate the truth of the Divine Unity, God has sent messengers or prophets to men, whose weakness of nature makes them ever prone to forget or even willfully reject the Divine Unity under the promptings of Satan. According to the Qurʾānic teaching, the being who became Satan (Shayṭān or Iblīs) had previously occupied a high station but fell from divine grace by his act of disobedience in refusing to honour Adam when he, along with other angels, was ordered to do so. Since then, his work has been to beguile man into error and sin. Satan is, therefore, the contemporary of man, and Satan's own act of disobedience is construed by the Qurʾān as the sin of pride. Satan's machinations will cease only on the Last Day.

Judging from the accounts of the Qurʾān, the record of man's accepting the prophets' messages has been rather dismal. The whole universe is replete with signs of God; the human soul itself is viewed as a witness of the unity and grace of God. The messengers of God have, throughout history, been calling man back to God. Yet very few men have accepted the truth; most of them have rejected it and become disbelievers (*kāfir,* plural *kuffār:* literally "ungrateful"—*i.e.,* to God), and when man becomes so obdurate, his heart is sealed by God. Nevertheless, it is always possible for a sinner to repent (*tawbah*) and redeem himself by a genuine conversion to the truth. There is no point of no return, and God is always willing and ready to pardon. Genuine repentance has the effect of removing all sins and restoring a person to the state of sinlessness with which he started his life.

Prophecy. Prophets are men specially elected by God to be his messengers. Prophethood is indivisible, and the Qurʾān requires recognition of all prophets as such without discrimination. Yet they are not all equal, some of them being particularly outstanding in qualities of steadfastness and patience under trial. Abraham, Noah, Moses, and Jesus were such great prophets. As vindication of the truth of their mission, God often vests them with miracles: Abraham was saved from fire, Noah from the deluge, and Moses from the Pharaoh. Not only was Jesus born from the Virgin Mary, but God also saved him from crucifixion at the hands of the Jews. The conviction that God's messengers are ultimately vindicated and saved is an integral part of the Qurʾānic doctrine.

All prophets are human and never part of divinity: they are simply recipients of revelation from God. God never speaks directly to a human: he either sends an angel messenger to him or makes him hear a voice or inspires him. Muḥammad is accepted as the last prophet in this series and its greatest member, for in him all the messages of earlier prophets were consummated. He had no miracles except the Qurʾān, the like of which no human can produce. (Soon after the Prophet's death, however, a plethora of miracles was attributed to him by Muslims.) The angel Gabriel brought the Qurʾān down to the Prophet's "heart." Gabriel is represented by the Qurʾān as a spirit, but the Prophet could sometimes see and hear him. According to early traditions, the Prophet's revelations occurred in a state of trance when his normal consciousness was in abeyance. This state was accompanied by heavy sweating. The Qurʾān itself makes it clear that the revelations brought with them a sense of extraordinary weight: "If we were to send this Qurʾān down on a mountain, you would see it split asunder out of fear of God."

Muḥammad as the last prophet

This phenomenon at the same time was accompanied by an unshakable conviction that the message was from God, and the Qurʾān describes itself as the transcript of a heavenly "Mother Book" written on a "Preserved Tablet." The conviction was of such an intensity that the Qurʾān categorically denies that it is from any earthly source, for in that case it would be liable to "manifold doubts and oscillations."

Eschatology. In Islāmic doctrine, on the Last Day, when the world will come to an end, the dead will be resurrected and a judgment will be pronounced on every person in accordance with his deeds. Although the Qurʾān in the main speaks of a personal judgment, there are several verses that speak of the resurrection of distinct communities that will be judged according to "their own book." In conformity with this, the Qurʾān also speaks in several passages of the "death of communities," each one of which has a definite term of life. The actual evaluation, however, will be for every individual, whatever the terms of reference of his performance. In order to prove that the resurrection will occur, the Qurʾān uses a moral and a physical argument. Because not all requital is meted out in this life, a final judgment is necessary to bring it to completion. Physically, God, who is all-powerful, has the ability to destroy and bring back to life all creatures, who are limited and are, therefore, subject to God's limitless power.

According to strict Qurʾānic doctrine, there is no intercession, although God himself, in his mercy, may forgive certain sinners. Those condemned will burn in hellfire, and those who are saved will enjoy the abiding pleasures of paradise. Hell and heaven are both spiritual and physical. Besides suffering in physical fire, the damned will also experience fire "in their hearts"; similarly, the blessed, besides physical enjoyment, will experience the greatest happiness of divine pleasure. Quite early, however, Islāmic tradition developed the notion of intercession, probably in answer to the Christian doctrine of redemption.

Social service. Because the purpose of the existence of man, as of every other creature, is submission to the Divine Will, God's role in relation to man is that of the commander. Whereas the rest of nature obeys God automatically, man alone possesses the choice to obey or disobey. With the deep-seated belief in Satan's existence, man's fundamental role becomes one of moral struggle, which constitutes the essence of human endeavour. Recognition of the unity of God does not simply rest in the intellect but entails consequences in terms of the moral struggle, which consists primarily in freeing oneself of narrowness of mind and smallness of heart. One must go out of oneself and expend one's best possessions for the sake of others.

God in the role of commander

The doctrine of social service, in terms of alleviating suffering and helping the needy, constitutes an integral part of the Islāmic teaching. Praying to God and other religious acts are deemed to be a pure facade in the absence of active welfare service to the needy. In regard to this matter, the Qurʾānic criticisms of human nature become very sharp: "Man is by nature timid; when evil befalls him, he panics, but when good things come to him he prevents them from reaching others." It is Satan who whispers into man's ears that by spending for others he will become poor. God, on the contrary, promises prosperity in exchange for such expenditure, which constitutes a credit with God and grows much more than the money people invest in usury. Hoarding of wealth without recognizing the rights of the poor is threatened with the direst punishment in the hereafter and is declared to be one of the main causes of the decay of societies in this world. The practice of usury is forbidden.

The concept of a community of the faithful

With this socioeconomic doctrine cementing the bond of faith, the idea of a closely knit community of the faithful who are declared to be "brothers unto each other" emerges. Muslims are described as "the middle community bearing witness on mankind," "the best community produced for mankind," whose function it is "to enjoin good and forbid evil" (Qurʾān). Cooperation and "good advice" within the community are emphasized, and a person who deliberately tries to harm the interests of the community is to be given exemplary punishment. Opponents from within the community are to be fought and reduced with armed force, if issues cannot be settled by persuasion and arbitration.

Because the mission of the community is to "enjoin good and forbid evil" so that "there is no mischief and corruption" on earth, the doctrine of *jihād,* in view of the constitution of the community as the power base, is the logical outcome. For the early community it was a basic religious concept. *Jihād,* or holy war, means an active struggle using armed force whenever necessary. The object of *jihād* is not the conversion of individuals to Islām but rather the gaining of political control over the collective affairs of societies to run them in accordance with the principles of Islām. Individual conversions occur as a by-product of this process when the power structure passes into the hands of the Muslim community. In fact, according to strict Muslim doctrine, conversions "by force" are forbidden, because after the revelation of the Qurʾān "good and evil have become distinct," so that one may follow whichever one may prefer (Qurʾān), and it is also strictly prohibited to wage wars for the sake of acquiring worldly glory, power, and rule. With the establishment of the Muslim empire, however, the doctrine of the *jihād* was modified by the leaders of the community. Their main concern had become the consolidation of the empire and its administration, and thus they interpreted the teaching in a defensive rather than in an expansive sense. The Khārijite sect (see below *Theology and sectarianism*) which held that "decision belongs to God alone," insisted on continuous and relentless *jihād,* but its followers were virtually destroyed during the internecine wars in the 8th century.

Qurʾān and *jihād*

Besides a measure of economic justice and the creation of a strong community ideal, the Prophet Muḥammad effected a general reform of the Arab society, in particular protecting its weaker segments—the poor, the orphans, women, and slaves. Slavery was not legally abolished, but emancipation of slaves was religiously encouraged as an act of merit. Slaves were given legal rights, including the right of acquiring their freedom against payment, in installments, of a sum agreed upon by the slave and his master out of his earnings. A slave woman who bore a child by her master became automatically free after her master's death. The infanticide of girls that was practiced among certain tribes—out of fear of poverty or a sense of shame—was forbidden.

Distinction and privileges based on tribal rank or race were repudiated in the Qurʾān and in the celebrated "Farewell Pilgrimage Address" of the Prophet shortly before his death. All men are therein declared to be "equal children of Adam," and the only distinction recognized in the sight of God is to be based on piety and good acts. The age-old Arab institution of intertribal revenge (called *thaʾr*)—whereby it was not necessarily the killer who was executed but a person equal in rank to the slain person—was abolished. The pre-Islāmic ethical ideal of manliness was modified and replaced by a more humane ideal of moral virtue and piety.

(F.R./Ed.)

ORIGINS AND COMPILATION OF THE QURʾĀN

Revelation of the Qurʾān to the Prophet

Muslim tradition. According to Muslim tradition the Qurʾān was revealed to Muḥammad in separate pieces over some 20 years. On such occasions, Muḥammad, it is said, was in a kind of trance or ecstasy, during which the revelations were brought to him by the angel Gabriel. On his return to normal consciousness he recited the words of revelation to those present. There are many traditions about the occasions on which a certain *sūrah* or part of a *sūrah* was revealed. Thus the revelation of the Qurʾān is connected with events in the life of the Prophet. Even the traditional recension (version) of the Qurʾān itself classifies the *sūrah*s as Meccan or Medinan.

Obviously, many people learned the words of the revelation by heart, but there are also traditions that, at the time of their revelation, Muḥammad had them written down on "pieces of paper, stones, palm-leaves, shoulder-blades, ribs, and bits of leather," *i.e.,* whatever writing-material there was at hand. It is believed that the Prophet indicated to the scribes the context in which a certain passage should be placed.

Establishment of an authoritative text

After the Prophet's death, and especially after the battle of Yamāmah (633), in which a great number of those who knew the Qurʾān by heart had fallen, fear arose that the knowledge of the Qurʾān might disappear. So it was decided to collect the revelations from all available written sources and, as Muslim tradition has it, "from the hearts [*i.e.,* memories] of people." A companion of the Prophet, Zayd ibn Thābit, is said to have copied on sheets whatever he could find and to have handed it over to the caliph ʿUmar. After ʿUmar's death the collection was left in the care of his daughter Ḥafṣah. Other copies of the Qurʾān appear to have been written later, and different versions were used in different parts of the Muslim empire. So that there would be no doubt about the correct reading of the Qurʾān, the caliph ʿUthmān (644–656) is reported to have commissioned Zayd ibn Thābit and some other learned men to revise the Qurʾān using the "sheets" of Ḥafṣah, comparing them with whatever material was at hand, and consulting those who knew the Qurʾān by heart. It was decided that in case of doubt about the pronunciation, the dialect of Quraysh, the Prophet's tribe, was to be given preference. Thus an authoritative text of the Qurʾān (now known as the ʿuthmānic recension) was established.

These traditions may have been reworked and changed to some extent to suit certain dogmatic theories concerning the Qurʾān, but in the main they reflect historical truth. It is obvious that the description of the method of revelation has been somewhat simplified. The Qurʾān itself states (42:50–52) that God spoke to Muḥammad "by suggestion, or from behind a veil, or by sending a messenger to suggest what he pleases." The first term (Arabic *waḥy*) denotes a "suggestion" or "inspiration" of the kind that is well known by many poets; the Qurʾān also uses a term meaning "it was sent down." The second term seems to suggest some kind of imaginative locution without any accompanying vision. Only the third expression alludes to an angel but without mentioning the name of Gabriel.

Views of those outside Islām. The chronology of the *sūrah*s is a much debated problem. The existing traditions concerning the occasions for the revelation of certain passages cannot always be controlled and may or may not be reliable. European scholars have applied the criteria of style and contents to establish the relative order of the *sūrah*s or parts of *sūrah*s. From the time when Theodor Nöldeke published his *History of the Qurʾān* (1860), it has been common to arrange the *sūrah*s in four groups, deriving from three subsequent periods at Mecca and from Medina. The above exposition of the content of the Qurʾān roughly follows this arrangement.

In the Muslim view, Muḥammad received every word of the Qurʾān directly from God. The Qurʾān describes, and indignantly rejects, accusations that the Prophet had reproduced things that he had drawn from other sources. Western scholars who have analyzed the contents of the various revelations have shown that much of the narrative material concerning biblical persons and events differs from the biblical account and seems to have come from later Christian and, above all, from Jewish sources (*e.g.,* Midrash). Other motifs, such as the idea of the impending judgment and the descriptions of paradise agree with standard topics in the missionary preaching of the contemporary Syriac church fathers. The dependence need not, however, be of a literary kind, but might be due to influence from oral traditions.

Question of influence of oral traditions on text

It would appear that learning the words of the revelation by heart was the normal way of preserving them, and that only on special occasions were the words written down immediately. The existence of various early collections of Qurʾānic material seems to be a warranted fact, although their nature and contents cannot be determined. Some of the *sūrah*s beginning with separate letters (*al-fawātiḥ*)—certain consonant combinations detached from the main text (mentioned above under the heading *Form*)—occur together in the present Qurʾān and in the order of decreasing length in such a way as to suggest that they once formed separate collections. The establishment of a vulgate recension (a standard version) was not sufficient to secure the uniform and correct reading of the Qurʾān in all details. The Arabic script was incomplete; several consonants were easy to confuse, and there was no way of indicating the vowels to differentiate the variety of possible meanings inherent in a particular combination of consonants. To assure the correct recitation, therefore, it was necessary to know the text more or less by heart. In this way, differing variant readings arose, warranted by this or that "reader" of the Qurʾān.

The recorded variations, however, turned out to be remarkably few, and though no complete listing of the textual variants exists, it can safely be said that the textual tradition of the Qurʾān is much firmer and more uniform than that of the New Testament. The Arabic script was gradually improved. Diacritical signs were introduced to distinguish the letters that were similar in form, and long vowels were indicated by the letters *alif* (for *ā*), *wāw* (for *ū*), and *yā* (for *ī*). It is known that this vowel system was still disputed at the beginning of the 9th century. The special vowel signs placed above or beneath the letters were added in a different colour and did not count as part of the text itself.

INTERPRETATIONS

The "readers" (*qurrāʾ,* singular *qāriʾ*) were the specialists of the text of the Qurʾān. They were at the same time philologians, and it was to a great extent from their dealings with the language of the Qurʾān that the science of Arabic grammar grew. Two schools developed, one at Baṣra (in present-day Iraq), which was especially interested in systematizing and ordering the material to set up the rules governing the language, and a rival one at Kūfa (also in Iraq), which took more interest in the exceptional. It was theorized that several variant readings could be accepted only if they were based on the ʿUthmānic recension (version). It was also important that a reading be based on the authority of some renowned reader.

There was also theological speculation as to the true nature of the Qurʾān. In the discussions initiated by the Muʿtazilites (Seceders; literally, "those who stand apart"; a group that sought to introduce philosophical principles from Greek rationalism into Islāmic thought) the question of the eternity of the Qurʾān (*i.e.,* of its heavenly prototype) was one of the main points. The Muʿtazilites, who wanted to avoid everything that might compromise or encroach upon the oneness of God, denied the doctrine that the Qurʾan was uncreated and eternal, because this would mean that something else besides the God of eternity would exist eternally and thus create an eternal and irreconcilable "dualism." Consequently they asserted that the Qurʾān was created by God. This doctrine, however, was rejected by orthodox adherents of Islām. In popular belief, the reverence for the Qurʾān is often directed toward the visible, physical book or parts of it. Oaths are taken on it, and passages are sometimes copied out of it to be used for magical or superstitious purposes.

In these and other doctrinal disputes the parties sought support for their opinions in the sayings of the Qurʾān, since it was considered as the ultimate authority in all legal and religious questions. The correct interpretation of the Qurʾān became the object of a special branch of learning, the so-called *tafsīr,* or Qurʾānic exegesis. All kinds of resources were utilized in order to elucidate the meaning of a Qurʾānic passage. Traditions concerning the circumstances surrounding the revelation of certain passages or containing interpretative utterances of the Prophet that had been transmitted orally were recorded and collected, together with other traditions deriving from and concerning the Prophet (Ḥadīth). At times, in order to provide authority for a certain theory, traditions were simply invented. Any interpretation of a Qurʾānic passage that could not be supported by Ḥadīth was originally rejected. The results of the study of grammar and lexicography were also utilized; examples from contemporary poetry were often quoted in order to elucidate the grammatical structure or the lexical meaning of a passage. Thus, work on the Qurʾān, whose ultimate goal was the correct understanding and application of its teachings, went hand in hand with the development of Arabic grammar and lexicography.

Qurʾānic exegesis

Two works are especially renowned in the field of *tafsīr,* namely the commentary of aṭ-Ṭabarī (839–923), a huge encyclopaedic collection that sums up everything that had been done so far in the field, and the *Kashshāf* of Zamakhsharī (1075–1143), which has gained almost canonical reputation, though its author was a Muʿtazilite and began his work with the words, "Praise be to God who created the Qurʾān." A handy commentary of Bayḍāwī (d. *c.* 1280), which is often quoted as authoritative, is merely an abridged revision of the latter work.

The theological schools of medieval Islām all sought to support their doctrines with the aid of Qurʾānic exegesis, and each of them produced their own commentaries. There are also examples of allegorical interpretation (*taʾwīl*) especially in Ṣūfī (Islāmic mystical) literature, in which the doctrines of mysticism are found to be hidden behind the literal sense of the Qurʾānic word.

Qurʾānic exegesis gained new significance with the appearance of modernism toward the end of the 19th century. The modernists, who sought to revive Islām from its degradation and to reconcile it with what they found valuable in Western scientific traditions, set up the principle of returning to the pure and uncorrupted Islām of the "ancestors." As a consequence, the interpretation of the oldest and original source of Islām was regarded as imperative, and attempts were made to establish the principles necessary for a correct understanding of the Qurʾān. Traditional exegesis was accused of having introduced Israelite legends and false traditions that had nothing to do with the original teachings of the Prophet. On the other hand, the authority of the Qurʾān was never called in question.

Modern commentaries

Muḥammad ʿAbduh, the founder of modernism in Egypt, for several years published exegetical lectures in the journal *al-Manār;* and they were later published in book form by his Syrian disciple Rashīd Riḍā. In them he accepts the Qurʾān as the literally inspired word of God, in which there can be nothing false or antiquated, and tries to show that the results of modern science and many modern views are already present in the Qurʾān. This is often achieved by twisted interpretations, reading modern ideas into the words of the Qurʾān. For instance, the *jinn* (genii) of *sūrah* 2:176 that cause disease are interpreted as "microbes," and the words in 2:250, "How often a little company has overcome a numerous company; and God is with those who endure," is taken to refer to ideas reminiscent of Darwin's theory of the struggle for life and the survival of the fittest. Allegorical interpretation is also used when it can serve the purpose of the author. Other modernistic interpreters of the Qurʾān have continued along the same lines. The Qurʾān is, however, left untouched by criticism; as the infallible word of God it cannot have been influenced by the circumstances under which it was revealed, it can contain no mistake, and it cannot be superseded by any new discovery.

Later developments, however, have brought some new ideas to the fore. In an Urdu commentary on the Qur'ān, which has in part been made available in English, Maulana Abul Kalam Azad (1888–1958), an Indian Muslim scholar (minister of education of the Republic of India at the time of his death), developed some new principles for the interpretation of the Qur'ān. He argues that it is necessary to interpret the Qur'ān against the background of its environment; therefore it is necessary to study the cultures and the languages of ancient Arabia and other Semitic peoples. Study of the historical circumstances in which the Qur'ān came into being is said to facilitate the understanding of what it meant to those who received the revelation.

Scholars have no doubt, however, that there are new developments in the field of Qur'ānic exegesis. D. Rahbar, in his study *The God of Justice* (1960), argues that in order to elucidate a passage in the Qur'ān one should quote traditional exegesis and medieval dogmatics and, above all, use other Qur'ānic passages for comparison, letting one passage throw light on another. Though such ideas are looked upon with suspicion by orthodox Muslims and are fervently rejected by most Muslim leaders, they may indicate the inception of a more historical view of the Qur'ān, one that tries to distinguish between central religious ideas and those outward things that are dependent on the historical environment.

TRANSLATIONS

The Qur'ān was revealed to Muḥammad as "an Arabic book" or an Arabic reading (*qur'ān*), to provide the Arabs with a holy book in their own language, comparable with the Scriptures of Judaism and Christianity. As has been noted, the language of the Qur'ān is regarded as surpassing everything that can be written in Arabic. The Qur'ān itself is a miracle and cannot be imitated by man.

As a consequence of this, it is regarded as unfitting to translate the Qur'ān. In countries in which other languages are spoken, the Qur'ān is still recited in Arabic. There exist Muslim translations of the Qur'ān; *e.g.*, into Turkish, Urdu, and English (the latter during the Aḥmadiyah movement founded in 1889 by Mirza Ghulam Aḥmad in the Punjab region of India), but on principle these are regarded as paraphrases, not as translations that can be used for ritual purposes.

The Qur'ān was first printed in Arabic at Rome by Pagninus Brixiensis (1530), but the edition was never circulated. A. Hinckelmann published an Arabic text at Hamburg in 1694. Since then several European editions have appeared; one of the best was that of G. Flügel (1834), the first critical edition, often reprinted. It is from this edition that Western scholars have usually quoted the Qur'ān. Several editions are today printed in Muslim countries, and an official Egyptian edition is gaining more and more ground among Western scholars.

First critical edition

The first Latin translation was made in 1143 at the request of an abbot of the monastery of Cluny and was published at Basle in 1543 by Theodor Bibliander and afterward rendered into Italian, German, and Dutch. The first French translation was by A. du Ryer (1647); it was translated into English by Alexander Ross (1649–88). G. Sale's English translation first appeared in 1734 and has passed through many new editions. It has become something of a classic and can still be useful in many respects. A translation by J.M. Rodwell, with the *sūrah*s arranged in chronological order, appeared in 1861. E.H. Palmer's translation was published in Sacred Books of the East in 1880. Bell's translation "with a critical rearrangement of the sūrahs" (1937–39) tries to analyze the *sūrah*s into their smallest units and show how these were joined together to form the present Qur'ān. (See *Bibliography* for contemporary English translations.)

The Qur'ān has also been translated into most other European languages. Special mention should be made of R. Blachère's French translation (1949–50) because of its rather detailed notes, and of R. Paret's German rendering (1962), which is very accurate and makes extensive use of parallel passages within the Qur'ān itself, but is rather dry in its style. (H.R./Ed.)

Ḥadīth, traditions of the Prophet

Ḥadīth is the record of the traditions or sayings of the Prophet Muḥammad, revered and received as a major source of religious law and moral guidance, second only to the authority of the Qur'ān, or scripture of Islām. It might be defined as the biography of Muḥammad perpetuated by the long memory of his community for their exemplification and obedience. The development of Ḥadīth is a vital element during the first three centuries of Islāmic history, and its study provides a broad index to the mind and ethos of Islām.

NATURE AND ORIGINS

The term Ḥadīth derives from the Arabic root *ḥdth*, meaning "to happen," and so, "to tell a happening," "to report," "to have, or give, as news," or "to speak of." It means tradition seen as narrative and record. From it comes *sunnah* (literally, a "well-trodden path," *i.e.*, taken as precedent and authority or directive), to which the faithful conform in submission to the sanction that Ḥadīth possesses and that legalists, on that ground, can enjoin. Tradition in Islām is thus both content and constraint, Ḥadīth as the biographical ground of law and *sunnah* as the system of obligation derived from it. In and through Ḥadīth, Muḥammad may be said to have shaped and determined from the grave the behaviour patterns of the household of Islām by the posthumous leadership his personality exercised. There were, broadly, two factors operating to this end. One was the unique status of Muḥammad in the genesis of Islām; the other was the rapid geographical expansion of the new faith in the first two centuries of its history into various areas of cultural confrontation. Ḥadīth cannot be rightly assessed unless the measure of these two elements and their interaction is properly taken.

Sunnah

The experience of Muslims in the conquered territories of west and middle Asia and of North Africa was related to their earlier tradition. Islāmic tradition was firmly grounded in the sense of Muḥammad's personal destiny as the Prophet—the instrument of the Qur'ān and the apostle of God. The clue to tradition as an institution in Islām may be seen in the recital of the *Shahādah* or "witness" ("There is no god but God; Muḥammad is the prophet of God"), with its twin items as inseparable convictions—God and the messenger. Islāmic tradition follows from the primary phenomenon of the Qur'ān, received personally by Muḥammad and thus inextricably bound up with his person and the agency of his vocation. Acknowledgment of the Qur'ān as scripture by the Islāmic community was inseparable from acknowledgment of Muḥammad as its appointed recipient. In that calling, he had neither fellow nor partner, for God, according to the Qur'ān, spoke only to Muḥammad. When Muḥammad died, therefore, in AD 632, the gap thus created in the emotions and the mental universe of Muslims was shatteringly wide. It was also permanent. Death had also terminated the revelation embodied in the Qur'ān. By the same stroke scriptural mediation had ended, as well as prophetic presence.

The Prophet's death was said to have coincided with the perfection of revelation. But the perfective closure of both the book and the Prophet's life, though in that sense triumphant, was also onerous, particularly in view of the new changing circumstances, both of space and time, in the geographical expansion of Islām. In all the new pressures of historical circumstance, where was direction to be sought? Where, if not from the same source as the scriptural mouthpiece, who by virtue of that consummated status had become the revelatory instrument of the divine word and could therefore be taken as an everlasting index to the divine counsel? The instinct for and the growth of tradition are thus integral elements in the very nature of Islām, Muḥammad, and the Qur'ān. Ongoing history and the extending dispersion of Muslim believers provided the occasion and spur for the compilation of Ḥadīth.

HISTORICAL DEVELOPMENT

The appeal of the ordered recollection of Muḥammad to the Islāmic mind did not become immediately formalized and sophisticated. On the contrary, there is evidence that the full development of Ḥadīth was slow and uneven. Time and distance had to play their role before memory became stylized and official.

Relation to early Arabic poetry

Literary tradition in pre-Islāmic Arabia. The first generation had its own immediacy of Islāmic experience, both within the life span of the Prophet and in the first quarter century afterward. It had also the familiar patterns of tribal chronicle in song and saga. Pre-Islāmic poetry celebrated the glory of each tribe and their warriors. Such poetry was recited in honour of each tribe's ancestors. The vigour and élan of original Islām took up these postures and baptized them into Muslim lore. The proud history of which Muḥammad was the crux was, naturally, the ardent theme, first of chronicle, and then of history writing. Both needed and stimulated the cherishing of tradition. The lawyers, in turn, took their clues from the same source. While the Qurʾān was being received, there had been reluctance and misgiving about recording the words and acts of the Prophet, lest they be confused with the uniquely constituted contents of the scripture. Knowledge of Muḥammad's disapproval of the practice of recording his words is evidence enough that the practice existed. With the Qurʾān complete and canonized, those considerations no longer obtained; and time and necessity turned the instinct for Ḥadīth into a process of gathering momentum.

Developments of the 1st and 2nd centuries AH. Within the first century of the Prophet's death, tradition had come to be a central factor in the development of law and the shape of society. Association by Ḥadīth with Muḥammad's name and example became increasingly the ground of authority. The 2nd century brought the further elaboration of this relationship by increasing formalism in its processes. Traditions had to be sustained by an expert "science" of attestation able to satisfy rigorous formal criteria of their connection with the person of Muḥammad through his "companions," by an unbroken sequence of "reportage" (see below). This science became so meticulous that it is fair (even if also paradoxical) to suspect that the more complete and formally satisfactory the attestation claimed to be, the more likely it was that the tradition was of late and deliberate origin. The developed requirements of acceptability that the tradition boasted simply did not exist in the early, more haphazard and spontaneous days.

It is clear that many customs and usages native to non-Arab societies prior to their Islāmization found their way into Islām in the form of reputed or alleged traditions of Muḥammad, though always on the condition of their general compatibility with the Islāmic religion. Implicit in this sense in Muḥammad's personal example and genius, tradition inferred an elasticity and an embrace large enough to comprehend and anticipate all that Islām in its wide geographical experience was to become.

Qurʾānic commentary

Qurʾānic commentary, as it developed in the wake of these other factors of law and custom, also leaned heavily on traditional material, for the incidents of the Qurʾānic narrative and the occasions of revelation could best be understood by what tradition had to say in its reporting of them. Further, since the patterns of Qurʾānic commentary were largely hortatory, Ḥadīth was a ready mine of word and story calculated to exemplify and reinforce what exhortation commended. Except in rare and controversial cases (the so-called Ḥadīth Qudsī, or Holy Tradition), these traditional factors in Qurʾānic interpretation were only elucidatory, and the substance of tradition could in no way dispute or displace the essential, primary, authority of the Qurʾānic text. For the *obiter dicta* (incidental observations) of Muḥammad, though sacrosanct, lacked the hallmark of revelation, which belonged solely to the Qurʾān. Among earliest developed examples of Ḥadīth are the narratives of the biographer Ibn Isḥāq (died AH 150 [AD 767]) and the compilation of laws by Mālik ibn Anas, known as al-Muwaṭṭaʾ (died AH 179 [AD 795]). But they preceded by less than half a century the success of the theory that made tradition indispensable to the valid development of Islāmic law.

3rd century AH and subsequent developments. The chief protagonist of the view correlating tradition and law was Muḥammad ash-Shāfiʿī (died AH 204 [AD 820]) who claimed for tradition a divine imprint as an extension of the revelation of the Qurʾān. It was in line with this conviction that the phrase "the Qurʾān and the *sunnah*" became current to describe the fount of authority in Sunnī Islām (the major traditionalist sect). By this mandate and out of the needs and inventiveness of lawyers, the mass of tradition grew apace. When virtually no issues could be argued, still less settled, except by connection with cited acts and opinions of Muḥammad, the temptation to require or to imagine or to allege such traditions became irresistible. Supply approximated to demand, and the growth of both made more ingenious and pretentious the science of supporting attribution. The increasing volume and complexity of the material contained in Ḥadīth necessitated larger compilations and more detailed classification. These factors worked together to inspire a critical editorial activity that in the course of the 3rd century generated what have come to be regarded as the six canonical collections of Ḥadīth by Sunnī Muslims. The first two of them have acquired a status of great sanctity. Before noting these it is convenient to describe the editorial task and the editorial procedures that constitute the developed science of Ḥadīth criticism.

THE SCIENCE OF ḤADĪTH

Matn and *isnād*

The study of tradition distinguishes between the substance, or content, known as the "gist" (*matn*) of the matter, and the "leaning" (*isnād*) or chain of corroboration on which it hangs.

Form of Ḥadīth and criteria of authentication. That Muḥammad observed, "Seek knowledge, though it be in China" or "Beware of suspicion, for it is the falsest of falsehoods" reveals the *matn* or "the meat of the matter." The formula introducing such a Ḥadīth would speak in the first person: "It was related to me by A, on the authority of B, on the authority of C, on the authority of D, from E (here a companion of Muḥammad) that the Prophet said . . . ." This chain of names constituted the *isnād* on which the saying or event depended for its authenticity. The major emphases in editing and arguing from tradition always fell on the *isnād,* rather than on a critical attitude to the *matn* itself. The question was not, "Is this the sort of thing Muḥammad might credibly be imagined to have said or done?" but "Is the report that he said or did it well supported in respect of witnesses and transmitters?" The first question would have introduced too great a danger of subjective judgment or independence of mind, though it may be suspected that issues were in fact often decided by such critical appraisal in the form of decisions ostensibly relating only to *isnād.* The second question certainly allowed a theoretically objective and reasonably precise pattern of criteria.

If the adjacent names in the chain of transmission overlapped in life, there was certainty that they could have listened to one another. Their travels were also investigated to see if their paths could have really crossed. Biographies could be built up to show that they were honest men and spoke truly. Comparative study could be made of their reputations for veracity as acknowledged by their contemporaries or indicated by their traditions when compared. The frequency of currency through several sources was yet another element in the testing of traditions. Most important of all was the final link with the "companion," who in the first instance had the tradition from his or her contact with the Prophet.

Weight of traditions

Classifications. In all these ways, and others involving more minutiae, it was possible to establish categories of Ḥadīth quality. Traditions might be sound (*ṣaḥīḥ*), good (*ḥasan*), or weak (*ḍaʿīf*). Other terms, such as healthy (*ṣāliḥ*) and infirm (*saqīm*), were also current. Each of the three classifications was liable to subdivisions, depending on refinements of assessment and, later, on their standing with the classic compilers. Distinctions were less rigorously seen if the traditions were cited not for legal definitions but merely for moral purposes. A *ḍaʿīf* tradition, for example, might well be salutary for exhortation, even if lawyers were required to exclude or ignore it. Traditions also varied in strength according to whether one or more "companions" could be adduced, whether the *isnād* had parallels, whether they were continuous back to Muḥammad (*muttaṣil*), or intermitted (*mawqūf*). The subtleties in these and other questions were part of the active competence that attended the whole science.

The repute and authority of the canonical collections did much to stabilize the situation, but only because their emergence demonstrated that the zest for tradition had overreached itself. By the end of the 3rd century AH it was sorely necessary to solidify Ḥadīth into a stable corpus of material to which no new element could credibly be added and from which extravagances had been purged. The Ḥadīth tradition within the various traditions had by then become a permanent and disciplined element in the authority structure of Islām—the second great source of law and practice, complementary to the Qur'ān and available for analogical handling (*qiyās*) and for consensus (*ijtihād*) as further sources of legislation, arguing from the Qur'ān and the Sunnah as primary. Shī'ah tradition (see below) stands apart from this structure of authority.

THE COMPILATIONS

The most revered of all traditionalists was Muḥammad ibn Ismā'īl al-Bukhārī (AH 194–56 [AD 810–870]), whose *Kitāb al-Jāmi' aṣ-Ṣaḥīḥ* (*The Book of the Authentic Collection*) has a unique place in the awe and esteem of Muslims as a work of great historical import and deep piety. While a boy he made the pilgrimage to Mecca and gathered traditions in wide travels. According to tradition, he was inspired to his task by a vision of Muḥammad pestered by flies while asleep—flies that he (al-Bukhārī) fanned from the Prophet's face. The flies represented the cloud of spurious traditions darkening the true image, and the fan was its tireless rescuer. Whatever the truth of this narrative, it captures the temper of al-Bukhārī's vocation. His *Ṣaḥīḥ* occupied 16 years of editorial pains and scrutiny. He included 7,397 traditions with full *isnād.* Allowing for repetitions, the net total was 2,762, gathered, it is said, from over 600,000 memorized items. He arranged the whole into 97 books and 3,450 chapters or topics, repeating the traditions that bore on several themes.

Of comparable stature was the *Ṣaḥīḥ* of Muslim ibn al-Ḥajjāj (AH 202–261 [AD 817–875]), to which the compiler prefaced a discussion of the criteria of Ḥadīth. The material largely confirms his contemporaries, and all such traditions common to these two authorities are known as agreed (*muttafaq*). It became characteristic to give freer rein to prevailing or communal assent in matters of *isnād.*

There are four other classical collections of tradition, all belonging within the 3rd century AH, and interdependent in part. Abū Dā'ūd al-Sijistānī (AH 202–275 [AD 817–889]) produced his *Kitāb as-Sunan* ("Book of traditions"), containing 4,800 traditions relating to matters of jurisprudence (as the term *Sunan* indicates, in contradistinction to a *Jāmi',* or collection embracing all fields). Abū 'Isā Muḥammad at-Tirmidhī (died AH 279 [AD 892]) edited the *Jāmi'aṣ-Ṣaḥīḥ,* adding notes on the distinctive interpretations of the schools of law (*madhāhib*). Abū' Abd ar-Raḥmān an-Nasā'ī (AH 216–303 [AD 830–915]) produced another *Kitāb as-Sunan* with special concern for the religious law relating to ritual acts. Abū 'Abdallāh ibn Mājā (AH 210–273 [AD 824–886]), a pupil of Abū Dā'ūd, compiled another with the same title but tended to a readier tolerance of less than satisfactory traditions. Preferences shifted between these four, and some were slower of recognition than others. Nor did they oust the earlier collection of Mālik ibn Anas, which maintained, if intermittently, its wide appeal. But they formed the increasing reliance of generations of Muslims, within the unique eminence of the master "pair," and formed the sources of later popular editions, intended to conflate material for didactic purposes. One such was the work of Abū Muḥammad al-Baghawī (died AH 516 [AD 1122]) called *Maṣābīḥ as-Sunnah* ("The Lamps of the Sunnah"). Commentaries on all these classical *musannafāt,* or compilations, were many, and important in education and piety.

SECTARIAN VARIATIONS

The tradition of the Shī'ah, a minority branch of Islām, (distinguished from the tradition of the Sunnah majority by belief in the special role of the Prophet's cousin 'Alī and his descendants) diverges sharply from a very early date, though the emphasis on the personality of Muḥammad was identical. The Shī'ah broke away from the (to be) dominant Sunnī stream of Islām for deep reasons of politics, emotion, and theology. There was the dispute about caliphal succession and the role of 'Alī, cousin and son-in-law of Muḥammad and fourth caliph, and bitter cleavage because of the tragic fate of his two sons and especially of Ḥusayn in the massacre of Karbalā', from which there ultimately evolved the theology of vicarious suffering epitomized in Shī'ī devotion and ritual. (Sectarian disputes are treated in detail; see below *Theology and sectarianism.*) All these factors inevitably involved the business of tradition. The schism read the origins according to the divided loyalties, and there was little that was not potentially contentious, apart from obvious matters; *e.g.,* Muḥammad's intentions for 'Alī and the caliphate. The issues were fought out in rivalry for the mind of the Prophet, the authority of which was the sole agreement in the very disputing of it. The Shī'ah thus rejected the tradition of the Sunnīs and developed their own corpus of tradition (though there is evidence that an-Nasā'ī, at least, among the classical compilers, had sympathy with aspects of their cause). They also questioned the Sunnī notions of *isnād* and of the community as a locus of authority and evolved their own system of submission to their *imāms* (Shī'ah leaders). This altered the whole role that tradition might play. The major Shī'ī compilations date from the 4th and 5th centuries and allow only traditions emanating from the house of 'Alī. The first of them is that of Abū Ja'far Muḥammad al-Qulīnī (died AH 328 [AD 939]), *Kāfī fī 'Ilm ad-Dīn,* which might be translated: "All You Need About the Science of Religious Practice."

Shī'ah and Sunnah

SIGNIFICANCE OF ḤADĪTH

Canonical collections of Ḥadīth are, for the non-Muslim, an introduction to a world of faith, of behaviour and authority, a world of almost encyclopaedic inclusiveness. Provisions of law are the primary element, enlarging Qur'ānic legislation. They contain a whole array of moral, social, commercial, and personal matters, as well as the themes of eschatology. All reaches of public and private conduct may be found there, from the disposal of a date stone to the crisis of the deathbed, from the manner of ablution to the duties of forgiveness, from the physical routines of digestion to the description of the day of judgment. There is a Talmudic capacity for detail and scrupulousness in legal and ethical prescriptions and precepts. There are stories of integrity and right action, for example, that of the purchaser of a plot of ground who subsequently unearthed in it a pot of gold, which he brought back to the former owner, protesting that it was not within his bargain. The vendor, likewise, refused to claim it since he had not known the gold was there when he sold his field. An arbitrator solved their dilemma of honesty by proposing the marriage of the son of one with the daughter of the other so that, after alms, the gold might be settled on the couple. Through and in tradition, Islām aligned itself authoritatively with all it found compatible in local usages and brought hospitably and masterfully within its purview the continuity of many cultures. There is wide evidence of the impact of Jewish and Christian elements, notably in the realm of eschatology, in the elaboration of the stark and urgent Qur'ānic doctrine of the last judgment. But always the imprint of Islām is clear. Tradition is at once a mine and a kind of currency, the source and the circulation of the values it makes and preserves. (A.K.C./Ed.)

Fundamental practices and institutions of Islām

THE FIVE PILLARS

During the earliest decades after the death of the Prophet, certain basic features of the religio-social organization of Islām were singled out to serve as anchoring points of the community's life and formulated as the "Pillars of Islām." To these five, the Khawārij sect added a sixth pillar, the *jihād,* which, however, was not accepted by the general community.

The shahādah or profession of faith. The first pillar is the profession of faith: "There is no god but God; Muḥammad is the prophet of God," upon which depends the membership in the community. The profession must be recited at least once in one's lifetime, aloud, correctly, and purposively, with an understanding of its meaning and with an assent from the heart. From this fundamental belief are derived beliefs in (1) angels (particularly Gabriel, the Angel of Revelation), (2) the revealed Books (the Qurʾān and the sacred books of Judeo-Christian revelation described in the Qurʾān), (3) a series of prophets (among whom Judeo-Christian figures are particularly eminent—although it is believed that God has sent messengers to every nation), and (4) the Last Day (Day of Judgment).

Basic beliefs deriving from the *shahādah*

Prayer. The second pillar consists of five daily congregational prayers, which may, however, be offered individually if one is unable to go to the mosque. The first prayer is performed in the morning before sunrise, the second just after noon, the third in the later afternoon, the fourth immediately after sunset, and the fifth before retiring to bed (only three prayers are mentioned in the Qurʾān: morning, evening, and the middle prayer in the afternoon).

Before a prayer, ablutions are performed by washing the hands, face, and feet. The muezzin (one who gives the call for prayer) chants aloud from a raised place (such as a tower) in the mosque. When prayer starts, the *imām,* or leader (of the prayer), stands in the front facing Mecca, and the congregation stands behind him in rows, following him in various postures. Each prayer consists of two to four genuflection units (*rakʿah*); each unit consists of a standing posture (during which verses from the Qurʾān are recited, in certain prayers aloud, in others silently), as well as a genuflection and two prostrations. At every change in posture, "God is great" is recited. Tradition has fixed the materials to be recited in each posture.

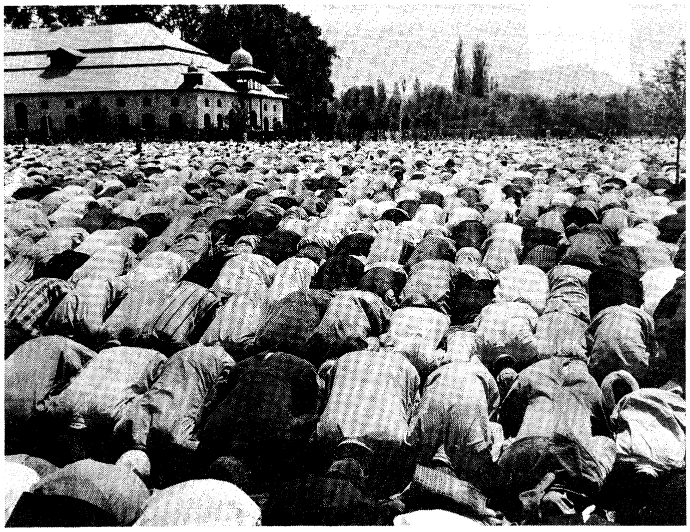
Special congregational prayers are offered on Friday instead of the prayer just after noon. The Friday service consists of a sermon (*khuṭbah*), part of which consists of preaching in the local language and part of recitation of certain formulas in Arabic. In the sermon, the preacher usually recites a verse of the Qurʾān and builds his address on it, which can be of a moral, social, or political content. Friday sermons have usually considerable impact on public opinion regarding sociopolitical questions.

Nature of sermons

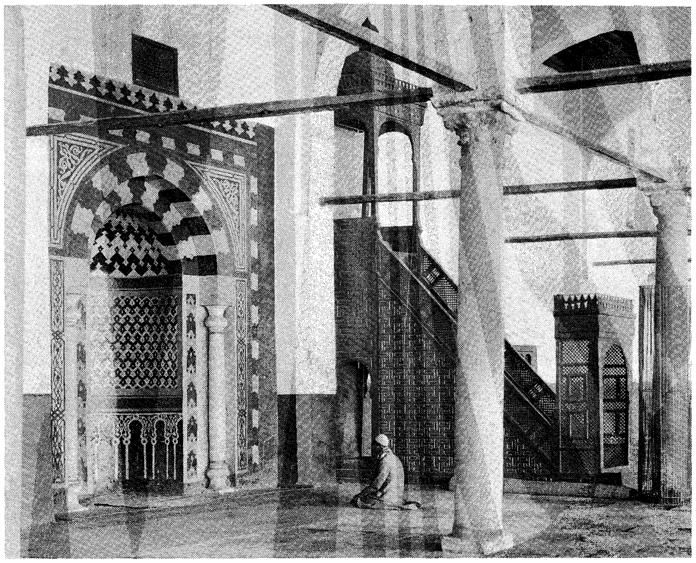
Although not ordained as an obligatory duty, nocturnal prayers (called *tahajjud*) are encouraged, particularly during the latter half of the night. During the month of Ramaḍān (see below *Fasting*) lengthy prayers are offered congregationally before retiring and are called *tarāwīḥ.*

In strict doctrine, the five daily prayers cannot be waived even for the sick, who may pray in bed and, if necessary, lying down. When on a journey, it is recommended that the two afternoon prayers be combined into one and the sunset and late evening prayers into one prayer as well. In practice, however, much laxity has occurred, particularly in modern times, although Friday prayers are still attended by large numbers.

Brian Brake—Magnum



Muslims at prayer, Kashmir, India.



Interior of the Mosque of ʿAmr ibn al-ʿĀṣ, Cairo, showing the *miḥrāb* (prayer niche) and the *minbar* (pulpit).

Lehnert & Landrock

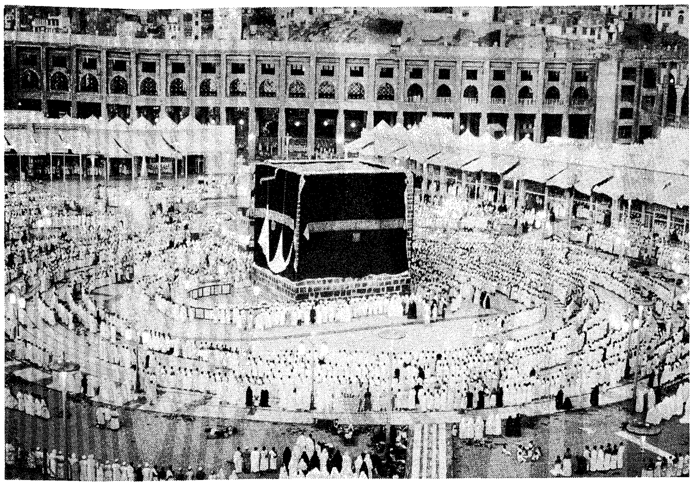
The zakāt. The third pillar is the obligatory tax called *zakāt* ("purification," indicating that such a payment makes the rest of one's wealth religiously and legally pure). This is the only permanent tax levied by the Qurʾān and is payable annually on food grains, cattle, and cash after one year's possession. The amount varies for different categories. Thus, on grains and fruits it is 10 percent if land is watered by rain, 5 percent if land is watered artificially. On cash and precious metals it is 2½ percent. *Zakāt* is collectable by the state and is to be used primarily for the poor, but the Qurʾān mentions other purposes: ransoming Muslim war captives, redeeming chronic debts of people, tax collectors' fees, *jihād* (and by extension, according to Qurʾān commentators, education and health), and creating facilities for travellers.

After the breakup of Muslim religio-political power, payment of *zakāt* has become a matter of voluntary charity dependent on individual conscience. Some Muslim countries are seeking to reintroduce it, and in several Middle Eastern countries *zakāt* is officially collected, but on a voluntary basis.

The fast of Ramaḍān

Fasting. Fasting during the month of Ramaḍān (ninth month of the Muslim lunar calendar), laid down in the Qurʾān (2:183–185), is the fourth pillar of the faith. Fasting begins at daybreak,and ends at sunset, and during the day eating, drinking, and smoking are forbidden. The Qurʾān (2:185) states that it was in the month of Ramaḍān that the Qurʾān was revealed. Another verse of the Qurʾān (97:1) states that it was revealed "on the night of determination," which Muslims generally observe on the night of 26–27 Ramaḍān. For a person who is sick or on a journey, fasting may be postponed until "another equal number of days." Daily feeding of one poor person is also prescribed "for those who can afford it."

The ḥajj. The fifth pillar is the annual pilgrimage (*ḥajj*) to Mecca prescribed for every Muslim once in a lifetime—"provided one can afford it" and provided a person has enough provisions to leave for his family in his absence. The pilgrimage rite begins every year on the 7th and ends on the 10th of the month of Dhū al-Ḥijjah (last month of the Muslim year). When the pilgrim is about six miles (ten kilometres) from the Holy City, he enters upon the state of *iḥrām:* he wears two seamless garments and neither shaves nor cuts his hair or nails until the ceremony ends. The principal activities consist of walking seven times around the Kaʿbah, a shrine within the Sacred Mosque; the kissing and touching of the Black Stone (Ḥajar al-Aswad); and the ascent of and running between Mt. Ṣafā and Mt. Marwah (which are now, however, mere elevations) seven times.



Worshippers encircling the holy Kaʿbah, Mecca, Saudi Arabia.
By courtesy of the Saudi Arabian Information Service, Royal Embassy of Saudi Arabia

At the second stage of the ritual, the pilgrim proceeds from Mecca to Minā, a few miles away; from there he goes to ʿArafāt, where it is essential to hear a sermon and to spend one afternoon. The last rites consist of spending the night at Muzdalifah (between ʿArafāt and Minā) and offering sacrifice on the last day of *iḥrām,* which is the *ʿīd* ("festival") of sacrifice.

Many countries have imposed restrictions on the number of outgoing pilgrims because of foreign-exchange difficulties. Because of the improvement of communications, however, the total number of visitors has greatly increased in recent years. In 1965 the number of visitors was estimated to be about 1,500,000, approximately 600,000 of them from outside Arabia. All Muslim countries send official delegations on the occasion, which is being increasingly used for religio-political congresses. At other times in the year, it is considered meritorious to perform the lesser pilgrimage (*ʿumrah*), which is not, however, a substitute for the *ḥajj* pilgrimage.

SACRED PLACES AND DAYS

The most sacred place for Muslims is the Kaʿbah sanctuary at Mecca, the object of the annual pilgrimage. It is much more than a mosque; it is believed to be the place where the heavenly bliss and power touches the earth directly. According to Muslim tradition, the Kaʿbah was built by Abraham. The Prophet's mosque in Medina is the next in sanctity. Jerusalem follows in third place in sanctity as the first *qiblah* (*i.e.,* direction in which the Muslims offered prayers at first, before the *qiblah* was changed to the Kaʿbah) and as the place from where Muḥammad, according to tradition, made his ascent (*miʿrāj*) to heaven. For the Shīʿah, Karbalāʾ in Iraq (the place of martyrdom of ʿAlī's son, Ḥusayn) and Meshed in Iran (where Imām ʿAlī ar-Riḍā is buried) constitute places of special veneration where the Shīʿah make pilgrimages.

Shrines of Ṣūfī saints. For the Muslim masses in general, shrines of Ṣūfī saints are particular objects of reverence and even veneration. In Baghdad, the tomb of the greatest saint of all, ʿAbd al-Qādir al-Jīlānī, is visited every year by large numbers of pilgrims from all over the Muslim world.

The Ṣūfī shrines, which were managed privately in earlier periods, are almost entirely owned by governments in the 20th century and are managed by departments of *awqāf* (plural of *waqf,* a religious endowment). The official appointed to care for a shrine is usually called a *mutawallī.* In Turkey, where such endowments formerly constituted a very considerable portion of the national wealth, all were confiscated by the regime of Mustafa Kemal Atatürk (president, 1928–38).

The mosque. The general religious life of the Muslims is centred around the mosque, and in the days of the Prophet and early caliphs the mosque was, indeed, the centre of all community life. Small mosques are usually supervised by the *imām* (one who administers the prayer service) himself, although sometimes also a muezzin is appointed. In larger mosques, where Friday prayers are offered, a *khaṭīb* (one who gives the *khuṭbah,* or sermon) is appointed for Friday service. Many large mosques also function as religious schools and colleges. Mosque officials are appointed by the government in most countries. In some countries, *e.g.,* Pakistan, most mosques are private and are run by the local community, although some of the larger ones are being increasingly taken over by the government departments of *awqāf.*

Holy days. The Muslim calendar (based on the lunar year) dates from the emigration (*hijrah*) of the Prophet from Mecca to Medina in AD 622. The two festive days in the year are the *ʿīds,* ʿĪd al-Fiṭr celebrating the end of the month of Ramaḍān and the other, ʿĪd al-Aḍḥā (the feast of sacrifice), marking the end of the pilgrimage. Because of the crowds, *ʿīd* prayers are offered either in very large mosques or on specially consecrated grounds. Other sacred times include the "night of determination" (believed to be the night in which God makes decisions about the destiny of individuals and the world as a whole) and the night of the ascension of the Prophet to heaven. The Shīʿah celebrate the 10th of Muḥarram (the first month of the Muslim year) to mark the day of the martyrdom of Ḥusayn. The Muslim masses also celebrate the death anniversaries of various saints in a ceremony called *ʿurs* (literally, "nuptial ceremony"). The saints, far from dying, are believed to reach the zenith of their spiritual life on this occasion. (F.R./Ed.)

ISLĀMIC THOUGHT

Islāmic theology (*kalām*) and philosophy (*falsafah*) are two traditions of learning developed by Muslim thinkers who were engaged, on the one hand, in the rational clarification and defense of the principles of the Islāmic religion (*mutakallimūn*) and, on the other, in the pursuit of the ancient (Greek and Hellenistic, or Greco-Roman) sciences (*falāsifah*). These thinkers took a position that was intermediate between the traditionalists, who remained attached to the literal expressions of the primary sources of Islāmic doctrines (the Qurʾān, or the Islāmic scripture, and the Ḥadīth, or the sayings and traditions of Muḥammad) and who abhorred reasoning, and those whose reasoning led them to abandon the Islāmic community (the *ummah*) altogether. The status of the believer in Islām remained in practice a juridical question, not a matter for theologians or philosophers to decide. Except in regard to the fundamental questions of the existence of God, Islāmic revelation, and future reward and punishment, the juridical conditions for declaring someone an unbeliever or beyond the pale of Islām were so demanding as to make it almost impossible to make a valid declaration of this sort about a professing Muslim. In the course of events in Islāmic history, representatives of certain theological movements, who happened to be jurists and who succeeded in converting rulers to their cause, made those rulers declare in favour of their movements and even encouraged them to persecute their opponents. Thus there arose in some localities and periods a semblance of an official, or orthodox, doctrine.

Origins, nature, and significance of Islāmic theology

EARLY DEVELOPMENTS

Beginnings of Islāmic theology

The beginnings of theology in the Islāmic tradition in the second half of the 7th century are not easily distinguishable from the beginnings of a number of other disciplines—Arabic philology, Qurʾānic interpretation, the collection of the sayings and deeds of the prophet Muḥammad (Ḥadīth), jurisprudence, and historiography. Together with these other disciplines, Islāmic theology is concerned with ascertaining the facts and context of the Islāmic revelation and with understanding its meaning and implications as to what Muslims should believe and do after the revelation had ceased and the Islāmic community had to chart its own way. During the first half of the 8th century, a number of questions—which centred on God's unity, justice, and other attributes and which were relevant to man's freedom, actions, and fate in the hereafter—formed the core of a more specialized discipline, which was called *kalām* ("speech"). This term (*kalām*) was used to designate the more specialized discipline because of the rhetorical and dialectical "speech" used in formulating the principal matters of Islāmic belief, debating them, and defending them against Muslim and non-Muslim opponents. Gradually, *kalām* came to include all matters directly or indirectly relevant to the establishment and definition of religious beliefs, and it developed its own necessary or useful systematic rational arguments about human knowledge and the makeup of the world. Despite various efforts by later thinkers to fuse the problems of *kalām* with those of philosophy (and mysticism), theology preserved its relative independence from philosophy and other nonreligious sciences. It remained true to its original traditional and religious point of view, confined itself within the limits of the Islāmic revelation, and assumed that these limits as it understood them were identical with the limits of truth.

THE HELLENISTIC LEGACY

Relationships to other religious communities

The pre-Islāmic and non-Islāmic legacy with which early Islāmic theology came into contact included almost all the religious thought that had survived and was being defended or disputed in Egypt, Syria, Iran, and India. It was transmitted by learned representatives of various Christian, Jewish, Manichaean (members of a dualistic religion founded by Mani, an Iranian prophet, in the 3rd century), Zoroastrian (members of a monotheistic, but later dualistic, religion founded by Zoroaster, a 7th-century-BC Iranian prophet), Indian (Hindu and Buddhist, primarily), and Ṣābian (star worshippers of Harran often confused with the Mandaeans) communities and by early converts to Islām conversant with the teachings, sacred writings, and doctrinal history of the religions of these areas. At first, access to this legacy was primarily through conversations and disputations with such men, rather than through full and accurate translations of sacred texts or theological and philosophic writings, although some translations from Pahlavi (a Middle Persian dialect), Syriac, and Greek must also have been available.

The characteristic approach of early Islāmic theology to non-Muslim literature was through oral disputations, the starting points of which were the statements presented or defended (orally) by the opponents. Oral disputation continued to be used in theology for centuries, and most theological writings reproduce or imitate that form. From such oral and written disputations, writers on religions and sects collected much of their information about non-Muslim sects. Much of Hellenistic (post-3rd century BC Greek cultural), Iranian, and Indian religious thought was thus encountered in an informal and indirect manner.

From the 9th century onward, theologians had access to an increasingly larger body of translated texts, but by then they had taken most of their basic positions. They made a selective use of the translation literature, ignoring most of what was not useful to them until the mystical theologian al-Ghazālī (flourished 11th–12th centuries) showed them the way to study it, distinguish between the harmless and harmful doctrines contained in it, and refute the latter. By this time Islāmic theology had coined a vast number of technical terms, and theologians (*e.g.*, al-Jāḥiẓ) had forged Arabic into a versatile language of science; Arabic philology had matured; and the religious sciences (jurisprudence, the study of the Qur'ān, Ḥadīth, criticism, and history) had developed complex techniques of textual study and interpretation. The 9th-century translators availed themselves of these advances to meet the needs of patrons. Apart from demands for medical and mathematical works, the translation of Greek learning was fostered by the early ʿAbbāsid caliphs (8th–9th centuries) and their viziers as additional weapons (the primary weapon was theology itself) against the threat of Manichaeanism and other subversive ideas that went under the name *zandaqah* ("heresy" or "atheism"). (M.S.M./Ed.)

Theology and sectarianism

Despite the notion of a unified and consolidated community, as taught by the Prophet, serious differences arose within the Muslim community immediately after his death. According to the Sunnah, or traditionalist faction—who now constitute the majority of Islām—the Prophet had designated no successor. Thus the Muslims at Medina decided to elect a separate chief. Because he would not have been accepted by the Quraysh, the *ummah,* or Muslim community, would have disintegrated. Therefore, two of Muḥammad's fathers-in-law, who were highly respected early converts as well as trusted lieutenants, prevailed upon the Medinans to elect a single leader, and the choice fell upon Abū Bakr, father of the Prophet's favoured wife, ʿĀʾ ishah. All of this occurred before the Prophet's burial (under the floor of ʿĀʾishah's hut, alongside the courtyard of the mosque).

According to the Shīʿah, or "Partisans" of ʿAlī, the Prophet had designated as his successor his son-in-law ʿAlī ibn Abī Ṭālib, husband of his daughter Fāṭimah and father of his only surviving grandsons, Ḥasan and Ḥusayn. His preference was general knowledge; yet, while ʿAlī and the Prophet's closest kinsmen were preparing the body for burial, Abū Bakr, ʿUmar, and Abū ʿUbaydah from Muḥammad's Companions in the Quraysh tribe, met with the leaders of the Medinans and agreed to elect the aging Abū Bakr as the successor (*khalīfah,* hence "caliph") of the Prophet. ʿAlī and his kinsmen were dismayed but agreed for the sake of unity to accept the *fait accompli* because ʿAlī was still young

After the murder of ʿUthmān, the third caliph, ʿAlī was invited by the Muslims at Medina to accept the caliphate. Thus ʿAli became the fourth caliph (656–661), but the disagreement over his right of succession brought about a major schism in Islām, between the Shīʿah, or "legitimists"—those loyal to ʿAlī—and the Sunnah, or "traditionalists." Athough their differences were in the first instance political, arising out of the question of leadership, theological differences developed over time.

THE KHAWĀRIJ

During the reign of the third caliph, ʿUthmān, certain rebellious groups accused the Caliph of nepotism and misrule, and the resulting discontent led to his assassination. The rebels then recognized the Prophet's cousin and son-in-law, ʿAlī, as ruler but later deserted him and fought against him, accusing him of having committed a grave sin in submitting his claim to the caliphate to arbitration. The word *khāraju,* from which *khārijī* is derived, means "to withdraw" and Khawārij were, therefore, seceders who believed in active dissent or rebellion against a state of affairs they considered to be gravely impious.

The basic doctrine of the Khawārij was that a person or a group who committed a grave error or sin and did not sincerely repent ceased to be Muslim. Mere profession of the faith—"there is no god but God; Muḥammad is the prophet of God"—did not make a person a Muslim unless this faith was accompanied by righteous deeds. In other words, good works were an integral part of faith and not extraneous to it. The second principle that flowed from their aggressive idealism was militancy, or *jihād,* which the Khawārij considered to be among the cardinal principles, or pillars, of Islām. Contrary to the orthodox view, they interpreted the Qurʾānic command about "enjoining good and forbidding evil" to mean the vindication of truth through the sword. The placing of these two principles together made the Khawārij highly inflammable fanatics, intolerant of almost any established political authority. They incessantly resorted to rebellion and as a result were virtually wiped out during the first two centuries of Islām.

Because the Khawārij believed that the basis of rule was righteous character and piety alone, any Muslim, irrespective of race, colour, and sex, could, in their view, become ruler—provided he or she satisfied the conditions of piety. This was in contrast to the claims of the Shīʿah (the party of Muḥammad's son-in-law, ʿAlī) that the ruler must belong to the family of the Prophet and to the doctrine of the Sunnah (followers of the Prophet's way) that the head of state must belong to the Prophet's tribe, *i.e.,* the Quraysh.

A moderate group of the Khawārij, the Ibāḏīs, avoided extinction, and its members are to be found today in North Africa and in Oman and other parts of East Africa, including Zanzibar Island. The Ibāḏīs do not believe in aggressive methods and, throughout medieval Islām, remained dormant. Because of the interest of 20th-century Western scholars in this sect, the Ibāḏīs have become active and have begun to publish their classical writings and their own journals.

The permanent influence of the Khawārij

Although Khārijism is now essentially a story of the past, it has left a permanent influence on Islām, because of reaction against it. It forced the religious leadership of the community to formulate a bulwark against religious intolerance and fanaticism. Positively, it has influenced the reform movements that have sprung up in Islām from time to time and that have treated spiritual and moral placidity and status quo with a quasi-Khawārij zeal and militancy.

THE MUʿTAZILAH

The question of whether works are an integral part of faith or independent of it, as raised by the Khawārij, led to another important theological question: are human acts the result of a free human choice, or are they predetermined by God? This question brought with it a whole series of questions about the nature of God and of man. Although the initial impetus to theological thought, in the case of the Khawārij, had come from within Islām, full-scale religious speculation resulted from the contact and confrontation of Muslims with other cultures and systems of thought.

As a consequence of translations of Greek philosophical and scientific works into Arabic during the 8th and 9th centuries and the controversies of Muslims with Dualists (*e.g.,* Gnostics and Manichaeans), Buddhists, and Christians, a more powerful movement of rational theology emerged; its representatives are called the Muʿtazilah (literally "those who stand apart," a reference to the fact that they dissociated themselves from extreme views of faith and infidelity). On the question of the relationship of faith to works, the Muʿtazilah—who called themselves "champions of God's unity and justice"—taught, like the Khawārij, that works were an essential part of faith but that a person guilty of a grave sin, unless he repented, was neither a Muslim nor yet a non-Muslim but occupied a "middle ground." They further defended the position, as a central part of their doctrine, that man was free to choose and act and was, therefore, responsible for his actions. Divine predestination of human acts, they held, was incompatible with God's justice and human responsibility. The Muʿtazilah, therefore, recognized two powers, or actors, in the universe—God in the realm of nature and man in the domain of moral human action. The Muʿtazilah explained away the apparently predeterministic verses of the Qurʾān as being metaphors and exhortations.

Emphasis on reason

They claimed that human reason, independent of revelation, was capable of discovering what is good and what is evil, although revelation corroborated the findings of reason. Man is, therefore, under moral obligation to do the right even if there were no prophets and no divine revelation. Revelation has to be interpreted, therefore, in conformity with the dictates of rational ethics. Yet revelation is neither redundant nor passive. Its function is twofold. First, its aim is to aid man in choosing the right, because in the conflict between good and evil man often falters and makes the wrong choice against his rational judgment. God, therefore, must send prophets, for he must do the best for man; otherwise, the demands of divine grace and mercy cannot be fulfilled. Secondly, revelation is also necessary to communicate the positive obligations of religion—*e.g.,* prayers and fasting—which cannot be known without revelation.

God is viewed by the Muʿtazilah as pure Essence, without eternal attributes, because they hold that the assumption of eternal attributes in conjunction with Essence will result in a belief in multiple coeternals and violate the pure, unadulterated unity of God. God knows, wills, and acts by virtue of his Essence and not through attributes of knowledge, will, and power. Nor does he have an eternal attribute of speech, of which the Qurʾān and other earlier revelations were effects; the Qurʾān was, therefore, created in time and was not eternal.

The promises of reward that God has made in the Qurʾān to righteous people and the threats of punishment he has issued to evildoers must be carried out by him on the Day of Judgment. For promises and threats are viewed as reports about the future, and if not fulfilled exactly those reports will turn into lies, which are inconceivable of God. Also, if God were to withhold punishment for evil and forgive it, this would be as unjust as withholding reward for righteousness. There can be neither undeserved punishment nor undeserved reward; otherwise, good may just as well turn into evil and evil into good. From this position it follows that there can be no intercession on behalf of sinners.

When, in the early 9th century, the ʿAbbāsid caliph al-Maʾmūn raised Muʿtazilism to the status of the state creed, the Muʿtazilite rationalists showed themselves to be illiberal and persecuted their opponents. Aḥmad ibn Ḥanbal (died 855), an eminent orthodox figure and founder of one of the four orthodox schools of Islāmic law, was subjected to flogging and imprisonment for his refusal to subscribe to the doctrine that the Qurʾān, the word of God, was created in time.

THE SUNNAH

In the 10th century a reaction began against the Muʿtazilah that culminated in the formulation and subsequent general acceptance of another set of theological propositions, which became Sunnī, or "orthodox" theology.

The issues raised by these early schisms and the positions adopted by them enabled the Sunnī orthodoxy to define its own doctrinal positions in turn. Much of the content of Sunnī theology was, therefore, supplied by its reactions to those schisms. The term *sunnah,* which means a "well-trodden path" and in the religious terminology of Islām normally signifies "the example set by the Prophet," in the present context simply means the traditional and well-defined way. In this context, the term *sunnah* usually is accompanied by the appendage "the consolidated majority" (*al-jamāʿah*). The term clearly indicates that the traditional way is the way of the consolidated majority of the community as against peripheral or "wayward" positions of sectarians, who by definition must be erroneous.

The way of the majority. With the rise of the orthodoxy, then, the foremost and elemental factor that came to be emphasized was the notion of the majority of the community. The concept of the community so vigorously pronounced by the earliest doctrine of the Qurʾān gained both a new emphasis and a fresh context with the rise of Sunnism. Whereas the Qurʾān had marked out the Muslim community from other communities, Sunnism now emphasized the views and customs of the majority of the community in contradistinction to peripheral groups. An abundance of tradition (Ḥadīth) came to be attributed to the Prophet to the effect that Muslims must follow the majority's way, that minority groups are all doomed to hell, and that God's protective hand is always on (the majority of) the community, which can never be in error. Under the impact of the new Ḥadīth, the community, which had been charged by the Qurʾān with a mission and commanded to accept a challenge, now became transformed into a privileged one that was endowed with infallibility.

Tolerance of diversity. At the same time, while condemning schisms and branding dissent as heretical, Sunnism developed the opposite trend of accommodation, catholicity, and synthesis. A putative tradition of the Prophet that says "differences of opinion among my community are a blessing" was given wide currency. This principle of toleration ultimately made it possible for diverse sects and schools of thought—notwithstanding a wide range of difference in belief and practice—to recognize and coexist with each other. No group may be excluded from the community unless it itself formally renounces Islām. As for individuals, tests of heresy may be applied to their beliefs, but, unless a person is found to flagrantly violate or deny the unity of God or expressly negate the prophethood of Muḥammad, such tests usually have no serious consequences. Catholicity was orthodoxy's answer to the intolerance and secessionism of the Khawārij and the severity of the Muʿtazilah. As a consequence, a formula was adopted in which good works were recognized as enhancing the quality of faith but not as entering into the definition and essential nature of faith. This broad formula saved the integrity of the community at the expense of moral strictness and doctrinal uniformity.

On the question of free will, Sunnī orthodoxy attempted a synthesis between man's responsibility and God's omnipotence. The champions of orthodoxy accused the Muʿtazilah of quasi-Magian Dualism (Zoroastrianism) insofar as the Muʿtazilah admitted two independent and original actors in the universe: God and man. To the orthodox it seemed blasphemous to hold that man could act wholly outside the sphere of divine omnipotence, which had been so vividly portrayed by the Qurʾān but which the Muʿtazilah had endeavoured to explain away in order to make room for man's free and independent action.

Formulators of Sunnī doctrine

Influence of Al-Ashʿarī and al-Māturīdī. The Sunnī formulation, however, as presented by al-Ashʿarī and al-Māturīdī, Sunnī's two main representatives in the 10th century, shows palpable differences despite basic uniformity. Al-Ashʿarī taught that human acts were created by God and acquired by man and that human responsibility depended on this acquisition. He denied, however, that man could be described as an actor in a real sense. Al-Māturīdī, on the other hand, held that although God is the sole Creator of everything, including human acts, nevertheless, man is an actor in the real sense, for acting and creating were two different types of activity involving different aspects of the same human act.

In conformity with their positions, al-Ashʿarī believed that man did not have the power to act before he actually acted and that God created this power in him at the time of action; and al-Māturīdī taught that before the action man has a certain general power for action but that this power becomes specific to a particular action only when the action is performed, because, after full and specific power comes into existence, action cannot be delayed.

Al-Ashʿarī and his school also held that human reason was incapable of discovering good and evil and that acts became endowed with good or evil qualities through God's declaring them to be such. Because man in his natural state regards his own self-interest as good and that which thwarts his interests as bad, natural human reason is unreliable. Independently of revelation, therefore, murder would not be bad nor the saving of life good. Furthermore, because God's Will makes acts good or bad, one cannot ask for reasons behind the divine law, which must be simply accepted. Al-Māturīdī takes an opposite position, not materially different from that of the Muʿtazilah: human reason is capable of finding out good and evil, and revelation aids human reason against the sway of human passions.

Despite these important initial differences between the two main Sunnī schools of thought, the doctrines of al-Māturīdī became submerged in course of time under the expanding popularity of the Ashʿarite school, which gained wide currency particularly after the 11th century because of the influential activity of the Ṣūfī theologian al-Ghazālī. Because these later theologians placed increasing emphasis on divine omnipotence at the expense of the freedom and efficacy of the human will, a deterministic outlook on life became characteristic of Sunnī Islām—reinvigorated by the Ṣūfī world view, which taught that nothing exists except God, whose being is the only real being. This general deterministic outlook produced, in turn, a severe reformist reaction in the teachings of Ibn Taymīyah, a 14th-century theologian who sought to rehabilitate human freedom and responsibility and whose influence has been strongly felt through the reform movements in the Muslim world since the 18th century.

THE SHĪʿAH

The Shīʿah are the only important surviving sect in Islām. As noted above, they owe their origin to the hostility between ʿAlī (the fourth caliph and son-in-law of the Prophet) and the Umayyad dynasty (661–750). After ʿAlī's death, the Shīʿah (Party; *i.e.,* of ʿAlī) demanded the restoration of rule to ʿAlī's family, and from that demand developed the Shīʿite legitimism, or the divine right of the holy family to rule. In the early stages, the Shīʿah used this legitimism to cover the protest against the Arab hegemony under the Umayyads and to agitate for social reform.

Emphasis on the *imām*

Gradually, however, Shīʿism developed a theological content for its political stand. Probably under Gnostic (esoteric, dualistic, and speculative) and old Iranian (dualistic) influences, the figure of the political ruler, the *imām* (exemplary "leader"), was transformed into a metaphysical being, a manifestation of God and the primordial light that sustains the universe and bestows true knowledge on man. Through the *imām* alone the hidden and true meaning of the Qurʾānic revelation can be known, because the *imām* alone is infallible. The Shīʿah thus developed a doctrine of esoteric knowledge that was adopted also, in a modified form, by the Ṣūfīs, or Islāmic mystics (see below *Islāmic mysticism, Ṣūfism*). The orthodox Shīʿah recognize 12 such *imāms*, the last (Muḥammad) having disappeared in the 9th century. Since that time, the *mujtahids* (*i.e.,* the Shīʿī divines) have been able to interpret law and doctrine under the putative guidance of the *imām,* who will return toward the end of time to fill the world with truth and justice.

On the basis of their doctrine of imamology, the Shīʿah emphasize their idealism and transcendentalism in conscious contrast with Sunnī pragmatism. Thus, whereas the Sunnīs believe in the *ijmāʿ* ("consensus") of the community as the source of decision making and workable knowledge, the Shīʿah believe that knowledge derived from fallible sources is useless and that sure and true knowledge can come only through a contact with the infallible *imām.* Again, in marked contrast to Sunnism, Shīʿism adopted the Muʿtazilite doctrine of the freedom of the human will and the capacity of human reason to know good and evil, although its position on the question of the relationship of faith to works is the same as that of the Sunnīs.

Parallel to the doctrine of an esoteric knowledge, Shīʿism, because of its early defeats and persecutions, also adopted the principle of *taqīyah,* or dissimulation of faith in a hostile environment. Introduced first as a practical principle, *taqīyah,* which is also attributed to ʿAlī and other *imāms,* became an important part of the Shīʿah religious teaching and practice. In the sphere of law, Shīʿism differs from Sunnī law mainly in allowing a temporary marriage, called *mutʿah,* which can be legally contracted for a fixed period of time on the stipulation of a fixed dower.

Introduction of the passion motive

From a spiritual point of view, perhaps the greatest difference between Shīʿism and Sunnism is the former's introduction into Islām of the passion motive, which is conspicuously absent from Sunnī Islām. The violent death (in 680) of ʿAlī's son, Ḥusayn, at the hands of the Umayyad troops is celebrated with moving orations, passion plays, and processions in which the participants, in a state of emotional frenzy, beat their breasts with heavy chains and sharp instruments, inflicting wounds on their bodies. This passion motive has also influenced the Sunnī masses in Afghanistan and the Indian subcontinent, who participate in passion plays called *taʿziyahs.* Such celebrations are, however, absent from Egypt and North Africa.

Although the Shīʿah number only about 40,000,000 (Shīʿism has been the official religion in Iran since the 16th century), Shīʿism has exerted a great influence on Sunnī Islām in several ways. The veneration in which all Muslims hold ʿAlī and his family and the respect shown to ʿAlī's descendants (who are called *sayyids* in the East and *sharīfs* in North Africa) are obvious evidence of this influence.

Ismāʿīlīs. Besides the main body of Twelver (Ithnā ʿAsharīyah) Shīʿah, Shīʿism has produced a variety of more or less extremist sects, the most important of them being the Ismāʿīlī. Instead of recognizing Mūsā as the seventh *imām,* as did the main body of the Shīʿah, the Ismāʿīlīs upheld the

Muslims carrying the *taʿzīyah* to their cremation during a procession commemorating the martyrdom of Ḥusayn, in Jaipur, India.

Foto Features

claims of his elder brother Ismāʿīl. One group of Ismāʿīlīs, called Seveners (Sabʿīyah), considered Ismāʿīl the seventh and last of the *imām*s. The majority of Ismāʿīlīs, however, believed that the imamate continued in the line of Ismāʿīl's descendants. The Ismāʿīlī teaching spread during the 9th century from North Africa to Sind, in India, and the Ismāʿīlī Fāṭimid dynasty succeeded in establishing a prosperous empire in Egypt. Ismāʿīlīs are subdivided into two groups—the Nizārīs, headed by the Aga Khan, and the Mustaʿlīs in Bombay, with their own spiritual head. The Ismāʿīlīs are to be found mainly in East Africa, Pakistan, India, and Yemen.

In their theology, the Ismāʿīlīs have absorbed the most extreme elements and heterodox ideas. The universe is viewed as a cyclic process, and the unfolding of each cycle is marked by the advent of seven "speakers"—messengers of God with Scriptures—each of whom is succeeded by seven "silents"—messengers without revealed scriptures; the last speaker (the Prophet Muḥammad) is followed by seven *imām*s who interpret the Will of God to man and are, in a sense, higher than the Prophet because they draw their knowledge directly from God and not from the Angel of Revelation. During the 10th century, certain Ismāʿīlī intellectuals formed a secret society called the Brethren of Purity, which issued a philosophical encyclopaedia, *The Epistles of the Brethren of Purity,* aiming at the liquidation of positive religions in favour of a universalist spirituality.

The late Aga Khan III (1887–1957) had taken several measures to bring his followers closer to the main body of the Muslims. The Ismāʿīlīs, however, still do not have mosques but *jamāʿat khānah*s ("gathering houses"), and their mode of worship bears little resemblance to that of the Muslims generally.

Other Shīʿī sects. Several other sects arose out of the general Shīʿite movement—*e.g.,* the Nuṣayrīs, the Yazīdīs, and the Druzes—out of which only the Druzes have any considerable following. The Druze sect, sometimes considered as independent from Islām, arose in the 11th century out of a cult of deification of the Fāṭimid caliph al-Ḥākim. Some authorities believe that the growth of the Freemasonry movement was influenced by Druze rituals.

Druzes and Bahāʾīs

During a 19th-century anticlerical movement in Iran, a certain ʿAlī Moḥammad of Shīrāz appeared, declaring himself to be the Bāb ("Gate"; *i.e.,* to God). At that time the climate in Iran was generally favourable to Messianic ideas. He was, however, bitterly opposed by the Shīʿah *ʿulamāʾ* (council of learned men) and was executed in 1850. After his death, his two disciples, Ṣobḥ-e Azal and Bahāʾ Ullāh, broke and went in different directions. Bahāʾ Ullāh eventually declared his religion—stressing a humanitarian pacificism and universalism—to be an independent religion outside Islām. The Bahāʾī faith won a considerable number of converts in North America during the early 20th century (see also in the *Micropædia:* DRUZE; BAHĀʾĪ FAITH).

Islāmic mysticism, or Ṣūfism, emerged out of early ascetic reactions on the part of certain religiously sensitive personalities against the general worldliness that had overtaken the Muslim community and the purely "externalist" expressions of Islām in law and theology. These persons stressed the Muslim qualities of moral motivation, contrition against overworldliness, and "the state of the heart" as opposed to the legalist formulations of Islām. For a complete exposition of Ṣūfī history, beliefs, and practices, see below *Islāmic mysticism, Ṣūfism.*

OTHER GROUPS

The Aḥmadīyah. In the latter half of the 19th century in Punjab, India, Mirza Ghulam Ahmad claimed to be an inspired prophet. At first a defender of Islām against Christian missionaries, he adopted certain doctrines of the Indian Muslim modernist Sayyid Ahmad Khan—namely, that Jesus died a natural death and was not assumed into heaven as the Islāmic orthodoxy believed and that *jihād* "by the sword" had been abrogated and replaced with *jihād* "of the pen." His aim appears to have been to synthesize all religions under Islām, for he declared himself to be not only the manifestation of the Prophet Muḥammad but also the Second Advent of Jesus, as well as Ḳrishna for the Hindus, among other claims. He did not announce, however, any new revelation or new law.

In 1914 a schism over succession occurred among the Aḥmadīyah. One group that seceded from the main body, which was headed by a son of the founder, disowned the prophetic claims of Ghulam Ahmad and established their centre in Lahore (in modern Pakistan). The main body of the Aḥamadīyah evolved a separatist organization and, after the partition of India in 1947, moved their headquarters to Rabwah in what was then West Pakistan.

Both groups are noted for their missionary work, particularly in the West and in Africa. Within the Muslim countries, however, there is fierce opposition to the main group because of its claim that Ghulam Ahmad was a prophet (the Muslim community believes in the finality of prophethood with Muḥammad) and because of its separatist organization. Outside the Muslim countries, however, the Qadiani group (as the main body is called, Qadian being the birthplace of the founder and first centre of the sect) acts more like a movement than a sect, with a relatively loose connection with its centre in Pakistan.

The "Black Muslims." After World War II an Islāmic movement arose among blacks in the U.S.; members called themselves the Nation of Islam, but they were popularly known as Black Muslims. Although they adopted some Islāmic social practices, the group was in large part a black separatist and social-protest movement. Their leader, Elijah Muhammad, who claimed to be an inspired prophet, interpreted the doctrine of Resurrection in an unorthodox sense as the revival of oppressed ("dead") peoples. The popular leader and spokesman Malcolm X (el-Hajj Malik el-Shabazz) broke with Elijah Muhammad and adopted more orthodox Islāmic views. He was assassinated in 1965. After the death of Elijah Muhammad in 1975, the group was renamed World Community of Islam in the West and officially abandoned its separatist aims. The name was again changed in the late 1970s, to American Muslim Mission. (F.R./Ed.)

Islāmic mysticism, Ṣūfism

Mysticism is that aspect of Islāmic belief and practice in which Muslims seek to find the truth of divine love and knowledge through direct personal experience of God. It consists of a variety of mystical paths that are designed to ascertain the nature of man and God and to facilitate the experience of the presence of divine love and wisdom in the world.

Islāmic mysticism is called *taṣawwuf* (literally, "to dress in wool") in Arabic, but it has been called Ṣūfism in Western languages since the early 19th century. An abstract word, Ṣūfism derives from the Arabic term for a mystic, *ṣūfī,* which is in turn derived from *ṣūf,* "wool," plausibly a reference to the woollen garment of early Islāmic ascetics. The Ṣūfīs are also generally known as "the poor," *fuqarāʾ*, plural of the Arabic *faqīr*, in Persian *darvīsh*, whence the English words fakir and dervish.

Origins and influence

Though the roots of Islāmic mysticism formerly were supposed to have stemmed from various non-Islāmic sources in ancient Europe and even India, it now seems established that the movement grew out of early Islāmic asceticism that developed as a counterweight to the increasing worldliness of the expanding Muslim community; only later were foreign elements that were compatible with mystical theology and practices adopted and made to conform to Islām.

By educating the masses and deepening the spiritual concerns of the Muslims, Ṣūfism has played an important role in the formation of Muslim society. Opposed to the dry casuistry of the lawyer-divines, the mystics nevertheless scrupulously observed the commands of the divine law. The Ṣūfīs have been further responsible for a large-scale missionary activity all over the world, which still continues. Ṣūfīs have elaborated the image of the prophet Muḥammad—the founder of Islām—and have thus largely influenced Muslim piety by their Muḥammad-mysticism. Without the Ṣūfī vocabulary, Persian and other literatures related to it, such as Turkish, Urdu, Sindhi, Pashto, and Panjabi, would lack their special charms. Through the poetry of these literatures mystical ideas spread widely among the Muslims. In some countries Ṣūfī leaders were also active politically.

HISTORY

Islāmic mysticism had several stages of growth, including (1) the appearance of early asceticism, (2) the development of a classical mysticism of divine love, and (3) the rise and proliferation of fraternal orders of mystics. Despite these general stages, however, the history of Islāmic mysticism is largely a history of individual mystic experience.

The first stage of Ṣūfism appeared in pious circles as a reaction against the worldliness of the early Umayyad period (AD 661–749). From their practice of constantly meditating on the Qurʾānic words about Doomsday, the ascetics became known as "those who always weep" and those who considered this world "a hut of sorrows." They were distinguished by their scrupulous fulfillment of the injunctions of the Qurʾān and tradition, by many acts of piety, and especially by a predilection for night prayers.

Classical mysticism. The introduction of the element of love, which changed asceticism into mysticism, is ascribed to Rābiʿah al-ʿAdawīyah (died 801), a woman from Basra who first formulated the Ṣūfī ideal of a love of God that was disinterested, without hope for paradise and without fear of hell. In the decades after Rābiʿah, mystical trends grew everywhere in the Islāmic world, partly through an exchange of ideas with Christian hermits. A number of mystics in the early generations had concentrated their efforts upon *tawakkul*, absolute trust in God, which became a central concept of Ṣūfism. An Iraqi school of mysticism became noted for its strict self-control and psychological insight. The Iraqi school was initiated by al-Muḥāsibī (died 857)—who believed that purging the soul in preparation for companionship with God was the only value of asceticism. Its teachings of classical sobriety and wisdom were perfected by Junayd of Baghdad (died 910), to whom all later chains of the transmission of doctrine and legitimacy go back. In an Egyptian school of Ṣūfism, the Nubian Dhū an-Nūn (died 859) reputedly introduced the technical term *maʿ rifah* ("interior knowledge"), as contrasted to learnedness; in his hymnical prayers he joined all nature in the praise of God—an idea based on the Qurʾān and later elaborated in Persian and Turkish poetry. In the Iranian school, Abū Yazīd al-Bisṭāmī (died 874) is usually considered to have been representative of the important doctrine of annihilation of the self, *fanāʾ* (see below); the strange symbolism of his sayings prefigures part of the terminology of later mystical poets. At the same time the concept of divine love became more central, especially among the Iraqi Ṣūfīs. Its main representatives are Nūrī, who offered his life for his brethren, and Sumnūn "the Lover."

Mystical love

The first of the theosophical speculations based on mystical insights about the nature of man and the essence of the Prophet were produced by such Ṣūfīs as Sahl at-Tustarī (died *c.* 896). Some Hellenistic ideas were later adopted by al-Ḥakīm at-Tirmidhī (died 898). Sahl was the master of al-Ḥusayn ibn Manṣūr al-Ḥallāj, who has become famous for his phrase *anā al-ḥaqq*, "I am the Creative Truth" (often rendered "I am God"), which was later interpreted in a pantheistic sense but is, in fact, only a condensation of his theory of *huwa huwa* ("He he"): God loved himself in his essence, and created Adam "in his image." Ḥallāj was executed in 922 in Baghdad as a result of his teachings; he is, for later mystics and poets, the "martyr of Love" par excellence, the enthusiast killed by the theologians. His few poems are of exquisite beauty; his prose, which contains an outspoken Muḥammad-mysticism—*i.e.*, mysticism centred on the prophet Muḥammad—is as beautiful as it is difficult.

Ṣūfī thought was in these early centuries transmitted in small circles. Some of the *shaykhs*, Ṣūfī mystical leaders or guides of such circles, were also artisans. In the 10th century, it was deemed necessary to write handbooks about the tenets of Ṣūfism in order to soothe the growing suspicions of the orthodox; the compendiums composed in Arabic by Abū Ṭālib Makkī, Sarrāj, and Kalābādhī in the late 10th century, and by Qushayrī and, in Persian, by Hujvīrī in the 11th century reveal how these authors tried to defend Ṣūfism and to prove its orthodox character. It should be noted that the mystics belonged to all schools of Islāmic law and theology of the times.

The last great figure in the line of classical Ṣūfism is Abū Ḥāmid al-Ghazālī (died 1111), who wrote, among numerous other works, the *Iḥyāʾ ʿulūm ad-dīn* ("The Revival of the Religious Sciences"), a comprehensive work that established moderate mysticism against the growing theosophical trends—which tended to equate God and the world—and thus shaped the thought of millions of Muslims. His younger brother, Aḥmad al-Ghazālī, wrote one of the subtlest treatises (*Sawāniḥ*; "Occurrences" [*i.e.*, stray thoughts]) on mystical love, a subject that then became the main subject of Persian poetry.

Rise of fraternal orders. Slightly later, mystical orders (fraternal groups centring around the teachings of a leader-founder) began to crystallize. The 13th century, though politically overshadowed by the invasion of the Mongols into the Eastern lands of Islām and the end of the ʿAbbāsid caliphate, was also the golden age of Ṣūfism: the Spanish-born Ibn alʿArabī created a comprehensive theosophical system (concerning the relation of God and the world) that was to become the cornerstone for a theory of "Unity of Being." According to this theory all existence is one, a manifestation of the underlying divine reality. His Egyptian contemporary Ibn al-Fāriḍ wrote the finest mystical poems in Arabic. Two other important mystics, who died *c.* AD 1220, were a Persian poet, Farīd od-Dīn ʿAṭṭār, one of the most fertile writers on mystical topics, and a Central Asian master, Najmuddīn Kubrā, who presented elaborate discussions of the psychological experiences through which the mystic adept has to pass.

The mystical poetry of Jalāl ad-Dīn ar-Rūmī

The greatest mystical poet in the Persian language, Jalāl ad-Dīn ar-Rūmī (1207–73), was moved by mystical love to compose his lyrical poetry that he attributed to his mystical beloved, Shams ad-Dīn of Tabriz, as a symbol of their union. Rūmī's didactic poem *Mas̄navī* in about 26,000 couplets—a work that is for the Persian-reading mystics second in importance only to the Qurʾān—is an encyclopaedia of mystical thought in which everyone can find his own religious ideas. Rūmī inspired the organization of the whirling dervishes—who sought ecstasy through an elaborate dancing ritual, accompanied by superb music. His younger contemporary Yunus Emre inaugurated Turkish mystical poetry with his charming verses that were transmitted by the Bektāshīyah (Bektaşi) order of dervishes and are still admired in modern Turkey. In Egypt, among many other mystical trends, an order—known as Shādhilīyah—was founded by ash-Shādhilī (died 1258); its main literary representative, Ibn ʿAṭāʾ Allāh of Alexandria, wrote sober aphorisms (*ḥikam*).

At that time, the basic ideals of Ṣūfism permeated the whole world of Islām; and at its borders as, for example, in

India, Ṣūfīs largely contributed to shaping Islāmic society. Later some of the Ṣūfīs in India were brought closer to Hindu mysticism by an overemphasis on the idea of divine unity which became almost monism—a religiophilosophic perspective according to which there is only one basic reality, and the distinction between God and the world (and man) tends to disappear. The syncretistic attempts of the Mughal emperor Akbar (died 1605) to combine different forms of belief and practice, and the religious discussions of the crown prince Dārā Shukōh (executed for heresy, 1659) were objectionable to the orthodox. Typically, the countermovement was again undertaken by a mystical order, the Naqshbandīyah, a Central Asian fraternity founded in the 14th century. Contrary to the monistic trends of the school of *waḥdat al-wujūd* ("existential unity of being"), the later Naqshbandīyah defended the *waḥdat ash-shuhūd* ("unity of vision"), a subjective experience of unity, occurring only in the mind of the believer, and not as an objective experience. Aḥmad Sirhindī (died 1624) was the major protagonist of this movement in India. His claims of sanctity were surprisingly daring: he considered himself the divinely invested master of the universe. His refusal to concede the possibility of union between man and God (characterized as "servant" and "Lord") and his sober law-bound attitude gained him and his followers many disciples, even at the Mughal court and as far away as Turkey. In the 18th century, Shāh Walī Allāh of Delhi was connected with an attempt to reach a compromise between the two inimical schools of mysticism; he was also politically active and translated the Qur'ān into Persian, the official language of Mughal India. Other Indian mystics of the 18th century, such as Mīr Dard, played a decisive role in forming the newly developing Urdu poetry.

Trends in modern Ṣūfism

In the Arabic parts of the Islāmic world, only a few interesting mystical authors are found after 1500. They include ash-Shaʿrānī in Egypt (died 1565) and the prolific writer ʿAbd al-Ghanī an-Nābulusī in Syria (died 1731). Turkey produced some fine mystical poets in the 17th and 18th centuries. The influence of the mystical orders did not recede; rather new orders came into existence, and most literature was still tinged with mystical ideas and expressions. Political and social reformers in the Islāmic countries have often objected to Ṣūfism because they have generally considered it as backward, hampering the free development of society. Thus, the orders and dervish lodges in Turkey were closed by Kemal Atatürk in 1925. Yet, their political influence is still palpable, though under the surface. Such modern Islāmic thinkers as the Indian philosopher Muḥammad Iqbāl have attacked traditional monist mysticism and have gone back to the classical ideals or divine love as expressed by Ḥallāj and his contemporaries. The activities of modern Muslim mystics in the cities are mostly restricted to spiritual education.

SUFI LITERATURE

Though a prophetic saying (Ḥadīth) claims that "he who knows God becomes silent," the Ṣūfīs have produced a literature of impressive extent and could defend their writing activities with another Ḥadīth: "He who knows God talks much." The first systematic books explaining the tenets of Ṣūfism date from the 10th century; but earlier, Muḥāsibī had already written about spiritual education, Ḥallāj had composed meditations in highly concentrated language, and many Ṣūfīs had used poetry for conveying their experiences of the ineffable mystery or had instructed their disciples in letters of cryptographic density. The accounts of Ṣūfism by Sarrāj and his followers, as well as the *ṭabaqāt* (biographical works) by Sulamī, Abū Nuʿaym al-Iṣfahānī, and others, together with some biographies of individual masters, are the sources for knowledge of early Ṣūfism.

Early mystical commentaries on the Qur'ān are only partly extant, often preserved in fragmentary quotations in later sources. With the formation of mystical orders, books about the behaviour of the Ṣūfī in various situations became important, although this topic had already been touched on in such classical works as *Adāb al-murīdīn* ("The Adepts' Etiquette") by Abū Najīb as-Suhrawardī (died 1168), the founder of the Suhrawardīyah order and uncle of the author of the oft translated *ʿAwārif al-maʿārif* ("The Well-known Sorts of Knowledge"). The theosophists had to condense their systems in readable form; Ibn al-ʿArabī's *al-Futūḥāt al-Makkīyah* ("The Meccan Revelations") is the textbook of *waḥdat al-wujūd* (God and creation as two aspects of one reality); his smaller work on the peculiar character of the prophets—*Fuṣūṣ al-ḥikam* ("The Bezels—or cutting edges—of Wisdom")—became even more popular.

Later mystics commented extensively upon the classical sources and, sometimes, translated them into their mother tongues. A literary type that has flourished especially in India since the 13th century is the *malfūẓāt,* a collection of sayings of the mystical leader, which are psychologically interesting and allow glimpses into the political and social situation of the Muslim community. Collections of letters of the *shaykh*s are similarly revealing. Ṣūfī literature abounds in hagiography, either biographies of all known saints from the Prophet to the day of the author, or of saints of a specific order, or of those who lived in a certain town or province, so that much information on the development of Ṣūfī thought and practice is available if sources are critically sifted.

Poetical, national, and regional literature

The greatest contribution of Ṣūfism to Islāmic literature, however, is poetry—beginning with charming, short Arabic love poems (sometimes sung for a mystical concert, *samāʿ*) that express the yearning of the soul for union with the beloved. The love-relation prevailing in most Persian poetry is that between a man and a beautiful youth; less often, as in the writings of Ibn al-ʿArabī and Ibn al-Fāriḍ, eternal beauty is symbolized through female beauty; in Indo-Muslim popular mystical songs the soul is the loving wife, God the longed-for husband. Long mystic–didactic poems (*mas̄navīs*) were written to introduce the reader to the problems of unity and love by means of allegories and parables. After Sanāʾī's (died 1131?) *Ḥadīqat al-ḥaqīqah wa sharīʿat aṭ-ṭariqah* ("The Garden of Truth and the Law of Practice"),came ʿAṭṭar's *Manṭeq oṭ-ṭeyr* ("The Birds' Conversation") and Rūmī's *Mas̄navī-ye maʿnavī* ("Spiritual Couplets"). These three works are the sources that have furnished poets for centuries with mystical ideas and images. Typical of Ṣūfī poetry is the hymn in praise of God, expressed in chains of repetitions.

The mystics also contributed largely to the development of national and regional literatures, for they had to convey their message to the masses in their own languages: in Turkey as well as in the Panjabi-, the Sindhi-, and the Urdu-speaking areas of South Asia, the first true religious poetry was written by Ṣūfīs, who blended classical Islāmic motifs with inherited popular legends and used popular rather than Persian metres. Ṣūfī poetry expressing divine love and mystical union through the metaphors of profane love and union often resembled ordinary worldly love poetry; and nonmystical poetry made use of the Ṣūfī vocabulary, thus producing an ambiguity that is felt to be one of the most attractive and characteristic features of Persian, Turkish, and Urdu literatures. Ṣūfī ideas thus permeated the hearts of all those who hearkened to poetry. An example is al-Ḥusayn ibn Manṣūr al-Ḥallāj, the 10th-century martyr–mystic, who is as popular in modern progressive Urdu poetry as he was with the "God-intoxicated" Ṣūfīs; he has been converted into a symbol of suffering for one's ideals.

SUFI THOUGHT AND PRACTICE

Important aspects. The mystics drew their vocabulary largely from the Qur'ān, which for Muslims contains all divine wisdom and has to be interpreted with ever-increasing insight. In the Qur'ān, mystics found the threat of the Last Judgment, but they also found the statement that God "loves them and they love him," which became the basis for love-mysticism. Strict obedience to the religious law and imitation of the Prophet were basic for the mystics. By rigid introspection and mental struggle the mystic tried to purify his baser self from even the smallest signs of selfishness, thus attaining *ikhlāṣ,* absolute purity of intention and act. *Tawakkul* (trust in God) was sometimes practiced to such an extent that every thought of tomorrow was considered irreligious. "Little sleep, little talk, little food" were fundamental; fasting became one of the most important preparations for the spiritual life.

Central concern of Ṣūfism

The central concern of the Ṣūfīs, as of every Muslim, was *tawḥīd,* the witness that "There is no deity but God." This truth had to be realized in the existence of each individual, and so the expressions differ: early Ṣūfism postulated the approach to God through love and voluntary suffering until a unity of will was reached; Junayd spoke of "recognizing God as He was before creation"; God is seen as the One and only actor; He alone "has the right to say 'I'." Later, *tawḥīd* came to mean the knowledge that there is nothing existent but God, or the ability to see God and creation as two aspects of one reality, reflecting each other and depending upon each other (*waḥdat al-wujūd*).

The mystics realized that beyond the knowledge of outward sciences intuitive knowledge was required in order to receive that illumination to which reason has no access. *Dhawq,* direct "tasting" of experience, was essential for them. But the inspirations and "unveilings" that God grants such mystics by special grace must never contradict the Qurʾān and tradition and are valid only for the person concerned. Even the Malāmatīs, who attracted public contempt upon themselves by outwardly acting against the law, in private life strictly followed the divine commands. Mystics who expressed in their poetry their disinterest in, and even contempt of, the traditional formal religions never forgot that Islām is the highest manifestation of divine wisdom.

The idea of the manifestation of divine wisdom was also connected with the person of the prophet Muḥammad. Though early Ṣūfism had concentrated upon the relation between God and the soul, from AD 900 onward a strong Muḥammad-mysticism developed. In the very early years, the alleged divine address to the Prophet—"If thou hadst not been I had not created the worlds"—was common among Ṣūfīs. Muḥammad was said to be "Prophet when Adam was still between water and clay." Muḥammad is also described as light from light, and from his light all the prophets are created, constituting the different aspects of this light. In its fullness such light radiated from the historical Muḥammad and is partaken of by his posterity and by the saints; for Muḥammad has the aspect of sanctity in addition to that of prophecy. An apocryphal tradition makes even God attest: "I am Aḥmad (= Muḥammad) without 'm' (*i.e.,* Aḥad, 'One')."

The *walī,* or saint

A mystic may also be known as *walī.* By derivation the word *walī* ("saint") means "one in close relation; friend." The *awlīyāʾ* (plural of *walī*) are "friends of God who have no fear nor are they sad." Later the term *walī* came to denote the Muslim mystics who had reached a certain stage of proximity to God, or those who had reached the highest mystical stages. They have their "seal" (*i.e.,* the last and most perfect personality in the historical process; with this person, the evolution has found its end—as in Muḥammad's case), just as the prophets have. Woman saints are found all over the Islāmic world.

The invisible hierarchy of saints consists of the 40 *abdāl* ("substitutes"; for when any of them dies another is elected by God from the rank and file of the saints), seven *awtād* ("stakes," or "props," of faith), three *nuqabāʾ* ("leader"; "one who introduces people to his master"), headed by the *quṭb* ("axis, pole"), or *ghawth* ("help")—titles claimed by many Ṣūfī leaders. Saint worship is contrary to Islām, which does not admit of any mediating role for human beings between man and God; but the cult of living and even more of dead saints—visiting their tombs to take vows there—responded to the feeling of the masses, and thus a number of pre-Islāmic customs were absorbed into Islām under the cover of mysticism. The advanced mystic was often granted the capacity of working miracles called *karāmāt* (*charismata* or "graces"); not *muʿjizāt* ("that which men are unable to imitate"), like the miracles of the prophets. Among them are "cardiognosia" (knowledge of the heart), providing food from the unseen, presence in two places at the same time, and help for the disciples, be they near or far. In short, a saint is one "whose prayers are heard" and who has *taṣarruf,* the power of materializing in this world possibilities that still rest in the spiritual world. Many great saints, however, considered miracle working as a dangerous trap on the path that might distract the Ṣūfī from his real goal.

The path. The path (*ṭarīqah*) begins with repentance. A mystical guide (*shaykh, pīr*) accepts the seeker as disciple (*murīd*), orders him to follow strict ascetic practices, and suggests certain formulas for meditation. It is said that the disciple should be in the hands of the master "like a corpse in the hand of the washer." The master teaches him constant struggle (the real "Holy War") against the lower soul, often represented as a black dog, which should, however, not be killed but merely tamed and used in the way of God. The mystic dwells in a number of spiritual stations (*maqām*), which are described in varying sequence, and, after the initial repentance, comprise abstinence, renunciation, and poverty—according to Muḥammad's saying, "Poverty is my pride"; poverty was sometimes interpreted as having no interest in anything apart from God, the Rich One, but the concrete meaning of poverty prevailed, which is why the mystic is often denoted as "poor," fakir or dervish. Patience and gratitude belong to higher stations of the path, and consent is the loving acceptance of every affliction.

On his way to illumination the mystic will undergo such changing spiritual states (*ḥāl*) as *qabḍ* and *basṭ,* constraint and happy spiritual expansion, fear and hope, and longing and intimacy, which are granted by God and last for longer or shorter periods of time, changing in intensity according to the station in which the mystic is abiding at the moment. The way culminates in *maʿrifah* ("interior knowledge," "gnosis") or in *maḥabbah* ("love"), the central subject of Ṣūfism since the 9th century, which implies a union of lover and beloved, and was therefore violently rejected by the orthodox, for whom "love of God" meant simply obedience. The final goal is *fanāʾ* ("annihilation"), primarily an ethical concept of annihilating one's own qualities, according to the prophetic saying "Take over the qualities of God," but slowly developing into a complete extinction of the personality. Some mystics taught that behind this negative unity where the self is completely effaced, the *baqāʾ,* ("duration, life in God") is found: the ecstatic experience, called intoxication, is followed by the "second sobriety"; *i.e.,* the return of the completely transformed mystic into this world where he acts as a living witness of God or continues the "journey in God." The mystic has reached *ḥaqīqah* ("realty"), after finishing the *ṭarīqah* ("path"), which is built upon the *sharīʿah* ("law"). Later, the disciple is led through *fanāʾ fī ashshaykh* ("annihilation in the master") to *fanāʾ fiar-Rasūl* ("annihilation in the Prophet") before reaching, if at all, *fanāʾ fī-Allāh* ("annihilation in God").

The practice of *dhikr*

One of the means used on the path is the ritual prayer, or *dhikr* ("remembrance"), derived from the Qurʾānic injunction "And remember God often" (*sūrah* 62:10). It consists in a repetition of either one or all of the most beautiful names of God, of the name "Allāh," or of a certain religious formula, such as the profession of faith: "There is no God but Allāh and Muḥammad is his prophet." The rosary with 99 or 33 beads was in use as early as the 8th century for counting the thousands of repetitions. Man's whole being should eventually be transformed into remembrance of God.

In the mid-9th century some mystics introduced sessions with music and poetry recitals (*samāʿ*) in Baghdad in order to reach the ecstatic experience—and since then debates about the permissibility of *samāʿ,* filling many books, have been written. Narcotics were used in periods of degeneration, coffee by the "sober" mystics (first by the Shādhilīyah after 1300).

Besides the wayfarers (*sālik*) on the path, Ṣūfīs who have no master but are attracted solely by divine grace are also found; they are called Uwaysī, after Uways al-Qaranī, the Yemenite contemporary of the Prophet who never saw him but firmly believed in him. There are also the so-called *majdhūb* ("attracted") who are often persons generally agreed to be more or less mentally deranged.

Symbolism in Ṣūfism. The divine truth was at times revealed to the mystic in visions, auditions, and dreams, in colours and sounds, but to convey these nonrational and ineffable experiences to others the mystic had to rely upon such terminology of worldly experience as that of love and intoxication—often objectionable from the orthodox viewpoint. The symbolism of wine, cup, and cupbearer, first expressed by Abū Yazīd al-Bisṭāmī in the 9th century, became popular everywhere, whether in the verses of the Arab Ibn al-Fāriḍ, or the Persian ʿIrāqī, or the Turk Yunus Emre, and their followers. The hope for the union of the soul with the divine had to be expressed through images of human yearning and love. The love for lovely boys in which the divine beauty manifests itself—according to the alleged Ḥadīth "I saw my Lord in the shape of a youth with a cap awry"—was commonplace in Persian poetry. Union was described as the submersion of the drop in the ocean, the state of the iron in the fire, the vision of penetrating light, or the burning of the moth in the candle (first used by Ḥallāj). Worldly phenomena were seen as black tresses veiling the radiant beauty of the divine countenance. The mystery of unity and diversity was symbolized, for example, under the image of mirrors that reflect the different aspects of the divine, or as prisms colouring the pure light. Every aspect of nature was seen in relation to God. The symbol of the soulbird—in which the human soul is likened to a flying bird—known everywhere, was the centre of ʿAṭṭār's *Manṭeq oṭ-ṭeyr* ("The Birds' Conversation"). The predilection of the mystical poets for the symbolism of the nightingale and rose (the red rose = God's perfect beauty; nightingale = soul; first used by Baqli [died 1206]) stems from the soulbird symbolism. For spiritual education, symbols taken from medicine (healing of the sick soul) and alchemy (changing of base matter into gold) were also used. Many descriptions that were originally applied to God as the goal of love were, in later times, used also for the Prophet, who is said to be like the "dawn between the darkness of the material world and the sun of Reality."

Symbolism of union with the divine

Allusions to the Qurʾān were frequent, especially so to verses that seem to imply divine immanence (God's presence in the world), such as "Whithersoever ye turn, there is the Face of God" (*sūrah* 2:109), or that God is "Closer than your neck-vein" (*sūrah* 50:8). *Sūrah* 7:172—*i.e.,* God's address to the uncreated children of Adam ("Am I not your Lord" [*alastu birabbikum*])—came to denote the pre-eternal love relation between God and man. As for the prophets before Muḥammad, the vision of Moses was considered still imperfect, for the mystic wants the actual vision of God, not His manifestation through a burning bush. Abraham, for whom fire turned into a rose garden, resembles the mystic in his afflictions; Joseph, in his perfect beauty, the mystical beloved after whom the mystic searches. The apocryphal traditions used by the mystics are numerous; such as "Heaven and earth do not contain me, but the heart of my faithful servant contains Me"; and the possibility of a relation between man and God is also explained by the traditional idea: "He (God) created Adam in His image."

THEOSOPHICAL SUFISM

Ṣūfism, in its beginnings a practical method of spiritual education and self-realization, grew slowly into a theosophical system by adopting traditions of Neoplatonism, the Hellenistic world, Gnosticism (an ancient esoteric religiophilosophical movement that viewed matter as evil and spirit as good), and spiritual currents from Iran and various countries in the ancient agricultural lands from the eastern Mediterranean to Iraq. One master who contributed to this development was the Persian as-Suhrawardī, called al-Maqtūl ("killed"), executed in 1191 in Aleppo. To him is attributed the philosophy of *ishrāq* ("illumination"), and he claimed to unite the Persian (Zoroastrian) and Egyptian (Hermetic) traditions. His didactic and doctrinal works in Arabic among other things taught a complicated angelology (theory of angels); some of his smaller Persian treatises depict the journey of the soul across the cosmos; the "Orient" (East) is the world of pure lights and archangels, the "Occident" (West) that of darkness and matter; and man lives in the "Western exile."

At the time of Suhrawardī's death the greatest representative of theosophic Ṣūfism was in his 20s: Ibn al-ʿArabī, born at Murcia, Spain, where speculative tendencies had been visible since Ibn Masarrah's philosophy (died 931). Ibn al-ʿArabī was instructed in mysticism by two Spanish woman saints. Performing the traditional pilgrimage to Mecca, he met there an accomplished young Persian lady who represented for him the divine wisdom. This experience resulted in the charming poems of the *Tarjumān al-ashwāq* ("Interpreter of Yearning"), which the author later explained mystically. Ibn al-ʿArabī composed at least 150 volumes. His magnum opus is *al-Futūḥāt al-Makkīyah* ("The Meccan Revelations") in 560 chapters, in which he expounds his theory of unity of being.

Contributions of Ibn al-ʿArabī

The substance of theosophic Ṣūfism is as follows. According to the Ḥadīth *qudsī*, or "holy tradition"—"I was a hidden treasure and wanted to be known"—the absolute, or God, yearned in his loneliness for manifestation and created the world by effusing being upon the heavenly archetypes, a "theophany (a physical manifestation of deity) through God's imaginative power." The universe is annihilated and created every moment. Every divine name is reflected in a named one. The world and God are said to be like ice and water, or like two mirrors contemplating themselves in each other, joined by a sympathetic union. The Prophet Muḥammad is the universal man, the perfect man, the total theophany of the divine names, the prototype of creation. Muḥammad is the "word," each particular dimension of which is identified with a prophet, and he is also the model for the spiritual realization of the possibilities of man. The mystic has to pass the stages of the Qurʾānic prophets as they are explained in the *Fuṣūṣ al-ḥikam* ("Bezels of Wisdom") until he becomes united with the *ḥaqīqa Muḥammadīya* (the first individualization of the divine in the "Muḥammadan Reality"). Man can have vision only of the form of the faith he professes, and Ibn al-ʿArabī's oft-quoted verse, "I follow the religion of love wherever its camels turn," with its seeming religious tolerance means, as S.H. Nasr puts it: "the form of God is for him no longer the form of this or that faith exclusive of all others but his own eternal form which he encounters." The theories of the perfect man were elaborated by Jīlī (died *c.* 1424) in his compendium *Al-insān al-kāmil* ("The Perfect Man") and became common throughout the Muslim world.

Ibn al-ʿArabī's theosophy has been attacked by orthodox Muslims and mystics of the "sober" school as incongruent with Islām because "a thoroughly monistic system cannot take seriously the objective validity of moral standards." Even the adversaries of the "greatest master" could not, however, help using part of his terminology. Innumerable mystics and poets propagated his ideas, though they only partly understood them, and this circumstance led also to a misinterpretation of the data of early Ṣūfism in the light of existential monism. Later Persian poetry is permeated by the pantheistic feeling of *hama ost* ("everything is He").

Ibn al-ʿArabī's contemporary in Egypt, the poet Ibn al-Fāriḍ, is usually mentioned together with him; Ibn al-Fāriḍ, however, is not a systematic thinker but a full-fledged poet who used the imagery of classical Arabic poetry to describe the state of the lover in extremely artistic verses and has given, in his *Tāʾiyat al-kubrā* ("Poem of the Journey"), glimpses of the way of the mystic, using, as many poets before and after him did, for example, the image of the shadow play for the actions of the creatures who are dependent upon the divine playmaster. His unifying experience is personal and is not the expression of a theosophical system.

SUFI ORDERS

Organization. Mystical life was first restricted to the relation between a master and a few disciples; the foundations of a monastic system were laid by the Persian Abū Saʿīd ebn Abī ol-Kheyr (died 1049), but real orders or fraternities came into existence only from the 12th century onward: ʿAbd al-Qādir al-Jīlānī (died 1166) gathered the first and still most important order around himself; then followed the Suhrawardīyah, and the 13th century saw the formation of large numbers of different orders in the East (for example, Kubrawīya in Khvārezm) and West (Shādhilīyah). Thus, Ṣūfism ceased to be the way of the chosen few and influenced the masses. A strict ritual was

Initiation and investiture

elaborated: when the adept had found a master for whom he had to feel a preformed affinity, there was an initiation ceremony in which he swore allegiance (*bayʿat*) into the master's hand; similarities to the initiation in Ismāʿīlism, the 9th-century sect, and in the guilds suggest a possible interaction. The disciple (*murīd*) had to undergo a stern training; he was often ordered to perform the lowest work in the community, to serve the brethren, to go out to beg (many of the old monasteries subsisted upon alms). A seclusion period of 40 days under hard conditions was common for the adepts in most orders.

Investiture with the *khirqah,* the frock of the master, originally made from shreds and patches, was the decisive act by which the disciple became part of the *silsilah,* the chain of mystical succession and transmission, which leads back—via Junayd—to the Prophet himself and differs in every order. Some mystical leaders claimed to have received their *khirqah* directly from al-Khiḍr, a mysterious immortal saint.

In the earliest times, allegiance was sworn exclusively to one master who had complete power over the disciple, controlling each of his movements, thoughts, visions, and dreams; but later many Ṣūfīs got the *khirqah* from two or more *shaykh*s. There is consequently a differentiation between the *shaykh at-tarbiyah,* who introduces the disciple into the ritual, forms, and literature of the order, and the *shaykh aṣ-ṣuḥbah,* who steadily watches him and with whom the disciple lives. Only a few members of the fraternity remained in the centre (*dargāh, khānqāh, tekke*), close to the *shaykh,* but even those were not bound to celibacy. Most of the initiated returned to their daily life and partook in mystic services only during certain periods. The most mature disciple was invested as *khalīfah* ("successor") to the *shaykh* and was often sent abroad to extend the activities of the order. The *dargāh*s were organized differently in the various orders; some relied completely upon alms, keeping their members in utmost poverty; others were rich, and their *shaykh* was not very different from a feudal lord. Relations with rulers varied—some masters refused contacts with the representatives of political power; others did not mind friendly relations with the grandees.

Discipline and ritual. Each order has peculiarities in its ritual. Most start the instruction with breaking the lower soul; others, such as the later Naqshbandīyah, stress the purification of the heart by constant *dhikr* ("remembrance") and by discourse with the master (*ṣuḥbah*). The forms of *dhikr* vary in the orders. Many of them use the word Allāh, or the profession of faith with its rhythmical wording, sometimes accompanied by movements of the body, or by breath control up to complete holding of the breath. The Mawlawīs, the whirling dervishes, are famous for their dancing ritual, an organized variation of the earlier *samāʿ* practices, which were confined to music and poetry. The Rifāʿīs, the so-called Howling Dervishes, have become known for their practice of hurting themselves while in an ecstatic state that they reach in performing their loud *dhikr.* (Such practices that might well degenerate into mere jugglery are not approved by most orders.) Some orders also teach the *dhikr khafī,* silent repetition of the formulas, and meditation, concentrating upon certain fixed points of the body; thus the Naqshbandīs do not allow any emotional practices and prefer contemplation to ecstasy, perhaps as a result of Buddhist influence from Central Asia. Other orders have special prayers given to the disciples, such as the protective *ḥizb al-baḥr* ("The protective armour of the sea"; *i.e.,* for seafaring people—then extended to all travellers) in the Shādhilīyah order. Most of them prescribe for their disciples additional prayers and meditation at the end of each ritual prayer.

Adaptation to levels of society

Function and role in Islāmic society. The orders formed an excellent means of bringing together the spiritually interested members of the community. They acted as a counterweight against the influence of hairsplitting lawyer-divines and gave the masses an emotional outlet in enthusiastic celebrations (*ʿurs,* "marriage") of the anniversaries of the deaths of founders of mystic orders or similar festivals in which they indulged in music and joy. The orders were adaptable to every social level; thus, some of them were responsible for adapting a number of un-Islāmic folkloristic practices such as veneration of saints. Their way of life often differed so much from Islāmic ideals that one distinguishes in Iran and India between orders *bā sharʿ* (law-bound) and *bī sharʿ* (not following the injunctions of the Qurʾān). Some orders were more fitting for the rural population, such as the Aḥmadīyah (after Aḥmad al-Badawī; died 1286) in Egypt. The Aḥmadīyah, however, even attracted some Mamlūk rulers. The Turkish Bektāshīyah (Haci Bektaş, early 14th century), together with strange syncretistic cults, showed a prevalence of the ideals of the Shīʿites (from Shīʿah—the followers of ʿAlī, son-in-law of the prophet Muḥammad, whose descendants claimed to be rightful successors to the religious leadership of Islām). The figure of ʿAlī played a role also in other fraternities, and the relations between Ṣūfism in the 14th and 15th centuries and the Shīʿah still have to be explored, as is also true of the general influence of Shīʿite ideas on Ṣūfism. Other orders, such as the Shādhilīyah, an offshoot of which still plays an important role among Egyptian officials and employees, are typically middle class. This order demands not a life in solitude but strict adherence to one's profession and fulfillment of one's duty. Still other orders were connected with the ruling classes, such as, for a time, the Chishtīyah in Mughal India, and the Mawlawīyah, whose leader had to invest the Ottoman sultan with the sword. The Mawlawīyah is also largely responsible for the development of classical Turkish poetry, music, and fine arts, just as the Chishtīyah contributed much to the formation of classical Indo-Muslim music.

Missionary activity

The main contribution of the orders, however, is their missionary activity. The members of different orders who settled in India from the early 13th century attracted thousands of Hindus by their example of love of both God and their own brethren and by preaching the equality of men. Missionary activity was often joined with political activity, as in 17th- and 18th-century Central Asia, where the Naqshbandīyah exerted strong political influence. In North Africa the Tijānīyah, founded in 1781, and the Sanūsīyah, active since the early 19th century, both heralded Islām and engaged in politics; the Sanūsīyah fought against Italy, and the former king of Libya was the head of the order. The Tijānīyah extended the borders of Islām toward Senegal and Nigeria, and their representatives founded large kingdoms in West Africa. Their influence, as well as that of the Qādirīyah (see below), is still an important sociopolitical factor in those areas.

Geographical extent of Ṣūfī orders. It would be impossible to number the members of mystical orders in the Islāmic world. Even in such countries as Turkey, where the orders have been banned since 1925, many people still cling to the mystical tradition and feel themselves to be links in the spiritual chains of the orders and try to implement their ideals in modern society. The most widely spread group is, no doubt, the Qādirīyah, whose adherents are found from West Africa to India—the tomb of ʿAbd al-Qādir al-Jīlānī in Baghdad still being a place of pilgrimage. The areas where the Sanūsīyah live are restricted to the Maghrib, the Atlas Massif, and the coastal plain from Morocco to Tunisia, whereas the Tijānīyah has some offshoots in Turkey. Such rural orders as the Egyptian Aḥmadīyah and Dasūqīyah (named after Ibrāhīm ad-Dasūqī; died 1277) are bound to their respective countries, as are the Mawlawīs and Bektāshīyah to the realms of the former Ottoman Empire. The Bektāshīyah had gained political importance in the empire because of its relations with the Janissaries, the standing army. Albania, since 1929, has had a strong and officially recognized group of Bektāshīyah who were even granted independent status after World War II. The Shaṭṭārīyah (derived from ʿAbd ash-Shaṭṭār; died 1415) extends from India to Java, whereas the Chishtīyah (derived from Khwājah Muʿīnud-Dīn Chishtīp; died 1236 in Ajmer) and Suhrawardīyah remain mainly inside the Indo-Pakistan subcontinent. The Kubrāwīyah reached Kashmir through ʿAlī Hama-dhānī (died 1385), a versatile author, but the order later lost its influence.

The great variety of possible forms may be seen by comparing the Haddāwah, vagabonds in Morocco, who "do not spoil God's day by work" and the Shādhilīyah with a sober attitude toward professional life and careful introspection. Out of the Shādhilīyah developed the austere Darqāwīyah, who, in turn, produced the ʿAlāwīyah, whose master has attracted even a number of Europeans. The splitting up and formation of suborders is a normal process, but most of the subgroups have only local importance. The High Ṣūfī Convent in Egypt counts 60 registered orders.

SIGNIFICANCE

Ṣūfism has helped to shape large parts of Muslim society. The orthodox disagree with such aspects of Ṣūfism as saint worship, visiting of tombs, musical performances, miracle mongering, degeneration into jugglery, and the adaptation of pre-Islāmic and un-Islāmic customs; and the reformers object to the influences of the monistic interpretation of Islām upon moral life and human activities. The importance given to the figure of the master is accused of yielding negative results; the *shaykh* as the almost infallible leader of his disciples and admirers could gain dangerous authority and political influence, for the illiterate villagers in backward areas used to rely completely upon the "saint." Yet, other masters have raised their voices against social inequality and have tried, even at the cost of their lives, to change social and political conditions for the better and to spiritually revive the masses. The missionary activities of the Ṣūfīs have enlarged the fold of the faithful. The importance of Ṣūfism for spiritual education, and inculcation in the faithful of the virtues of trust in God, piety, faith in God's love, and veneration of the Prophet, cannot be overrated. The *dhikr* formulas still preserve their consoling and quieting power even for the illiterate. Mysticism permeates Persian literature and other literatures influenced by it. Such poetry has always been a source of happiness for millions, although some modernists have disdained its "narcotic" influence on Muslim thinking.

Industrialization and modern life have led to a constant decrease in the influence of Ṣūfī orders in many countries. The spiritual heritage is preserved by individuals who sometimes try to show that mystical experience conforms to modern science. Today in the West, Ṣūfism is popularized, but the genuinely and authentically devout are aware that it requires strict discipline, and that its goal can be reached—if at all—as they say, only by throwing oneself into the consuming fire of divine love. (An.Sc.)

Islāmic philosophy

Origin and inspiration of Islāmic philosophy

The origin and inspiration of philosophy in Islām are quite different from those of Islāmic theology. Philosophy developed out of and around the nonreligious practical and theoretical sciences; it recognized no theoretical limits other than those of human reason itself; and it assumed that the truth found by unaided reason does not disagree with the truth of Islām when both are properly understood. Islāmic philosophy was not a handmaid of theology. The two disciplines were related, because both followed the path of rational inquiry and distinguished themselves from traditional religious disciplines and from mysticism, which sought knowledge through practical, spiritual purification. Islāmic theology was Islāmic in the strict sense: it confined itself within the Islāmic religious community, and it remained separate from the Christian and Jewish theologies that developed in the same cultural context and used Arabic as a linguistic medium. No such separation is observable in the philosophy developed in the Islāmic cultural context and written in Arabic: Muslims, Christians, and Jews participated in it and separated themselves according to the philosophic rather than the religious doctrines they held.

THE EASTERN PHILOSOPHERS

Background and scope of philosophical interest in Islām. The background of philosophic interest in Islām is found in the earlier phases of theology. But its origin is found in the translation of Greek philosophic works. By the middle of the 9th century, there were enough translations of scientific and philosophic works from Greek, Pahlavi, and Sanskrit to show those who read them with care that scientific and philosophic inquiry was something more than a series of disputations based on what the theologians had called sound reason. Moreover, it became evident that there existed a tradition of observation, calculation, and theoretical reflection that had been pursued systematically, refined, and modified for over a millennium.

The basis of Islāmic philosophy in Greek philosophical and scientific works

The scope of this tradition was broad: it included the study of logic, the sciences of nature (including psychology and biology), the mathematical sciences (including music and astronomy), metaphysics, ethics, and politics. Each of these disciplines had a body of literature in which its principles and problems had been investigated by classical authors, whose positions had been, in turn, stated, discussed, criticized, or developed by various commentators. Islāmic philosophy emerged from its theological background when Muslim thinkers began to study this foreign tradition, became competent students of the ancient philosophers and scientists, criticized and developed their doctrines, clarified their relevance for the questions raised by the theologians, and showed what light they threw on the fundamental issues of revelation, prophecy, and the divine law.

Relation to the Muʿtazilah and interpretation of theological issues. *The teachings of al-Kindī.* Although the first Muslim philosopher, al-Kindī, who flourished in the first half of the 9th century, lived during the triumph of the Muʿtazilah of Baghdad and was connected with the ʿAbbāsid caliphs who championed the Muʿtazilah and patronized the Hellenistic sciences, there is no clear evidence that he belonged to a theological school. His writings show him to have been a diligent student of Greek and Hellenistic authors in philosophy and point to his familiarity with Indian arithmetic. His conscious, open, and unashamed acknowledgment of earlier contributions to scientific inquiry was foreign to the spirit, method, and purpose of the theologians of the time. His acquaintance with the writings of Plato and Aristotle was still incomplete and technically inadequate. He improved the Arabic translation of the "Theology of Aristotle" but made only a selective and circumspect use of it.

Al-Kindī's interest in scientific inquiry

Devoting most of his writings to questions of natural philosophy and mathematics, al-Kindī was particularly concerned with the relation between corporeal things, which are changeable, in constant flux, infinite, and as such unknowable, on the one hand, and the permanent world of forms (spiritual or secondary substances), which are not subject to flux yet to which man has no access except through things of the senses. He insisted that a purely human knowledge of all things is possible, through the use of various scientific devices, learning such things as mathematics and logic, and assimilating the contributions of earlier thinkers. The existence of a "supernatural" way to this knowledge in which all these requirements can be dispensed with was acknowledged by al-Kindī: God may choose to impart it to his prophets by cleansing and illuminating their souls and by giving them his aid, right guidance, and inspiration; and they, in turn, communicate it to ordinary men in an admirably clear, concise, and comprehensible style. This is the prophets' "divine" knowledge, characterized by a special mode of access and style of exposition. In principle, however, this very same knowledge is accessible to man without divine aid, even though "human" knowledge may lack the completeness and consummate logic of the prophets' divine message.

Reflection on the two kinds of knowledge—the human knowledge bequeathed by the ancients and the revealed knowledge expressed in the Qurʾān—led al-Kindī to pose a number of themes that became central in Islāmic philosophy: the rational–metaphorical exegesis of the Qurʾān and the Ḥadīth; the identification of God with the first being and the first cause; creation as the giving of being and as a kind of causation distinct from natural causation and Neoplatonic emanation; and the immortality of the individual soul.

The teachings of Abū Bakr ar-Rāzī. The philosopher whose principal concerns, method, and opposition to authority were inspired by the extreme Muʿtazilah was the physician Abū Bakr ar-Rāzī (flourished 9th–10th centuries). He adopted the Muʿtazilah's atomism and was intent

The concept of the five eternal principles

on developing a rationally defensible theory of creation that would not require any change in God or attribute to him responsibility for the imperfection and evil prevalent in the created world. To this end, he expounded the view that there are five eternal principles—God, Soul, prime matter, infinite, or absolute, space, and unlimited, or absolute, time—and explained creation as the result of the unexpected and sudden turn of events (*faltah*). *Faltah* occurred when Soul, in her ignorance, desired matter and the good God eased her misery by allowing her to satisfy her desire and to experience the suffering of the material world, and then gave her reason to make her realize her mistake and deliver her from her union with matter, the cause of her suffering and of all evil. Ar-Rāzī claimed that he was a Platonist, that he disagreed with Aristotle, and that his views were those of the Ṣābians of Harran and the Brahmins (Hindu teachers).

Ismāʿīlī theologians became aware of the kinship between certain elements of his cosmology and their own. They disputed with him during his lifetime and continued afterward to refute his doctrines in their writings. According to their account of his doctrines, he was totally opposed to authority in matters of knowledge, believed in the progress of the arts and sciences, and held that all reasonable men are equally able to look after their own affairs, equally inspired and able to know the truth of what earlier men had taught, and equally able to improve upon it. Ismāʿīlī theologians were incensed, in particular, by his wholesale rejection of prophecy, particular revelation, and divine laws. They were likewise opposed to his criticisms of religion in general as a device employed by evil men and a kind of tyranny over men that exploits their innocence and credulity, perpetuates ignorance, and leads to conflicts and wars.

Although the fragmentary character of al-Kindī's and ar-Rāzī's surviving philosophic writings does not permit passing firm and independent judgment on their accomplishments, they tend to bear out the view of later Muslim students of philosophy that both lacked competence in the logical foundation of philosophy, were knowledgeable in some of the natural sciences but not in metaphysics, and were unable to narrow the gap that separated philosophy from the new religion, Islām.

The relationship of law and theology to the community

The teachings of al-Fārābī. *Political philosophy and the study of religion.* The first philosopher to meet this challenge was al-Fārābī (flourished 9th–10th centuries). He saw that theology and the juridical study of the law were derivative phenomena that function within a framework set by the prophet as lawgiver and founder of a human community. In this community, revelation defines the opinions the members of the community must hold and the actions they must perform if they are to attain the earthly happiness of this world and the supreme happiness of the other world. Philosophy could not understand this framework of religion as long as it concerned itself almost exclusively with its truth content and confined the study of practical science to individualistic ethics and personal salvation.

In contrast to al-Kindī and ar-Rāzī, al-Fārābī recast philosophy in a new framework analogous to that of the Islāmic religion. The sciences were organized within this philosophic framework so that logic, physics, mathematics, and metaphysics culminated in a political science whose subject matter is the investigation of happiness and how it can be realized in cities and nations. The central theme of this political science is the founder of a virtuous or excellent community. Included in this theme are views concerning the supreme rulers who follow the founder, their qualifications, and how the community must be ordered so that its members attain happiness as citizens rather than isolated human beings. Once this new philosophical framework was established, it became possible to conduct a philosophical investigation of all the elements that constituted the Islāmic community: the prophet-lawgiver, the aims of the divine laws, the legislation of beliefs as well as actions, the role of the successors to the founding legislator, the grounds of the interpretation or reform of the law, the classification of human communities according to their doctrines in addition to their size, and the critique of "ignorant" (pagan), "transgressing," "falsifying," and "erring" communities. Philosophical cosmology, psychology, and politics were blended by al-Fārābī into a political theology whose aim was to clarify the foundations of the Islāmic community and defend its reform in a direction that would promote scientific inquiry and encourage philosophers to play an active role in practical affairs.

Significance of al-Fārābī in the dissemination of Greek philosophical thought

Interpretation of Plato and Aristotle. Behind this public, or exoteric, aspect of al-Fārābī's work stood a massive body of more properly philosophic or scientific inquiries, which established his reputation among Muslims as the greatest philosophical authority after Aristotle, a great interpreter of the thought of Plato and Aristotle and their commentators, and a master to whom almost all major Muslim as well as a number of Jewish and Christian philosophers turned for a fuller understanding of the controversial, troublesome, and intricate questions of philosophy. Continuing the tradition of the Hellenistic masters of the Athenian and Alexandrian philosophical schools, al-Fārābī broadened the range of philosophical inquiry and fixed its form. He paid special attention to the study of language and its relation to logic. In his numerous commentaries on Aristotle's logical works, he expounded for the first time in Arabic the entire range of the scientific and nonscientific forms of argument and established the place of logic as an indispensable prerequisite for philosophic inquiry. His writings on natural science exposed the foundation and assumptions of Aristotle's physics and dealt with the arguments of Aristotle's opponents, both philosophers and scientists, pagan, Christian, and Muslim.

The analogy of religion and philosophy. Al-Fārābī's theological and political writings showed later Muslim philosophers the way to deal with the question of the relation between philosophy and religion and presented them with a complex set of problems that they continued to elaborate, modify, and develop in different directions. Starting with the view that religion is analogous or similar to philosophy, al-Fārābī argued that the idea of the true prophet-lawgiver ought to be the same as that of the true philosopher-king. Thus, he challenged both al-Kindī's view that prophets and philosophers have different and independent ways to the highest truth available to man and ar-Rāzī's view that philosophy is the only way to that knowledge. That a man could combine the functions of prophecy, lawgiving, philosophy, and kingship did not necessarily mean that these functions were identical; it did mean, however, that they all are legitimate subjects of philosophic inquiry. Philosophy must account for the powers, knowledge, and activities of the prophet, lawgiver, and king, which it must distinguish from and relate to those of the philosopher. The public, or political, function of philosophy was emphasized. Unlike Neoplatonism, which had for long limited itself to the Platonic teaching that the function of philosophy is to liberate the soul from the shadowy existence of the cave—in which knowledge can only be imperfectly comprehended as shadows reflecting the light of the truth beyond the cave (the world of senses)—al-Fārābī insisted with Plato that the philosopher must be forced to return to the cave, learn to talk to its inhabitants in a manner they can comprehend, and engage in actions that may improve their lot.

The use of al-Fārābī's theological–political writings to reform Ismāʿīlī thought

Impact on Ismāʿīlī theology. Although it is not always easy to know the immediate practical intentions of a philosopher, it must be remembered that in al-Fārābī's lifetime the fate of the Islāmic world was in the balance. The Sunnī caliphate's power hardly extended beyond Baghdad, and it appeared quite likely that the various Shīʿī sects, especially the Ismāʿīlīs, would finally overpower it and establish a new political order. Of all the movements in Islāmic theology, Ismāʿīlī theology was the one that was most clearly and massively penetrated by philosophy. Yet, its Neoplatonic cosmology, revolutionary background, antinomianism (antilegalism), and general expectation that divine laws were about to become superfluous with the appearance of the *qāʾim* (the imam of the "resurrection") all militated against the development of a coherent political theory to meet the practical demands of political life and present a viable practical alternative to the Sunnī caliphate. Al-Fārābī's theologico-political writings helped point out this basic defect of Ismāʿīlī theology. Under the Fāṭimids in Egypt (969–1171), Ismāʿīlī theology modified its cosmology in the direction suggested by al-Fārābī, returned to the view that the community must continue to live under the divine law, and postponed the prospect of the abolition of divine laws and the appearance of the *qāʾim* to an indefinite point in the future.

The teachings of Avicenna. *The "Oriental Philosophy."* Even more indicative of al-Fārābī's success is the fact that his writings helped produce a philosopher of the stature of Avicenna (flourished 10th–11th centuries), whose versatility, imagination, inventiveness, and prudence shaped philosophy into a powerful force that gradually penetrated Islāmic theology and mysticism and Persian poetry in eastern Islām and gave them universality and theoretical depth. His own personal philosophic views, he said, were those of the ancient sages of Greece (including the genuine views of Plato and Aristotle), which he had set forth in the "Oriental Philosophy," a book that has not survived and probably was not written or meant to be written. They were not identical with the common Peripatetic (Aristotelian) doctrines and were to be distinguished from the learning of his contemporaries, the Christian "Aristotelians" of Baghdad, which he attacked as vulgar, distorted, and falsified. His most voluminous writing, *Kitāb ash-shifāʾ* ("The Book of Healing"), was meant to accommodate the doctrines of other philosophers as well as hint at his own personal views, which are elaborated elsewhere in more imaginative and allegorical forms.

Distinction between essence and existence and the doctrine of creation. Avicenna had learned from certain hints in al-Fārābī that the exoteric teachings of Plato regarding "forms," "creation," and the immortality of individual souls were closer to revealed doctrines than the genuine views of Aristotle, that the doctrines of Plotinus and later Neoplatonic commentators were useful in harmonizing Aristotle's views with revealed doctrines, and that philosophy must accommodate itself to the divine law on the issue of creation and of reward and punishment in the hereafter, which presupposes some form of individual immortality. Following al-Fārābī's lead, Avicenna initiated a full-fledged inquiry into the question of being, in which he distinguished between essence and existence. He argued that the fact of existence cannot be inferred from or accounted for by the essence of existing things and that form and matter by themselves cannot interact and originate the movement of the universe or the progressive actualization of existing things. Existence must, therefore, be due to an agent-cause that necessitates, imparts, gives, or adds existence to an essence. To do so, the cause must be an existing thing and coexist with its effect. The universe consists of a chain of actual beings, each giving existence to the one below it and responsible for the existence of the rest of the chain below. Because an actual infinite is deemed impossible by Avicenna, this chain as a whole must terminate in a being that is wholly simple and one, whose essence is its very existence, and therefore is self-sufficient and not in need of something else to give it existence. Because its existence is not contingent on or necessitated by something else but is necessary and eternal in itself, it satisfies the condition of being the necessitating cause of the entire chain that constitutes the eternal world of contingent existing things.

Avicenna's investigation of being

All creation is necessarily and eternally dependent upon God. It consists of the intelligences, souls, and bodies of the heavenly spheres, each of which is eternal, and the sublunary sphere, which is also eternal, undergoing a perpetual process of generation and corruption, of the succession of form over matter, very much in the manner described by Aristotle.

The immortality of individual souls. There is, however, a significant exception to this general rule: the human rational soul. Man can affirm the existence of his soul from direct consciousness of his self (what he means when he says "I"); and he can imagine this happening even in the absence of external objects and bodily organs. This proves, according to Avicenna, that the soul is indivisible, immaterial, and incorruptible substance, not imprinted in matter, but created with the body, which it uses as an instrument. Unlike other immaterial substances (the intelligences and souls of the spheres), it is not pre-eternal but is generated, or made to exist, at the same time as the individual body, which can receive it, is formed. The composition, shape, and disposition of its body and the soul's success or failure in managing and controlling it, the formation of moral habits, and the acquisition of knowledge all contribute to its individuality and difference from other souls. Though the body is not resurrected after its corruption, the soul survives and retains all the individual characteristics, perfections or imperfections, that it achieved in its earthly existence and in this sense is rewarded or punished for its past deeds. Avicenna's claim that he has presented a philosophic proof for the immortality of generated ("created") individual souls no doubt constitutes the high point of his effort to harmonize philosophy and religious beliefs.

The doctrine of individual souls

Philosophy, religion, and mysticism. Having accounted for the more difficult issues of creation and the immortality of individual souls, Avicenna proceeded to explain the faculty of prophetic knowledge (the "sacred" intellect), revelation (imaginative representation meant to convince the multitude and improve their earthly life), miracles, and the legal and institutional arrangements (acts of worship and the regulation of personal and public life) through which the divine law achieves its end. Avicenna's explanation of almost every aspect of Islām is pursued on the basis of extensive exegesis of the Qurʾān and the Ḥadīth. The primary function of religion is to assure the happiness of the many. This practical aim of religion (which Avicenna saw in the perspective of Aristotle's practical science) enabled him to appreciate the political and moral functions of divine revelation and account for its form and content. Revealed religion, however, has a subsidiary function also—that of indicating to the few the need to pursue the kind of life and knowledge appropriate to rare individuals endowed with special gifts. These men must be dominated by the love of God to facilitate the achievement of the highest knowledge. In many places Avicenna appears to identify these men with the mystics. The identification of the philopher as a kind of mystic conveyed a new image of the philosopher as a member of the religious community who is distinguished from his coreligionists by his otherworldliness, dedicated to the inner truth of religion, and consumed by the love of God.

Prophetic, revelatory, and social knowledge

Avicenna's allegorical and mystical writings are usually called "esoteric" in the sense that they contain his personal views cast in an imaginative, symbolic form. The esoteric works must, then, be interpreted. Their interpretation must move away from the explicit doctrines contained in "exoteric" works such as the *Shifāʾ* and recover "the unmixed and uncorrupted truth" set forth in the "Oriental Philosophy." The "Oriental Philosophy," however, has never been available to anyone, and it is doubtful that it was written at all. This dilemma has made interpretation both difficult and rewarding for Muslim philosophers and modern scholars alike.

THE WESTERN PHILOSOPHERS

Background and characteristics of the western Muslim philosophical tradition. Andalusia (in Spain) and western North Africa contributed little of substance to Islāmic theology and philosophy until the 12th century. Legal strictures against the study of philosophy were more effective than in the east. Scientific interest was channelled into medicine, pharmacology, mathematics, astronomy, and logic. More general questions of physics and metaphysics were treated sparingly and in symbols, hints, and allegories. By the 12th century, however, the writings of al-Fārābī, Avicenna, and al-Ghazālī had found their way to the west. A philosophical tradition emerged, based primarily on the study of al-Fārābī. It was critical of Avicenna's philosophic innovations and not convinced that al-Ghazālī's critique of Avicenna touched philosophy as such, and it refused to acknowledge the position assigned by both to mysticism. The survival of philosophy in the west required extreme prudence, emphasis on its scientific character, abstention from meddling in political or religious matters, and abandonment of the hope of effecting extensive doctrinal or institutional reform.

Significance of Islāmic philosophers in the West

The duty of the philosopher

The teachings of Ibn Bājjah. *Theoretical science and intuitive knowledge.* Ibn Bājjah (died 1138) initiated this tradition with a radical interpretation of al-Fārābī's political philosophy that emphasized the virtues of the perfect but nonexistent city and the vices prevalent in all existing cities. He concluded that the philosopher must order his own life as a solitary individual, shun the company of nonphilosophers, reject their opinions and ways of life, and concentrate on reaching his own final goal by pursuing the theoretical sciences and achieving intuitive knowledge through contact with the Active Intelligence. The multitude live in a dark cave and see only dim shadows. Their ways of life and their imaginings and beliefs consist of layers of darkness that cannot be known through reason alone. Therefore, the divine law has been revealed to enable man to know this dark region. The philosopher's duty is to seek the light of the sun (the intellect). To do so, he must leave the cave, see all colours as they truly are and see light itself, and finally become transformed into that light. The end, then, is contact with Intelligence, not with something that transcends Intelligence (as in Plotinus, Ismāʿīlism, and mysticism), a doctrine criticized by Ibn Bājjah as the way of imagination, motivated by desire, and aiming at pleasure. Philosophy, he claimed, is the only way to the truly blessed state, which can be achieved only by going through theoretical science, even though it is higher than theoretical science.

Unconcern of philosophy with reform. Ibn Bājjah's cryptic style and the unfinished form in which he left most of his writings tend to highlight his departures from al-Fārābī and Avicenna. Unlike al-Fārābī, he is silent about the philosopher's duty to return to the cave and partake of the life of the city. He appears to argue that the aim of philosophy is attainable independently from the philosopher's concern with the best city and is to be achieved in solitude or, at most, in comradeship with philosophic souls. Unlike Avicenna, who prepared the way for him by clearly distinguishing between theoretical and practical science, Ibn Bājjah is concerned with practical science only insofar as it is relevant to the life of the philosopher. He is contemptuous of allegories and imaginative representations of philosophic knowledge, silent about theology, and shows no concern with improving the multitude's opinions and way of life.

Explanation of Avicenna's "Oriental Philosophy"

The teachings of Ibn Ṭufayl. *The philosopher as a solitary individual.* In his philosophic story *Ḥayy ibn Yaqẓān* ("*Alive Son of Awake*"), the philosopher Ibn Ṭufayl (died 1185) fills gaps in the work of his predecessor Ibn Bājjah. The story communicates the secrets of Avicenna's "Oriental Philosophy" as experienced by a solitary hero, who grows up on a deserted island, learns about the things around him, acquires knowledge of the natural universe (including the heavenly bodies), and achieves the state of "annihilation" (*fanāʾ*) of the self in the divine reality. This is the apparent and traditional secret of the "Oriental Philosophy." But the hero's wisdom is still incomplete, for he knows nothing about other human beings, their way of life, or their laws. When he chances to meet one of them—a member of a religious community inhabiting a neighbouring island, who is inclined to reflect on the divine law and seek its inner, spiritual meanings and who has abandoned the society of his fellow men to devote himself to solitary meditation and worship—he does not at first recognize that he is a human being like himself, cannot communicate with him, and frightens him by his wild aspect. After learning about the doctrines and acts of worship of the religious community, he understands them as alluding to and agreeing with the truth that he had learned by his own unaided effort, and he goes as far as admitting the validity of the religion and the truthfulness of the prophet who gave it. He cannot understand, however, why the prophet communicated the truth by way of allusions, examples, and corporeal representations or why religion permits men to devote much time and effort to practical, worldly things.

Concern for reform. His ignorance of the nature of most men and his compassion for them make the solitary hero insist on becoming their saviour. He persuades his companion to take him to his coreligionists and help him convert them to the naked truth by propagating among them "the secrets of wisdom." His education is completed when he fails in his endeavour. He learns the limits beyond which the multitude cannot ascend without becoming confused and unhappy. He also learns the wisdom of the divine lawgiver in addressing them in the way they can understand, enabling them to achieve limited ends through doctrines and actions suited to their abilities. The story ends with the hero taking leave of these people after apologizing to them for what he did and confessing that he is now fully convinced that they should not change their ways but remain attached to the literal sense of the divine law and obey its demands. He returns to his own island to continue his former solitary existence.

The hidden secret of Avicenna's "Oriental Philosophy." The hidden secret of Avicenna's "Oriental Philosophy" appears, then, to be that the philosopher must return to the cave, educate himself in the ways of nonphilosophers, and understand the incompatibility between philosophical life and the life of the multitude, which must be governed by religion and divine laws. Otherwise, his ignorance will lead him to actions dangerous to the well-being of both the community and philosophy. Because Ibn Ṭufayl's hero had grown up as a solitary human being, he lacks the kind of wisdom that could have enabled him to pursue philosophy in a religious community and be useful to such a community. Neither the conversion of the community to philosophy nor the philosopher's solitary life is a viable alternative.

Solution to the problem of the relation of philosophy to the community

The teachings of Averroës. *Philosophy.* To Ibn Ṭufayl's younger friend Averroës (Ibn Rushd, flourished 12th century) belongs the distinction of presenting a solution to the problem of the relation between philosophy and the Islāmic community in the west, a solution meant to be legally valid, theologically sound, and philosophically satisfactory. Here was a philosopher fully at home in what Ibn Bājjah had called the many layers of darkness. His legal training (he was a judge by profession) and his extensive knowledge of the history of the religious sciences (including theology) enabled him to speak with authority about the principles of Islāmic law and their application to theological and philosophic issues and to question the authority of al-Ghazālī and the Ashʿarīs to determine correct beliefs and right practices. He was able to examine in detail from the point of view of the divine law the respective claims of theology and philosophy to possess the best and surest way to human knowledge, to be competent to interpret the ambiguous expressions of the divine law, and to have presented convincing arguments that are theoretically tenable and practically salutary.

The intention of the divine law

The divine law. The intention of the divine law, he argued, is to assure the happiness of all members of the community. This requires everyone to profess belief in the basic principles of religion as enunciated in the Qurʾān, the Ḥādith, and the *ijmāʿ* (consensus) of the learned and to perform all obligatory acts of worship. Beyond this, the only just requirement is to demand that each pursue knowledge as far as his natural capacity and makeup permit. The few who are endowed with the capacity for the highest, demonstrative knowledge are under a divine legal obligation to pursue the highest wisdom, which is philosophy, and they need not constantly adjust its certain conclusions to what theologians claim to be the correct interpretation of the divine law. Being dialecticians and rhetoricians, theologians are not in a position to determine what is and is not correct interpretation of the divine law so far as philosophers are concerned. The divine law directly authorizes philosophers to pursue its interpretation according to the best—*i.e.,* demonstrative or scientific—method, and theologians have no authority to interfere with the conduct of this activity or judge its conclusions.

Theology. On the basis of this legal doctrine, Averroës judged the theologian al-Ghazālī's refutation of the philosophers ineffective and inappropriate because al-Ghazālī did not understand and even misrepresented the philosophers' positions and used arguments that only demonstrate his incompetence in the art of demonstration. He criticized al-Fārābī and Avicenna also for accommodating the theologians of their time and for departing from the path of the ancient philosophers merely to please the theologians. At the other extreme are the multitude for whom there are no more convincing arguments than those found in the divine law itself. Neither philosophers nor theologians are permitted to disclose to the multitude interpretations of the ambiguous verses of the Qur'ān or to confuse them with their own doubts or arguments. Finally, there are those who belong to neither the philosophers nor the multitude, either because they are naturally superior to the multitude but not endowed with the gift for philosophy or else are students in initial stages of philosophic training. For this intermediate group theology is necessary. It is an intermediate discipline that is neither strictly legal nor philosophic. It lacks their certain principles and sure methods. Therefore, theology must remain under the constant control of philosophy and the supervision of the divine law so as not to drift into taking positions that cannot be demonstrated philosophically or that are contrary to the intention of the divine law. Averroës himself composed a work on theology to show how these requirements can be met: *Kitāb al-kashf ʿan manāhij al-adillah* ("Exposition of the Methods of Proofs"). In the Latin West he was best known for his philosophical answer to al-Ghazālī, *Tahāfut at-tahāfut* ("Incoherence of the Incoherence"), and for his extensive commentaries on Aristotle, works that left their impact on medieval and renaissance European thought.

The new wisdom: synthesis of philosophy and mysticism

PHILOSOPHY, TRADITIONALISM, AND THE NEW WISDOM

The role of Arabic philosophical literature in the West

Philosophy. The western tradition in Islāmic philosophy formed part of the Arabic philosophic literature that was translated into Hebrew and Latin and that played a significant role in the development of medieval philosophy in the Latin West and the emergence of modern European philosophy. Its impact on the development of philosophy in eastern Islām was not as dramatic, but was important nevertheless. Students of this tradition—*e.g.,* the prominent Jewish philosopher Maimonides (flourished 12th century) and the historian Ibn Khaldūn (flourished 14th century)—moved to Egypt, where they taught and had numerous disciples. Most of the writings of Ibn Bājjah, Ibn Ṭufayl, and Averroës found their way to the east also, where they were studied alongside the writings of their eastern predecessors. In both regions thinkers who held to the idea of philosophy as formulated by the eastern and western philosophers thus far discussed continued to teach. They became isolated and overwhelmed, however, by the resurgence of traditionalism and the emergence of a new kind of philosophy whose champions looked on the earlier masters as men who had made significant contributions to the progress of knowledge but whose overall view was defective and had now become outdated.

The emergence of philosophy under a new garb

Traditionalism and the new wisdom. Resurgent traditionalism found effective defenders in men such as Ibn Taymīyah (13th–14th centuries) who employed a massive battery of philosophic, theological, and legal arguments against every shade of innovation and called for a return to the beliefs and practices of the pious ancestors. These attacks, however, did not deal a decisive blow to philosophy as such. It rather drove philosophy underground for a period, only to re-emerge in a new garb. A more important reason for the decline of the earlier philosophic tradition, however, was the renewed vitality and success of the program formulated by al-Ghazālī for the integration of theology, philosophy, and mysticism into a new kind of philosophy called wisdom (*ḥikmah*). It consisted of a critical review of the philosophy of Avicenna, preserving its main external features (its logical, physical, and, in part, metaphysical structure, and its terminology) and introducing principles of explanation for the universe and its relation to God based on personal experience and direct vision.

Characteristic features of the new wisdom. If the popular theology preached by the philosophers from al-Fārābī to Averroës is disregarded, it is evident that philosophy proper meant to them what al-Fārābī called a state of mind dedicated to the quest and the love for the highest wisdom. None of them claimed, however, that he had achieved this highest wisdom. In contrast, every leading exponent of the new wisdom stated that he had achieved or received it through a private illumination, dream (at times inspired by the Prophet), or vision and on this basis proceeded to give an explanation of the inner structure of natural and divine things. In every case, this explanation incorporated Platonic or Aristotelian elements but was more akin to some version of a later Hellenistic philosophy, which had found its way earlier into one or another of the schools of Islāmic theology, though, because of the absence of an adequate philosophic education on the part of earlier theologians, it had not been either elaborated or integrated into a comprehensive view. Like their late-Hellenistic counterparts, exponents of the new wisdom proceeded through an examination of the positions of Plato, Aristotle, and Plotinus. They also gave special attention to the insights of the pre-Socratic philosophers of ancient Greece and the myths and revelations of the ancient Near East, and they offered to resolve the fundamental questions that had puzzled earlier philosophers. In its basic movement and general direction, therefore, Islāmic philosophy between the 9th and the 19th centuries followed a course parallel to that of Greek philosophy from the 5th century BC to the 6th century AD.

Use of Greek philosophers by exponents of the new wisdom

Critiques of Aristotle in Islāmic theology. The critique of Aristotle that had begun in Muʿtazilī circles and had found a prominent champion in Abū Bakr ar-Rāzī was provided with a more solid foundation in the 10th and 11th centuries by the Christian theologians and philosophers of Baghdad, who translated the writings of the Hellenistic critics of Aristotle (*e.g.,* John Philoponus) and made use of their arguments in commenting on Aristotle and in independent theological and philosophic works. Avicenna's attack on these so-called Aristotelians and their Hellenistic predecessors (an attack that had been initiated by al-Fārābī and was to be continued by Averroës) did not prevent the spread of their theologically based anti-Aristotelianism among Jewish and Muslim students of philosophy in the 12th century, such as Abū al-Barakāt al-Baghdādī (died *c.* 1175) and Fakhr ad-Dīn ar-Rāzī. These theologians continued and intensified al-Ghazālī's attacks on Avicenna and Aristotle (especially their views on time, movement, matter, and form, the nature of the heavenly bodies, and the relation between the intelligible and sensible worlds). They suggested that a thorough examination of Aristotle had revealed to them, on philosophic grounds, that the fundamental disagreements between him and the theologies based on the revealed religions represented open options and that Aristotle's view of the universe was in need of explanatory principles that could very well be supplied by theology. This critique provided the framework for the integration of philosophy into theology from the 13th century onward.

Synthesis of philosophy and mysticism. Although it made use of such theological criticisms of philosophy, the new wisdom took the position that theology did not offer a positive substitute for and was incapable of solving the difficulties of "Aristotelian" philosophy. It did not question the need to have recourse to the Qur'ān and the Ḥadīth to find the right answers. It insisted (on the authority of a long-standing mystical tradition), however, that theology concerns itself only with the external expressions of this divine source of knowledge. The inner core was reserved for the adepts of the mystic path whose journey leads to the experience of the highest reality in dreams and visions. Only the mystical adepts are in possession of the one true wisdom, the ground of both the external expressions of the divine law and the phenomenal world of human experience and thought.

The inner core of divine knowledge as reserved for the adepts

PRIMARY TEACHERS OF THE NEW WISDOM

The teachings of as-Suhrawardī. The first master of the new wisdom, as-Suhrawardī (12th century), called it the "Wisdom of Illumination." He rejected Avicenna's distinction between essence and existence and Aristotle's distinction between substance and accidents, possibility and actuality, and matter and form, on the ground that they are mere distinctions of reason. Instead, he concentrated on the notion of being and its negation, which he called "light" and "darkness," and explained the gradation of beings as gradation of their mixture according to the degree of "strength," or "perfection," of their light. This gradation forms a single continuum that culminates in pure light, self-luminosity, self-awareness, self-manifestation, or self-knowledge, which is God, the light of lights, the true One. The stability and eternity of this single continuum result from every higher light overpowering and subjugating the lower, and movement and change in it result from each of the lower lights desiring and loving the higher.

As-Suhrawardī's "pan-lightism" is not particularly close to traditional Islāmic views concerning the creation of the world and God's knowledge of particulars. The structure of his universe remains largely that of the Platonists and the Aristotelians. And his account of the emanation process avoids the many difficulties that had puzzled Neoplatonists as they tried to understand how the second hypostasis (reality) proceeds from the One. He asserted that it proceeds without in any way affecting the One and that the One's self-sufficiency is enough to explain the giving out that seems to be both spontaneous and necessary. His doctrine is presented in a way that suggests that it is the inner truth behind the exoteric (external) teachings of Islām as well as Zoroastrianism, indeed the wisdom of all ancient sages, especially Iranians and Greeks, and the revealed religions as well. This neutral yet positive attitude toward the diversity of religions, which was not absent among Muslim philosophers and mystics, was to become one of the hallmarks of the new wisdom. Different religions were seen as different manifestations of the same truth, their essential agreement was emphasized, and various attempts were made to combine them into a single harmonious religion meant for all of mankind.

The positive attitude to the diversity of religions

As-Suhrawardī takes an important step in this direction through his doctrine of imaginative-bodily "resurrection." After their departure from the prison of the body, souls that are fully purified ascend directly to the world of separate lights. The ones that are only partially purified or are evil souls escape to a "world of images" suspended below the higher lights and above the corporeal world. In this world of images, or forms (not to be confused with the Platonic forms, which as-Suhrawardī identifies with higher and permanent intelligible lights), partially purified souls remain suspended and are able to create for themselves and by their own power of imagination pleasing figures and desirable objects in forms more excellent than their earthly counterparts and are able to enjoy them forever. Evil souls become dark shadows, suffer (presumably because their corrupt and inefficient power of imagination can create only ugly and frightening forms), and wander about as ghosts, demons, and devils. The creative power of the imagination, which as a human psychological phenomenon was already used by the philosophers to explain prophetic powers, was seized upon by the new wisdom as "divine magic." It was used to construct an eschatology, to explain miracles, dreams, and other saintly theurgic (healing) practices, to facilitate the movement between various orders of being, and for literary purposes.

The teachings of Ibn al-ʿArabī. The account of the doctrines of Ibn al-ʿArabī (12th–13th centuries) belongs properly to the history of Islāmic mysticism. Yet his impact on the subsequent development of the new wisdom was in many ways far greater than was that of as-Suhrawardī. This is true especially of his central doctrine of the "unity of being" and his sharp distinction between the absolute One, which is undefinable Truth (*ḥaqq*), and his self-manifestation (*ẓuhūr*), or creation (*khalq*), which is ever new (*jadīd*) and in perpetual movement, a movement that unites the whole of creation in a process of constant renewal. At the very core of this dynamic edifice stands nature, the "dark cloud" (*ʿamā*) or "mist" (*bukhār*), as the ultimate principle of things and forms: intelligence, heavenly bodies, and elements and their mixtures that culminate in the "perfect man." This primordial nature is the "breath" of the Merciful God in his aspect as Lord. It "flows" throughout the universe and manifests Truth in all its parts. It is the first mother through which Truth manifests itself to itself and generates the universe. And it is the universal natural body that gives birth to the translucent bodies of the spheres, to the elements, and to their mixtures, all of which are related to that primary source as daughters to their mother.

Doctrine of the "unity of being"

Ibn al-ʿArabī attempted to explain how Intelligence proceeds from the absolute One by inserting between them a primordial feminine principle, which is all things in potentiality but which also possesses the capacity, readiness, and desire to manifest or generate them, first, as archetypes in Intelligence, and then as actually existing things in the universe below. Ibn al-ʿArabī gave this principle numerous names, including prime "matter" (*ʿunṣur*), and characterized it as the principle "whose existence makes manifest the essences of the potential worlds." The doctrine that the first simple originated thing is not Intelligence but "indefinite matter" and that Intelligence was originated through the mediation of this matter was attributed to Empedocles, 5th-century-BC Greek philosopher, in doxographies (compilations of extracts from the Greek philosophers) translated into Arabic. It represented an attempt to bridge the gulf between the absolute One and the multiplicity of forms in Intelligence. The Andalusian mystic Ibn Masarrah (9th–10th centuries) is reported to have championed pseudo-Empedoclean doctrines, and Ibn al-ʿArabī (who studied under some of his followers) quotes Ibn Masarrah on a number of occasions. This philosophic tradition is distinct from the one followed by the Ismāʿīlī theologians, who explained the origination of Intelligence by the mediation of God's will.

The teachings of Twelver Shīʿism and the school of Isfahan. After Ibn al-ʿArabī, the new wisdom developed rapidly in intellectual circles in eastern Islām. Commentators on the works of Avicenna, as-Suhrawardī, and Ibn al-ʿArabī began the process of harmonizing and integrating the views of the masters. Great poets made them part of every educated man's literary culture. Mystical fraternities became the custodians of such works, spreading them into Central Asia and the Indian subcontinent and transmitting them from one generation to another. Following the Mongol khan Hülagü's entry into Baghdad (1258), the Twelver Shīʿah were encouraged by the Il Khanid Tatars and Naṣīr ad-Dīn aṭ-Ṭūsī (the philosopher and theologian who accompanied Hülagü as his vizier) to abandon their hostility to mysticism. Muʿtazilī doctrines were retained in their theology. Theology, however, was downgraded to "formal" learning that must be supplemented by higher things, the latter including philosophy and mysticism, both of earlier Shīʿī (including Ismāʿīlī) origin and of later Sunnī provenance. Al-Ghazālī, as-Suhrawardī, Ibn al-ʿArabī, and Avicenna were then eagerly studied and (except for their doctrine of the imamate) embraced with little or no reservation. This movement in Shīʿī thought gathered momentum when the leaders of a mystical fraternity established themselves as the Ṣafavid dynasty (1501–1732) in Iran, where they championed Twelver Shīʿism as the official doctrine of the new monarchy. During the 17th century, Iran experienced a cultural and scientific renaissance that included a revival of philosophic studies. There, Islāmic philosophy found its last creative exponents. The new wisdom as expounded by the masters of the school of Isfahan radiated throughout eastern Islām and continued as a vital tradition until modern times.

The role of mystical fraternities in the east

The major figures of the school of Isfahan were Mīr Dāmād (Muḥammad Bāqir ibn ad-Dāmād, died 1631/32) and his great disciple Mullā Ṣadrā (Ṣadr ad-Dīn ash-Shīrāzī, *c.* 1571–1640). Both were men of wide culture and prolific writers with a sharp sense for the history and development of philosophic ideas.

The teachings of Mīr Dāmād. Mīr Dāmād was the first to expound the notion of "eternal origination" (*ḥudūth dahrī*) as an explanation for the creation of the world. Muslim philosophers and their critics had recognized the crucial role played by the question of time in the discussion of the eternity of the world. The proposition that time is the measure of movement was criticized by Abū al-Barakāt al-Baghdādī, who argued that time is prior to movement and rest, indeed to everything except being. Time is the measure or concomitant of being, lasting and transient, enduring and in movement or rest. It character-

The concept of "eternal origination"

izes or qualifies all being, including God. God works in time, incessantly willing and directly creating everything in the world: his persistent will creates the eternal beings of the world, and his ever-renewed will creates the transient beings. The notion of a God who works in time was of course objectionable to theology, and Fakhr ad-Dīn ar-Rāzī refused to accept this solution despite its attractions. Ar-Rāzī also saw that it leads to the notion (attributed to Plato) that time is a self-subsistent substance, whose relation to God would further compromise his unity. Finally, ar-Rāzī explained that this self-subsistent substance will have to be related to different beings in different ways. It is called "everlastingness" (*sarmad*) when related to God and the Intelligences (angels) that are permanent and do not move or change in any way, "eternity" (*dahr*) when related to the totality of the world of movement and change, and "time" (*zamān*) when related to corporeal beings that make up the world of movement and change.

The problems of essence and existence

Mīr Dāmād returned to Avicenna and sought to harmonize his views with those of as-Suhrawardī on the assumption that what Avicenna meant by his "Oriental" (*mashriqīyah*) philosophy was identical with as-Suhrawardī's wisdom of "illumination" (*ishrāq*), which he interpreted as a Platonic doctrine that asserted the priority of essence (form) over being (existence). Time, for Mīr Dāmād, was neither a mere being of reason nor an accident of existing things. It belongs to the essence of things and describes their mode and rank of being. It is a "relation" that beings have to each other because of their essential nature. There must, therefore, be three ranks of order of time corresponding to the three ranks of order of being. Considered as the relation of God to the divine names and attributes (Intelligences or archetypes), the relation is "everlastingness." Considered as the relation between the Intelligences, or archetypes, and their reflections in the mutable things of the world below, the relation is "eternity." And considered as the relation between these mutable things, the relation is "time." Creation, or origination, is this very relation. Thus, the origination of the immutable Intelligences, or archetypes, is called "everlasting creation," the origination of the world of mutable beings as a whole is called "eternal creation," and the generation of mutable things within the world is called "temporal creation."

The teachings of Mullā Ṣadrā. Mullā Ṣadrā superimposed Ibn al-ʿArabī's mystical thought (whose philosophic implications had already been exposed by a number of commentators) on the "Aristotelian"-Illuminationist synthesis developed by Mīr Dāmād. Against his master, he argued with the Aristotelians for the priority of being (existence) over essence (form), which he called an abstraction; and, with Ibn al-ʿArabī, he argued for the "unity of being" within which beings differ only according to "priority and posteriority," "perfection and imperfection," and "strength and weakness." All being is thus viewed as a graded manifestation, or determination, of absolute, or pure, Being, and every level of being possesses all the attributes of pure Being, but with varying degrees of intensity or perfection.

Mullā Ṣadrā's doctrine of nature

Mullā Ṣadrā considered his unique contribution to Islāmic philosophy to be his doctrine of nature, which enabled him to assert that everything other than God and his knowledge—*i.e.,* the entire corporeal world, including the heavenly bodies—is originated "eternally" as well as "temporally." This doctrine of nature is an elaboration of the last manifestation of Ibn al-ʿArabī's "nature" or prime "matter," articulated on philosophic grounds and within the general framework of Aristotelian natural science and defended against every possible philosophic and theological objection.

Nature for Mullā Ṣadrā is the "substance" and "power" of all corporeal beings and the direct cause of their movement. Movement (and time, which measures it) is therefore not an accident of substance or an accompaniment of some of its accidents. It signifies the very change, renewal, and passing of being—itself being in constant "flow," or flux. The entire corporeal world, both the celestial spheres and the world of the elements, constantly renews itself. The "matter" of corporeal things has the power to become a new form at every instant; and the resulting matter–form complex is at every instant a new matter ready for, desiring, and moving toward another form. Men fail to observe this constant flux and movement in simple bodies not because of the endurance of the same form in them but because of the close similarity between their ever-new forms. What the philosophers call "movement" and "time" are not, as they believed, anchored in anything permanent—*e.g.,* in what they call "nature," "substance," or "essence"; essence is permanent only in the mind, and nature and substance are permanent activity. Nature as permanent activity is the very being of natural things and identical with their substance. Because nature is "permanent" in this sense, it is connected to a permanent principle that manifests activity in it permanently. Because nature constantly renews itself, all renewed and emergent things are connected to it. Thus, nature is the link between what is eternal and what is originated, and the world of nature is originated both eternally and temporarily.

The problem of movement, change, or renewal

Mullā Ṣadrā distinguishes this primary "movement-in-substance" (*al-ḥarakah fī al-jawhar*) from haphazard, compulsory, and other accidental movements that lack proper direction, impede the natural movement of substance, or reverse it. Movement-in-substance is not universal change or flux without direction, the product of conflict between two equally powerful principles, or a reflection of the nonbeing of the world of nature when measured against the world of permanent forms. It is, rather, the natural beings' innate desire to become more perfect, which directs this ceaseless self-renewal, self-origination, or self-emergence into a perpetual and irreversible flow upward in the scale of being—from the simplest elements to the human body–soul complex and the heavenly body–soul complex (both of which participate in the general instability, origination, and passing of being that characterizes the entire corporeal world). This flow upward, however, is by no means the end. For the indefinite "matter" (Ibn al-ʿArabī's "cloud" and the mystics' "created Truth") is the "substratum" of everything other than its Creator, the mysterious pure Truth. It "extends" beyond the body–soul complex to the Intelligences (divine names) that are Being's first, highest, and purest actualization or activity. This "extension" unites everything other than the Creator into a single continuum. The human body–soul complex and the heavenly body–soul complex are not moved externally by the Intelligences. Their movement is an extension of the process of self-perfection. Having reached the highest rank of order of substance in the corporeal world, they are now prepared, and still moved by their innate desire, to flow upward and transform themselves into pure intelligence.

IMPACT OF MODERNISM

The new wisdom lived on during the 18th and 19th centuries, conserving much of its vitality and strength but not cultivating new ground. It attracted able thinkers such as Shāh Walī Allāh of Delhi and Hādī Sabzevārī and became a regular part of the program of higher education in the cultural centres of the Ottoman Empire, Iran, and the Indian subcontinent, a status never achieved by the earlier tradition of Islāmic philosophy. In collaboration with its close ally Persian mystical poetry, the new wisdom determined the intellectual outlook and spiritual mood of educated Muslims in the regions where Persian had become the dominant literary language.

The problem of reform in modern Islāmic thought

The wholesale rejection of the new wisdom in the name of simple, robust, and more practical piety (which had been initiated by Ibn Taymīyah and which continued to find exponents among jurists) made little impression on its devotees. To be taken seriously, reform had to come from their own ranks and be espoused by such thinkers as the eminent theologian and mystic of Muslim India Aḥmad Sirhindī (flourished 16th–17th centuries)—a reformer who spoke their language and attacked Ibn al-ʿArabī's "unity of being" only to defend an older, presumably more orthodox form of mysticism. Despite some impact, however, attempts of this kind remained isolated and were either ignored or reintegrated into the mainstream, until the coming of the modern reformers. The 19th- and 20th-century reformers Jamāl ad-Dīn al-Afghānī, Muḥammad

Abduh, and Muḥammad Iqbāl were initially educated in this tradition, but they rebelled against it and advocated radical reforms.

The modernists attacked the new wisdom at its weakest point; that is, its social and political norms, its individualistic ethics, and its inability to speak intelligently about social, cultural, and political problems generated by a long period of intellectual isolation that was further complicated by the domination of the European powers. Unlike the earlier tradition of Islāmic philosophy from al-Fārābī to Averroës, which had consciously cultivated political science and investigated the political dimension of philosophy and religion and the relation between philosophy and the community at large, the new wisdom from its inception lacked genuine interest in these questions, had no appreciation for political philosophy, and had only a benign toleration for the affairs of the world.

None of the reformers was a great political philosopher. They were concerned with reviving their nations' latent energies, urging them to free themselves from foreign domination, and impressing on them the need to reform their social and educational institutions. They also saw that all this required a total reorientation, which could not take place so long as the new wisdom remained not only the highest aim of a few solitary individuals but also a social and popular ideal as well. Yet, as late as 1917, Iqbāl found that "the present-day Muslim prefers to roam about aimlessly in the valley of Hellenic-Persian mysticism, which teaches us to shut our eyes to the hard reality around, and to fix our gaze on what is described as 'illumination.' " His reaction was harsh: "To me this self-mystification, this nihilism, *i.e.,* seeking reality where it does not exist, is a physiological symptom, giving me a clue to the decadence of the Muslim world."

The call to return to the early masters of Islāmic theology and philosophy

To arrest the decadence and infuse new vitality in a society in which they were convinced religion must remain the focal point, the modern reformers advocated a return to the movements and masters of Islāmic theology and philosophy antedating the new wisdom. They argued that these, rather than the "Persian incrustation of Islām," represented Islām's original and creative impulse. The modernists were attracted, in particular, to the views of the Muʿtazilah: affirmation of God's unity and denial of all similarity between him and created things; reliance on human reason; emphasis on man's freedom; faith in man's ability to distinguish between good and bad; and insistence on man's responsibility to do good and fight against evil in private and public places. They were also impressed by the traditionalists' devotion to the original, uncomplicated forms of Islām and by their fighting spirit, and by the Ashʿarīs' view of faith as an affair of the heart and their spirited defense of the Muslim community. In viewing the scientific and philosophic tradition of eastern and western Islām prior to the Tatar and Mongol invasions, they saw an irrefutable proof that true Islām stands for the liberation of man's spirit, promotes critical thought, and provides both the impetus to grapple with the temporal and the demonstration of how to set it in order. These ideas initiated what was to become a vast effort to recover, edit, and translate into the Muslim national languages works of earlier theologians and philosophers, which had been long neglected or known only indirectly through later accounts.

The role of universities in the reform of Islāmic theology and philosophy

The modern reformers insisted, finally, that Muslims must be taught to understand the real meaning of what has happened in Europe, which in effect means the understanding of modern science and philosophy, including modern social and political philosophies. Initially, this challenge became the task of the new universities in the Muslim world. In the latter part of the 20th century, however, the originally wide gap between the various programs of theological and philosophic studies in religious colleges and in modern universities narrowed considerably.

(M.S.M./Ed.)

THE CULTURE OF ISLĀM

Islāmic Law, Sharīʿah

Total and unqualified submission to the will of Allāh (God) is the fundamental tenet of Islām: Islāmic law is therefore the expression of Allāh's command for Muslim society and, in application, constitutes a system of duties that are incumbent upon a Muslim by virtue of his religious belief. Known as the Sharīʿah (literally, "the path leading to the watering place"), the law constitutes a divinely ordained path of conduct that guides the Muslim toward a practical expression of his religious conviction in this world and the goal of divine favour in the world to come.

NATURE AND SIGNIFICANCE OF ISLĀMIC LAW

Muslim jurisprudence, the science of ascertaining the precise terms of the Sharīʿah, is known as *fiqh* (literally "understanding"). The historical process of the discovery of Allāh's law (see below) was regarded as completed by the end of the 9th century when the law had achieved a definitive formulation in a number of legal manuals written by different jurists. Throughout the medieval period this basic doctrine was elaborated and systematized in a large number of commentaries, and the voluminous literature thus produced constitutes the traditional textual authority of Sharīʿah law.

Scope of Sharīʿah

In classical form the Sharīʿah differs from Western systems of law in two principal respects. In the first place the scope of the Sharīʿah is much wider, since it regulates man's relationship not only with his neighbours and with the state, which is the limit of most other legal systems, but also with his God and his own conscience. Ritual practices, such as the daily prayers, almsgiving, fasting, and pilgrimage, are an integral part of Sharīʿah law and usually occupy the first chapters in the legal manuals. The Sharīʿah is also concerned as much with ethical standards as with legal rules, indicating not only what man is entitled or bound to do in law, but also what he ought, in conscience, to do or refrain from doing. Accordingly, certain acts are classified as praiseworthy (*mandūb*), which means that their performance brings divine favour and their omission divine disfavour, and others as blameworthy (*makrūh*), which means that omission brings divine favour and commission divine disfavour; but in neither case is there any legal sanction of punishment or reward, nullity or validity. The Sharīʿah is not merely a system of law, but a comprehensive code of behaviour that embraces both private and public activities.

The second major distinction between the Sharīʿah and Western legal systems is the result of the Islāmic concept of law as the expression of the divine will. With the death of the Prophet Muḥammad in 632, communication of the divine will to man ceased so that the terms of the divine revelation were henceforth fixed and immutable. When, therefore, the process of interpretation and expansion of this source material was held to be complete with the crystallization of the doctrine in the medieval legal manuals, Sharīʿah law became a rigid and static system. Unlike secular legal systems that grow out of society and change with the changing circumstances of society, Sharīʿah law was imposed upon society from above. In Islāmic jurisprudence it is not society that moulds and fashions the law, but the law that precedes and controls society.

Such a philosophy of law clearly poses fundamental problems of principle for social advancement in contemporary Islām. How can the traditional Sharīʿah law be adapted to meet the changing circumstances of modern Muslim society? This is now the central issue in Islāmic law. (See below *Reform of Sharīʿah law*).

HISTORICAL DEVELOPMENT OF SHARĪʿAH LAW

For the first Muslim community established under the leadership of the Prophet at Medina in 622, the Qurʾānic revelations laid down basic standards of conduct. But the Qurʾān is in no sense a comprehensive legal code. No more than 80 verses deal with strictly legal matters; while these verses cover a wide variety of topics and introduce many novel rules, their general effect is simply to modify the existing Arabian customary law in certain important particulars.

During his lifetime Muḥammad, as the supreme judge of the community, resolved legal problems as they arose by interpreting and expanding the general provisions of the Qurʾān, and the same *ad hoc* activity was carried on after his death by the caliphs (temporal and spiritual rulers) of Medina. But the foundation of the Umayyad dynasty in 661, governing from its centre of Damascus a vast military empire, produced a legal development of much broader dimensions. With the appointment of judges, or *qāḍīs,* to the various provinces and districts, an organized judiciary came into being. The *qāḍīs* were responsible for giving effect to a growing corpus of Umayyad administrative and fiscal law; and since they regarded themselves essentially as the spokesmen of the local law, elements and institutions of Roman-Byzantine and Persian-Sāsānian law were absorbed into Islāmic legal practice in the conquered territories. Depending upon the discretion of the individual *qāḍī,* decisions would be based upon the rules of the Qurʾān where these were relevant; but the sharp focus in which the Qurʾānic laws were held in the Medinian period had become lost with the expanding horizons of activity.

Development of a judiciary

Development of different schools of law. A reaction to this situation arose in the early 8th century when pious scholars, grouped together in loose, studious fraternities, began to debate whether or not Umayyad legal practice was properly implementing the religious ethic of Islām. Actively sponsored by the ʿAbbāsid rulers, who came to power in the mid-8th century pledged to build a truly Islāmic state and society, the activities of the jurists (*faqīh,* plural *fuqahāʾ*) in these early schools of law marked the real beginning of Islāmic jurisprudence. Their aim was to Islāmize the law by reviewing the current legal practice in the light of the Qurʾānic principles and then on this basis adopting, modifying, or rejecting the practice as part of their ideal scheme of law.

Of the many early schools of law the two most important were those of the Mālikīs in Medina and the Ḥanafīs in al-Kūfah, named after two outstanding scholars in the respective localities, Mālik ibn Anas and Abū Ḥanīfah. Inevitably the Mālikī and Ḥanafī doctrines, as they were then being recorded in the first compendiums of law, differed considerably from each other, not only because free juristic speculation was bound to produce varying results but also because the thought of the scholars was conditioned by their different social environments. A deep conflict of juristic principle emerged within the schools between those who maintained that outside the terms of the Qurʾān scholars were free to use their reason (*raʾy*) to ascertain the law and those who insisted that the only valid source of law outside the Qurʾān lay in the precedents set by the Prophet himself.

The jurist ash-Shāfiʿī (died 820) aimed to eliminate these schisms and produce greater uniformity in the law by expounding a firm theory of the sources from which the law must be derived. Ash-Shāfiʿī's fundamental teaching was that knowledge of the Sharīʿah could be attained only through divine revelation found either in the Qurʾān or in the divinely inspired traditions (*sunnah*) of the Prophet as ascertained through authentic reports (Ḥadīth). Human reason in law should be strictly confined to the process of analogical deduction, or *qiyās*—problems not specifically answered by the divine revelation were to be solved by applying the principles upon which closely parallel cases had been regulated by the Qurʾān or *sunnah.*

The role of divinely inspired traditions

Shāfiʿī's insistence upon the importance of the *sunnah* as a source of law produced a great activity in the collection and classification of Ḥadīths, particularly among his own supporters, who formed the Shāfiʿī school, and the followers of Aḥmad ibn Ḥanbal (died 855) who formed the Ḥanbalī school. Muslim scholarship maintained that the classical compilations of Ḥadīths—especially those of Bukhārī (died 870) and Muslim (died 875)—constituted an authentic record of the Prophet's precedents. The general view of Western orientalists, however, is that a considerable part of the *sunnah* represents the views of later jurists fictitiously ascribed to the Prophet to give the doctrine a greater authority.

Later developments. Shāfiʿī's thesis formed the basis of the classical theory of the roots of jurisprudence (*uṣūl al-fiqh*), which crystallized in the early 10th century. Juristic "effort" to comprehend the terms of the Sharīʿah is known as *ijtihād,* and legal theory first defines the course that *ijtihād* must follow. In seeking the answer to a legal problem the jurist must first consult the Qurʾān and the *sunnah.* Failing any specific solution in this divine revelation he must employ analogy (*qiyās*) or certain subsidiary principles of reasoning—*istiḥsān* (equitable preference) and *istiṣlāḥ* (the public interest). The legal theory then evaluates the results of *ijtihād* on the basis of the criterion of *ijmāʿ* (consensus). As an attempt to define Allāh's law, the *ijtihād* of individual scholars could result only in a tentative conclusion termed *ẓann* ("conjecture"). But where a conclusion became the subject of unanimous agreement by the qualified scholars, it became a certain (*yaqīn*) and infallible expression of Allāh's law.

Two major effects flowed from this classical doctrine of *ijmāʿ*. It served first as a permissive principle to admit the validity of variant opinions as equally probable attempts to define the Sharīʿah. Second, it operated as a restrictive principle to ratify the status quo; for once the *ijmāʿ* had cast an umbrella authority not only over those points that were the subject of a consensus but also over existing variant opinions, to propound any further variant was to contradict the infallible *ijmāʿ* and therefore tantamount to heresy.

Ijmāʿ set the final seal of rigidity upon the doctrine, and from the 10th century onward independent juristic speculation ceased. In the Arabic expression, "the door of *ijtihād* was closed." Henceforth jurists were *muqallids*, or imitators, bound by the doctrine of *taqlīd* ("clothing with authority" *i.e.,* unquestioned acceptance) to follow the doctrine as it was recorded in the authoritative legal manuals.

Sharīʿah law is a candidly pluralistic system, the philosophy of the equal authority of the different schools being expressed in the alleged dictum of the Prophet: "Difference of opinion among my community is a sign of the bounty of Allāh." But outside the four schools of Sunnī, or orthodox, Islām stand the minority sects of the Shīʿah and the Ibāḍīs whose own versions of the Sharīʿah differ considerably from those of the Sunnīs. Shīʿī law in particular grew out of a fundamentally different politico-religious system in which the rulers, or *imāms*, were held to be divinely inspired and therefore the spokesmen of the Lawgiver himself. Geographically, the division between the various schools and sects became fairly well defined as the *qāḍīs*' courts in different areas became wedded to the doctrine of one particular school. Thus Ḥanafī law came to predominate in the Middle East and the Indian subcontinent; Mālikī law in North, West, and Central Africa; Shāfiʿī law in East Africa, the southern parts of the Arabian peninsula, Malaysia, and Indonesia; Ḥanbalī law in Saudi Arabia, Shīʿī law in Iran and the Shīʿī communities of India and East Africa; Ibāḍī law in Zanzibar, ʿUman, and parts of Algeria.

Distribution of various schools of Islāmic law

Although Sharīʿah doctrine was all-embracing, Islāmic legal practice has always recognized jurisdictions other than that of the *qāḍīs*. Because the *qāḍis*' courts were hidebound by a cumbersome system of procedure and evidence, they did not prove a satisfactory organ for the administration of justice in all respects, particularly as regards criminal, land, and commercial law. Hence, under the broad head of the sovereign's administrative power (*siyāsah*), competence in these spheres was granted to other courts, known collectively as *maẓālim* courts, and the jurisdiction of the *qāḍīs* was generally confined to private family and civil law. As the expression of a religious ideal, Sharīʿah doctrine was always the focal point of legal activity, but it never formed a complete or exclusively authoritative expression of the laws that in practice governed the lives of Muslims.

THE SUBSTANCE OF TRADITIONAL SHARĪʿAH LAW

Sharīʿah duties are broadly divided into those that an individual owes to Allāh (the ritual practices or *ʿibādāt*) and those that he owes to his fellow men (*muʿāmalāt*). It is the latter category of duties alone, constituting law in the Western sense, that is described here.

Penal law. Offenses against the person, from homicide to assault, are punishable by retaliation (*qiṣāṣ*), the offender being subject to precisely the same treatment as his victim. But this type of offense is regarded as a civil injury rather than a crime in the technical sense, since it is not the state but only the victim or his family who have the right to prosecute and to opt for compensation or blood money (*diyah*) in place of retaliation.

Fixed punishment

For six specific crimes the punishment is fixed (*ḥadd*): death for apostasy and for highway robbery; amputation of the hand for theft; death by stoning for extramarital sex relations (*zinā*) where the offender is a married person and 100 lashes for unmarried offenders; 80 lashes for an unproved accusation of unchastity (*qadhf*) and for the drinking of any intoxicant.

Outside the *ḥadd* crimes, both the determination of offenses and the punishment therefore lies with the discretion of the executive or the courts.

Law of transactions. A legal capacity to transact belongs to any person "of prudent judgment" (*rāshid*), a quality that is normally deemed to arrive with physical maturity or puberty. There is an irrebuttable presumption of law (1) that boys below the age of 12 and girls below the age of 9 have not attained puberty, and (2) that puberty has been attained by the age of 15 for both sexes. Persons who are not *rāshid,* on account of minority, mental deficiency, simplicity, or prodigality, are placed under interdiction: their affairs are managed by a guardian and they cannot transact effectively without the guardian's consent.

The basic principles of the law are laid down in the four root transactions of (1) sale (*bayʿ*), transfer of the ownership or corpus of property for a consideration; (2) hire (*ijārah*), transfer of the usufruct (right to use) of property for a consideration; (3) gift (*hibah*), gratuitous transfer of the corpus of property, and (4) loan (*ʿāriyah*), gratuitous transfer of the usufruct of property. These basic principles are then applied to the various specific transactions of, for example, pledge, deposit, guarantee, agency, assignment, land tenancy, partnership, and *waqf* foundations. *Waqf* is a peculiarly Islāmic institution whereby the founder relinquishes his ownership of real property, which belongs henceforth to Allāh, and dedicates the income or usufruct of the property in perpetuity to some pious or charitable purpose, which may include settlements in favour of the founder's own family.

The Islāmic law of transactions as a whole is dominated by the doctrine of *ribā.* Basically, this is the prohibition of usury, but the notion of *ribā* was rigorously extended to cover, and therefore preclude, any form of interest on a capital loan or investment. And since this doctrine was coupled with the general prohibition on gambling transactions, Islāmic law does not, in general, permit any kind of speculative transaction the results of which, in terms of the material benefits accruing to the parties, cannot be precisely forecast.

Family law. A patriarchal outlook is the basis of the traditional Islāmic law of family relationships. Fathers have the right to contract their daughters, whether minor or adult, in compulsory marriage. Only when a woman has been married before is her consent to her marriage necessary; but even then the father, or other marriage guardian, must conclude the contract on her behalf. In Ḥanafī and Shīʿī law, however, only minor girls may be contracted in compulsory marriage, and adult women may conclude their own marriage contracts, except that the guardian may have the marriage annulled if his ward has married beneath her social status.

Husbands have the right of polygamy and may be validly married at the same time to a maximum of four wives. Upon marriage a husband is obliged to pay to his wife her dower, the amount of which may be fixed by agreement or by custom; and during the marriage he is bound to maintain and support her provided she is obedient to him, not only in domestic matters but also in her general social activities and conduct. A wife who rejects her husband's dominion by leaving the family home without just cause forfeits her right to maintenance.

Law of divorce

But it is in the traditional law of divorce that the scales are most heavily weighted against the wife. A divorce may be effected simply by the mutual agreement of the spouses, which is known as *khulʿ* when the wife pays some financial consideration to the husband for her release; and according to all schools except the Ḥanafīs a wife may obtain a judicial decree of divorce on the ground of some matrimonial offense—*e.g.,* cruelty, desertion, failure to maintain—committed by the husband. But the husband alone has the power unilaterally to terminate the marriage by repudiation (*ṭalāq*) of his wife. *Ṭalāq* is an extrajudicial process: a husband may repudiate his wife at will and his motive in doing so is not subject to scrutiny by the court or any other official body. A repudiation repeated three times constitutes a final and irrevocable dissolution of the marriage; but a single pronouncement may be revoked at will by the husband during the period known as the wife's *ʿiddah,* which lasts for three months following the repudiation (or any other type of divorce) or, where the wife is pregnant, until the birth of the child.

The legal position of children within the family group, as regards their guardianship, maintenance, and rights of succession, depends upon their legitimacy, and a child is legitimate only if it is conceived during the lawful wedlock of its parents. In Sunnī law no legal relationship exists between a father and his illegitimate child; but there is a legal tie, for all purposes, between a mother and her illegitimate child. Guardianship of the person (*e.g.,* control of education and marriage) and of the property of minor children belongs to the father or other close male, agnate relative, but the bare right of custody (*ḥaḍānah*) of young children, whose parents are divorced or separated, belongs to the mother or the female, maternal relatives.

Succession law. An individual's power of testamentary disposition is basically limited to one-third of his net estate (*i.e.,* the assets remaining after the payment of funeral expenses and debts) and two-thirds of the estate passes to the legal heirs of the deceased under the compulsory rules of inheritance.

Sunnī inheritance laws

There is a fundamental divergence between the Sunnī and the Shīʿī schemes of inheritance. Sunnī law is essentially a system of inheritance by male agnate relatives or *ʿaṣabah—i.e.,* relatives who, if they are more than one degree removed from the deceased, trace their connection with him through male links. Among the *ʿaṣabah,* priority is determined by: (1) class, descendants excluding ascendants, who in turn exclude brothers and their issue, who in turn exclude uncles and their issue; (2) degree, within each class the relative nearer in degree to the deceased excluding the more remote; (3) strength of blood tie, the germane, or full blood, connection excluding the half blood, or consanguine, connection among collateral relatives. This agnatic system is mitigated by allowing the surviving spouse and a limited number of females and nonagnates—the daughter; son's daughter; mother; grandmother; germane, consanguine, and uterine sisters; and uterine brother—to inherit a fixed fractional portion of the estate in suitable circumstances. But the females among these relatives only take half the share of the male relative of the same class, degree, and blood tie, and none of them excludes from inheritance any male agnate, however remote. No other female or non-agnatic relative has any right of inheritance in the presence of a male agnate. Where, for example, the deceased is survived by his wife, his daughter's son, and a distant agnatic cousin, the wife will be restricted to one-fourth of the inheritance, the grandson will be excluded altogether, and the cousin will inherit three-fourths of the estate.

Shīʿī law rejects the criterion of the agnatic tie and regards both the maternal and paternal connections as equally strong grounds of inheritance. In the Shīʿī system the surviving spouse always inherits a fixed portion, as in Sunnī law, but all other relatives, including females and nonagnates, are divided into three classes: (1) parents and lineal descendants; (2) grandparents, brothers and sisters, and their issue; (3) uncles and aunts and their issue. Any relative of class one excludes any relative of class two, who in turn excludes any relative of class three. Within each class the nearer in degree excludes the more remote, and the full blood excludes the half blood. While, therefore, a male relative normally takes double the share of the corresponding female relative, females and nonagnates are much more favourably treated than they are in Sunnī law. In the case mentioned above, for example, the wife would take one-fourth, but the remaining three-fourths would go to the daughter's son, or indeed to a daughter's daughter, and not to the agnatic cousin.

Under Shīʿī law the only restriction upon testamentary power is the one-third rule, but Sunnī law goes further and does not allow any bequest in favour of a legal heir. Under both systems, however, bequests that infringe these rules are not necessarily void and ineffective; the testator has acted beyond his powers, but the bequest may be ratified by his legal heirs.

Further protection is afforded to the rights of the legal heirs by the doctrine of death sickness. Any gifts made by a dying person in contemplation of his death are subject to precisely the same limitations as bequests, and, if they exceed these limits, will be effective only with the consent of the legal heirs.

Procedure and evidence. Traditionally, Sharīʿah law was administered by the court of a single *qāḍī*, who was the judge of the facts as well as the law, although on difficult legal issues he might seek the advice of a professional jurist, or *muftī*. There was no hierarchy of courts and no organized system of appeals. Through his clerk (*kātib*) the *qāḍī* controlled his court procedure, which was normally characterized by a lack of ceremony or sophistication. Legal representation was not unknown, but the parties would usually appear in person and address their pleas orally to the *qāḍī*.

Burden of proof

The first task of the *qāḍī* was to decide which party bore the burden of proof. This was not necessarily the party who brought the suit, but was the party whose contention was contrary to the initial legal presumption attaching to the case. In the case of an alleged criminal offense, for example, the presumption is the innocence of the accused, and in a suit for debt the presumption is that the alleged debtor is free from debt. Hence the burden of proof would rest upon the prosecution in the first case and upon the claiming creditor in the second. This burden of proof might, of course, shift between the parties several times in the course of the same suit, as, for example, where an alleged debtor pleads a counterclaim against the creditor.

The standard of proof required, whether on an initial, intermediate or final issue, was a rigid one and basically the same in both criminal and civil cases. Failing a confession or admission by the defendant, the plaintiff or prosecutor was required to produce two witnesses to testify orally to their direct knowledge of the truth of his contention. Written evidence and circumstantial evidence, even of the most compelling kind, were normally inadmissible. Moreover, the oral testimony (*shahādah*) had usually to be given by two male, adult Muslims of established integrity or character. In certain cases, however, the testimony of women was acceptable (two women being required in place of one man), and in most claims of property the plaintiff could satisfy the burden of proof by one witness and his own solemn oath as to the truth of his claim.

If the plaintiff or prosecutor produced the required degree of proof, judgment would be given in his favour. If he failed to produce any substantial evidence at all, judgment would be given for the defendant. If he produced some evidence, but the evidence did not fulfill the strict requirements of *shahādah*, the defendant would be offered the oath of denial. Properly sworn this oath would secure judgment in his favour; but if he refused it, judgment would be given for the plaintiff, provided, in some cases, that the latter himself would swear an oath.

In sum, the traditional system of procedure was largely self-operating. After his initial decision as to the incidence of the burden of proof, the *qāḍī* merely presided over the predetermined process of the law: witnesses were or were not produced, the oath was or was not administered and sworn, and the verdict followed automatically.

LAW IN CONTEMPORARY ISLĀM

The scope of Sharīʿah law and the mode of its administration. During the 19th century the impact of Western civilization upon Muslim society brought about radical changes in the fields of civil and commercial transactions and criminal law. In these matters the Sharīʿah courts were felt to be wholly out of touch with the needs of the time, not only because of their system of procedure and evidence but also because of the substance of the Sharīʿah doctrine, which they were bound to apply.

As a result, the criminal and general civil law of the Sharīʿah was abandoned in most Muslim countries and replaced by new codes based upon European models with a new system of secular tribunals to apply them. Thus, with the notable exception of the Arabian peninsula, where the Sharīʿah is still formally applied in its entirety, the application of Sharīʿah law in Islām has been broadly confined, from the beginning of the 20th century, to family law, including the law of succession at death and the particular institution of *waqf* endowments.

The decline of Sharīʿah

Nor, even within this circumscribed sphere, is Sharīʿah law today applied in the traditional manner. Throughout the Middle East generally Sharīʿah family law is now expressed in the form of modern codes, and it is only in the absence of a specific relevant provision of the code that recourse is had to the traditionally authoritative legal manuals. In India and Pakistan much of the family law is now embodied in statutory legislation, and since the law is there administered as a case-law system, the authority of judicial decisions has superseded that of the legal manuals.

In most countries, too, the court system has been, or is being, reorganized to include, for instance, the provision of appellate jurisdictions. In Egypt and Tunisia the Sharīʿah courts, as a separate entity, have been abolished, and Sharīʿah law is now administered through a unified system of national courts. In India, and, since partition, in Pakistan it has always been the case that Sharīʿah law has been applied by the same courts that apply the general civil and criminal law.

Finally, in many countries, special codes have been enacted to regulate the procedure and evidence of the courts that today apply Sharīʿah law. In the Middle East documentary and circumstantial evidence are now generally admissible; witnesses are put on oath and may be cross-examined, and the traditional rule that evidence is only brought by one side and that the other side, in suitable circumstances, takes the oath of denial has largely broken down. In sum, the court has a much wider discretion in assessing the weight of the evidence than it had under the traditional system of evidence. In India and Pakistan the courts apply the same rules of evidence to cases of Islāmic law as they do to civil cases generally. The system is basically English law, codified in the Indian Evidence Act, 1872.

Reform of Sharīʿah law. Traditional Islāmic family law reflected to a large extent the patriarchal scheme of Arabian tribal society in the early centuries of Islām. Not unnaturally certain institutions and standards of that law were felt to be out of line with the circumstances of Muslim society in the 20th century, particularly in urban areas where tribal ties had disintegrated and movements for the emancipation of women had arisen. At first this situation seemed to create the same apparent impasse between the changing circumstances of modern life and an allegedly immutable law that had caused the adoption of Western codes in civil and criminal matters. Hence, the only solution that seemed possible to Turkey in 1926 was the total abandonment of the Sharīʿah and the adoption of Swiss family law in its place. No other Muslim country, however, has as yet followed this example. Instead, traditional Sharīʿah law has been adapted in a variety of ways to meet present social needs.

From the outset the dominating issue in the Middle East has been the question of the juristic basis of reforms—*i.e.*, granted their social desirability, their justification in terms of Islāmic jurisprudential theory, so that the reforms appear as a new, but legitimate, version of the Sharīʿah.

In the early stages of the reform movement, the doctrine of *taqlīd* (unquestioning acceptance) was still formally observed and the juristic basis of reform lay in the doctrine of *siyāsah,* or "government," which allows the political authority (who, of course, has no legislative power in the real sense of the term) to make administrative regulations of two principal types.

Egyptian reforms

The first type concerns procedure and evidence and restricts the jurisdiction of the Sharīʿah courts in the sense that they are instructed not to entertain cases that do not fulfill defined evidential requirements. Thus, an Egyptian law was enacted in 1931 that no disputed claim of marriage was to be entertained where the marriage could not be proved by an official certificate of registration, and no such certificate could be issued if the bride was less than 16 or the bridegroom less than 18 years of age at the time of the contract. Accordingly the marriage of a minor contracted by the guardian was still perfectly valid but would not, if disputed, be the subject of judicial relief from the courts. In theory the doctrine of the traditional authorities was not contradicted, but in practice an attempt had been made to abolish the institution of child marriage. The second type of administrative regulation was a directive to the courts as to which particular rule among existing variants they were to apply. This directive allowed the political authority to choose from the views of the different schools and jurists the opinion that was deemed best suited to present social circumstances. For example, the traditional Ḥanafī law in force in Egypt did not allow a wife to petition for divorce on the ground of any matrimonial offense committed by the husband, a situation that caused great hardship to abandoned or ill-treated wives. Mālikī law, however, recognizes the wife's right to judicial dissolution of her marriage on grounds such as the husband's cruelty, failure to provide maintenance and support, and desertion. Accordingly, an Egyptian law of 1920 codified the Mālikī law as the law henceforth to be applied by the Sharīʿah courts.

By way of comparison, reform in the matters of child marriage and divorce was effected in the Indian subcontinent by statutory enactments that directly superseded the traditional Ḥanafī law. The Child Marriage Restraint Act, 1929, prohibited the marriage of girls below the age of 14 and boys below the age of 16 under pain of penalties; while the Dissolution of Muslim Marriages Act, 1939, modelled on the English Matrimonial Causes Acts, allowed a Ḥanafī wife to obtain judicial divorce on the standard grounds of cruelty, desertion, failure to maintain, etc.

In the Middle East, by the 1950s, the potential for legal reform under the principle of *siyāsah* had been exhausted. Since that time the basic doctrine of *taqlīd* has been challenged to an ever-increasing degree. On many points the law recorded in the medieval manuals, insofar as it represents the interpretations placed by the early jurists upon the Qurʾān and the *sunnah,* has been held no longer to have a paramount and exclusive authority. Contemporary jurisprudence has claimed the right to renounce those interpretations and to interpret for itself, independently and afresh in the light of modern social circumstances, the original texts of divine revelation: in short to reopen the door of *ijtihād* that had been in theory closed since the 10th century.

Syrian and Tunisian reforms

The developing use of *ijtihād* as a means of legal reform may be seen through a comparison of the terms of the Syrian law of Personal Status (1953) with those of the Tunisian Law of Personal Status (1957) in relation to the two subjects of polygamy and divorce by repudiation (*ṭalāq*).

As regards polygamy the Syrian reformers argued that the Qurʾān itself urges husbands not to take additional wives unless they are financially able to make proper provision for their maintenance and support. Classical jurists had construed this verse as a moral exhortation binding only on the husband's conscience. But the Syrian reformers maintained that it should be regarded as a positive legal condition precedent to the exercise of polygamy and enforced as such by the courts. This novel interpretation was then coupled with a normal administrative regulation that required the due registration of marriages after the permission of the court to marry had been obtained. The Syrian Law accordingly enacts: "The *qāḍī* may withhold permission for a man who is already married to marry a second wife, where it is established that he is not in a position to support them both." Far more extreme, however, is the approach of the Tunisian reformers. They argued that, in addition to a husband's financial ability to support a plurality of wives, the Qurʾān also required that co-wives should be treated with complete impartiality. This Qurʾānic injunction should also be construed, not simply as a moral exhortation, but as a legal condition precedent to polygamy, in the sense that no second marriage should be permissible unless and until adequate evidence was forthcoming that the wives would in fact be treated impartially. But under modern social and economic conditions such impartial treatment was a practical impossibility. And since the essential condition for polygamy could not be fulfilled the Tunisian Law briefly declares: "Polygamy is prohibited."

With regard to *ṭalāq* the Syrian law provided that a wife who had been repudiated without just cause might be awarded compensation by the court from her former husband to the maximum extent of one year's maintenance. The reform was once again represented as giving practical effect to certain Qurʾānic verses that had been generally regarded by traditional jurisprudence as moral rather than legally enforceable injunctions—namely, those verses that enjoin husbands to "make a fair provision" for repudiated wives and to "retain wives with kindness or release them with consideration." The effect of the Syrian law, then, is to subject the husband's motive for repudiation to the scrutiny of the court and to penalize him, albeit to a limited extent, for abuse of his power. Once again, however, the Tunisian *ijtihād* concerning repudiation is far more radical. Here the reformers argued that the Qurʾān orders the appointment of arbitrators in the event of discord between husband and wife. Clearly a pronouncement of repudiation by a husband indicated a state of discord between the spouses. Equally clearly the official courts were best suited to undertake the function of arbitration that then becomes necessary according to the Qurʾān. It is on this broad ground that the Tunisian law abolishes the right of a husband to repudiate his wife extrajudicially and enacts that: "Divorce outside a court of law is without legal effect." Although the court must dissolve the marriage if the husband persists in his repudiation, it has an unlimited power to grant the wife compensation for any damage she has sustained from the divorce—although in practice this power has so far been used most sparingly. In regard to polygamy and *ṭalāq* therefore, Tunisia has achieved by reinterpretation of the Qurʾān reforms hardly less radical than those effected in Turkey some 30 years previously by the adoption of the Swiss Civil Code.

Pakistani reforms

In Pakistan a new interpretation of the Qurʾān and *sunnah* was the declared basis of the reforms introduced by the Muslim Family Laws Ordinance of 1961, although the provisions of the Ordinance in relation to polygamy and *ṭalāq* are much less radical than the corresponding Middle Eastern reforms, since a second marriage is simply made dependent upon the consent of an Arbitration Council and the effect of a husband's repudiation is merely suspended for a period of three months to afford opportunity for reconciliation.

Judicial decisions in Pakistan have also unequivocally endorsed the right of independent interpretation of the Qurʾān. For example, in *Khurshīd Bībī* v. *Muḥammad Amīn* (1967) the Supreme Court held that a Muslim wife could as a right obtain a divorce simply by payment of suitable compensation to her husband. This decision was based on the Court's interpretation of a relevant Qurʾānic verse. But under traditional Sharīʿah law this form of divorce, known as *khul*, whereby a wife pays for her release, is a contract between the spouses and as such entirely dependent upon the husband's free consent.

These are but a few examples of the many far-reaching changes that have been effected in the Islāmic family law. But the whole process of legal reform as it has so far developed still involves great problems of principle and practice. A hard core of traditionalist opinion still adamantly rejects the validity of the process of reinterpretation of the basic texts of divine revelation. The traditionalists argue that the texts are merely being manipulated to yield the meaning that suits the preconceived purposes of the reformers, and that therefore, contrary to fundamental Islāmic ideology, it is social desirability and not the will of Allāh that is the ultimate determinant of the law.

As regards the practical effect of legal reform, there exists in many Muslim countries a deep social gulf between a Westernized and modernist minority and the conservative mass of the population. Reforms that aim at satisfying the standards of progressive urban society have little significance for the traditionalist communities of rural areas or for the Mulsim fundamentalists, whose geographical and social distribution crosses all apparent boundaries. It is also often the case that the *qāḍīs*, through their background and training, are not wholly sympathetic with the purposes of the modernist legislators—an attitude often reflected in their interpretations of the new codes.

Such problems are, of course, inevitable in the transitional stage of social evolution in which Islām finds itself. But the one supreme achievement of jurisprudence over the past few decades has been the emergence of a functional approach to the question of the role of law in society. Jurisprudence has discarded the introspective and idealistic attitude that the doctrine of *taqlīd* had imposed upon it since early medieval times and now sees its task to be the solution of the problems of contemporary society. It has emerged from a protracted period of stagnation to adopt again the attitude of the earliest Muslim jurists, whose aim was to relate the dictates of the divine will to their own social environment. It is this attitude alone that has ensured the survival of the Sharīʿah in modern times as a practical system of law and that alone provides its inspiration for the future. (N.J.C.)

Social and ethical principles

FAMILY LIFE

A basic social teaching of Islām is the encouragement of marriage, and the Qurʾān regards celibacy definitely as something exceptional—to be resorted to only under economic stringency. Thus, monasticism as a way of life was severely criticized by the Qurʾān. With the appearance of Ṣūfism, however, many Ṣūfīs preferred celibacy, and some even regarded women as an evil distraction from piety, although marriage remained the normal practice also with Ṣūfīs.

Pre-Islāmic practice of polygamy and the Qurʾān

Polygamy, which was practiced in pre-Islāmic Arabia, was permitted by the Qurʾān, which, however, limited the number of simultaneous wives to four, and this permission was made dependent upon the condition that justice be done among co-wives. The Qurʾān even suggests that "You shall never be able to do justice among women, no matter how much you desire." Medieval law and society, however, regarded this "justice" to be primarily a private matter between a husband and his wives, although the law did provide redress in cases of gross neglect of a wife. Right of divorce was also vested basically in the husband, who could unilaterally repudiate his wife, although the woman could also sue her husband for divorce before a court on certain grounds.

The virtue of chastity is regarded as of prime importance by Islām. The Qurʾān advanced its universal recommendation of marriage as a means to ensure a state of chastity (*iḥṣān*), which is held to be induced by a single free wife. The Qurʾān states that those guilty of adultery are to be severely punished with 100 lashes. Tradition has intensified this injunction and has prescribed this punishment for unmarried persons, but married adulterers are to be stoned to death. A false accusation of adultery is punishable by 80 lashes.

The general ethic of the Qurʾān considers the marital bond to rest on "mutual love and mercy," and the spouses are said to be "each other's garments." The detailed laws of inheritance prescribed by the Qurʾān also tend to confirm the idea of a central family—husband, wife, and children, along with the husband's parents. Easy access to polygamy (although the normal practice in Islāmic society has always been that of monogamy) and easy divorce on the part of the husband led, however, to frequent abuses in the family. In recent times, most Muslim countries have enacted legislation to tighten up marital relationships.

Rights of parents in terms of good treatment are stressed in Islām, and the Qurʾān extols filial piety, particularly tenderness to the mother, as an important virtue. A murderer of his father is automatically disinherited. The tendency of the Islāmic ethic to strengthen the immediate family on the one hand and the community on the other at the expense of the extended family or tribe did not succeed, however. Muslim society, until the encroachments upon it of modernizing influences, has remained basically one composed of tribes or quasi-tribes. Despite urbanization, tribal affiliations offer the greatest resistance to change and development of a modern polity. So strong, indeed, has been the tribal ethos that, in most Muslim societies, daughters are not given their inheritance share prescribed by the sacred law in order to prevent disintegration of the joint family's patrimony.

THE STATE

Because Islām draws no distinction between the religious and the temporal spheres of life, the Muslim state is by definition religious. The main differences between the Sunnī, Khawārij, and Shīʿī concepts of rulership have already been pointed out above. It should be noted that, although the office of the Sunnī caliph (*khalīfah,* one who is successor to the Prophet in rulership) is religious, this does not imply any functions comparable to those of the pope. The caliph has no authority either to define dogma or, indeed, even to legislate. He is the chief executive of a religious community, and his primary function is to implement the sacred law and work in the general interests of the community. He himself is not above the law and if necessary can even be deposed, at least in theory.

Sunnī political theory is essentially a product of circumstance—an after-the-fact rationalization of historical developments. Thus, between the Shīʿah legitimism that restricts rule to ʿAlī's family and the Khawārij democratism that allowed rulership to anyone, even to "an Ethiopian slave," Sunnism held the position that "rule belonged to the Quraysh" (the Prophet's tribe)—the condition that actually existed. Again, in view of the extremes represented by the Khawārij, who demanded rebellion against what they considered to be unjust or impious rule, and Shīʿites, who raised the *imām* to a metaphysical plane of infallibility, Sunnites took the position that a ruler has to satisfy certain qualifications but that rule cannot be upset on small issues. Indeed, under the impact of civil wars started by the Khawārij, Sunnism drifted to more and more conformism and actual toleration of injustice.

Political conformism of Sunnism

The first step taken in this direction by the Sunnites was the enunciation that "one day of lawlessness is worse than 30 years of tyranny." This was followed by the principle that "Muslims must obey even a tyrannical ruler." Soon, however, the sultan (ruler) was declared to be "shadow of God on earth." No doubt, the principle was also adopted—and insisted upon—that "there can be no obedience to the ruler in disobedience of God"; but there is no denying the fact that the Sunnī doctrine came more and more to be heavily weighted on the side of political conformism. This change is also reflected in the principles of legitimacy. Whereas early Islām had confirmed the pre-Islāmic democratic Arab principle of rule by consultation (*shūrā*) and some form of democratic election of the leader, those practices soon gave way to dynastic rule with the advent of the Umayyads. The *shūrā* was not developed into any institutionalized form and was, indeed, soon discarded. Soon the principle of "might is right" came into being, and later theorists frankly acknowledged that actual possession of effective power is one method of the legitimization of power.

In spite of this development, the ruler could not become absolute because a basic restraint was placed upon him by the Sharīʿah law under which he held his authority and which he dutifully was bound to execute and defend. When, in the latter half of the 16th century, the Mughal emperor Akbar in India wanted to arrogate to himself the right of administrative–legal absolutism, the strong reaction of the orthodox thwarted his attempt. In general, the *'ulamā'* (religious scholars) jealously upheld the sovereign position of the Sharī'ah against the political authority.

The effective shift of power from the caliph to the sultan was, again, reflected in the redefinition of the functions of the caliph. It was conceded that, if the caliph administered through *wazīrs* (viziers or ministers) or subordinate rulers (*amīrs*), it was not necessary for him to embody all the physical, moral, and intellectual virtues theoretically insisted upon earlier. In practice, however, the caliph was no more than a titular head from the middle of the 10th century onward, when real power passed to self-made and adventurous *amīrs* and sultans, who merely used the caliph's name for legitimacy.

EDUCATION

Muslim educational activity began in the 8th century, primarily in order to disseminate the teaching of the Qur'ān and the *sunnah* of the Prophet. The first task was to collect and systematize the knowledge that was handed down by the previous generations concerning the meaning of the Qur'ān and the activity and precepts of the Prophet. Thus, in the early period, the character of learning was traditional. that tradition was committed to writing in the 2nd century AH and systematized and developed in the 3rd century AH. This vast activity of "seeking knowledge" (*ṭalab al-'ilm*) resulted in the creation of specifically Arab sciences of tradition, history, and literature.

When the introduction of the Greek sciences—philosophy, medicine, and mathematics—created a formidable body of lay knowledge, a creative reaction on the traditional religious base resulted in the rationalist theological movement of the Mu'tazilah. Based on that Greek legacy, from the 9th to the 12th century AD a brilliant philosophical movement flowered and presented a challenge to orthodoxy on the issues of the eternity of the world, the doctrine of revelation, and the status of the Sharī'ah.

Distinction between religious and secular sciences

The orthodox met the challenges positively by formulating the religious dogma. At the same time, however, for fear of heresies, they began to draw a sharp distinction between religious and secular sciences. The custodians of the Sharī'ah developed an unsympathetic attitude toward the secular disciplines and excluded them from the curriculum of the *madrasah* (college) system. Their exclusion from the Sunnī system of education proved fatal, not only for those disciplines but, in the long run, for religious thought in general because of the lack of intellectual challenge and stimulation. A typical *madrasah* curriculum included logic (which was considered necessary as an "instrumental" science for the formal correctness of thinking procedure), Arabic literature, law, Ḥadīth, Qur'ān commentary, and theology. Despite sporadic criticism from certain quarters, the *madrasah* system remained impervious to change.

One important feature of Muslim education was that primary education (which consisted of Qur'ān reading, writing, and rudimentary arithmetic) did not feed candidates to institutions of higher education, and the two remained separate. In higher education, emphasis was on books rather than on subjects and on commentaries rather than on original works. This, coupled with the habit of learning by rote (which was developed from the basically traditional character of knowledge that encouraged learning more than thinking), impoverished intellectual creativity still further.

Despite these grave shortcomings, however, the *madrasah* produced one important advantage. Through the uniformity of its religio-legal content, it gave the *'ulamā'* the opportunity to effect that overall cohesiveness and unity of thought and purpose that, despite great variations in local Muslim cultures, has become a palpable feature of the world Muslim community. This uniformity has withstood even the serious tension created against the seats of formal learning by Ṣūfism through its peculiar discipline and its own centres.

In contrast to the Sunnī attitude toward it, philosophy continued to be seriously cultivated among the Shī'ah, even though it developed a strong religious character. Indeed, philosophy has enjoyed an unbroken tradition in Persia down to the present and has produced some highly original thinkers. Both the Sunnī and the Shī'ah medieval systems of learning, however, have come face to face with the greatest challenge of all—the impact of modern education and thought.

Organization of education developed naturally in the course of time. Evidence exists of small schools already established in the first century of Islām that were devoted to reading, writing, and instruction in the Qur'ān. These schools of "primary" education were called *kuttābs.* The well-known governor of Iraq at the beginning of the 8th century, the ruthless al-Ḥajjāj, had been a schoolteacher in his early career. When higher learning in the form of tradition grew in the 8th and 9th centuries, it was centred around learned men to whom students travelled from far and near and from whom they obtained a certificate (*ijāzah*) to teach what they had learned. Through the munificence of rulers and princes, large private and public libraries were built, and schools and colleges arose. In the early 9th century, a significant incentive to learning came from the translations made of scientific and philosophical works from the Greek (and partly Sanskrit) at the famous *bayt al-ḥikmah* ("house of wisdom") at Baghdad, which was officially sponsored by the caliph al-Ma'mūn. The Fāṭimid caliph al-Ḥākim set up a *dār alḥikmah* ("hall of wisdom") in Cairo in the 10–11th centuries. With the advent of the Seljuq Turks, the famous vizier Niẓām al-Mulk created an important college at Baghdad, devoted to Sunnī learning, in the latter half of the 11th century. The oldest surviving university, al-Azhar at Cairo, was established by the Fāṭimids, but Saladin (Ṣalāḥ ad-Dīn al-Ayyūbī), after ousting the Fāṭimids, consecrated it to Sunnī learning in the 12th century. Throughout subsequent centuries, colleges and quasi-universities (called *madrasah* or *dār al-'ulūm*) arose throughout the Muslim world from Spain (whence philosophy and science were transmitted to the Latin West) across Central Asia to India. In Turkey, a new style of *madrasah* came into existence; it had four wings, for the teaching of the four schools of Sunnī law. Professorial chairs were endowed in large colleges by princes and governments, and residential students were supported by college endowment funds. Myriads of smaller centres of learning were endowed by private donations.

CULTURAL DIVERSITY

Underneath the legal and creedal unity, the world of Islām harbours a tremendous diversity of cultures, particularly in the outlying regions. The expansion of Islām can be divided into two broad periods. In the first period of the Arab conquests, the assimilative activity of the conquering religion was far-reaching. Although Persia resurrected its own language and a measure of its national culture after the first three centuries of Islām, its culture and language had come under heavy Arab influence. Only after Ṣafavid rule installed Shī'ism as a distinctive creed in the 16th century did Persia regain a kind of religious autonomy. The language of religion and thought, however, continued to be Arabic.

Ṣūfī compromises with local customs

In the second period, the spread of Islām was not conducted by the state with *'ulamā'* influence but was largely the work of Ṣūfī missionaries. The Ṣūfīs, because of their latitudinarianism, compromised with local customs and beliefs and left a great deal of the pre-Islāmic legacy in every region intact. Thus, among the Central Asian Turks, shamanistic practices were absorbed, while in Africa the holy man and his *barakah* (an influence supposedly causing material and spiritual well-being) are survivors from the older cults. In India there are large areas geographically distant from the Muslim religio-political centre of power in which customs are still Hindu and even pre-Hindu and in which people worship a motley of saints and deities in common with the Hindus. The custom of *satī,* under which a widow burned herself alive along with her dead husband, persisted in India even among some Muslims until late into the Mughal period. The 18th- and 19th-century reform movements exerted themselves to "purify" Islām of these accretions and superstitions.

Indonesia affords a striking example of this phenomenon. Because Islām reached there late and soon thereafter came under European colonialism, the Indonesian society has retained its pre-Islāmic world view beneath an overlay of Islāmic practices. It keeps its customary law (called *adat*) at the expense of the Sharīʿah; many of its tribes are still matriarchal; and culturally the Hindu epics *Rāmāyaṇa* and *Mahābhārata* hold a high position in national life. Since the 19th century, however, orthodox Islām has gained steadily in strength because of fresh contacts with the Middle East.

Apart from regional diversity, the main internal division within Islāmic society is brought about by urban and village life. Islām originally grew up in the two cities of Mecca and Medina, and as it expanded, its peculiar ethos appears to have developed in urban areas. Culturally, it came under a heavy Persian influence in Iraq, where the Arabs learned the ways and style of life of their conquered people, who were culturally superior to them. The custom of veiling women (which originally arose as a sign of aristocracy but later served the purpose of segregating women from men—the *pardah*), for example, was acquired in Iraq.

Another social trait derived from outside cultures was the disdain for agriculture and manual labour in general. Because the people of the town of Medina were mainly agriculturists, this disdain could not have been initially present. In general, Islām came to appropriate a strong feudal ethic from the peoples it conquered. Also, because the Muslims generally represented the administrative and military aristocracy and because the learned class (the *ʿulamāʾ*) was an essential arm of the state, the higher culture of Islām became urban based.

This city orientation explains and also underlines the traditional cleavage between the orthodox Islām of the *ʿulamāʾ* and the folk Islām espoused by the Ṣūfī orders of the countryside. In the modern period, the advent of education and rapid industrialization threatened to make this cleavage still wider. With the rise of a strong and widespread fundamentalist movement in the second half of the 20th century, this dichotomy has decreased.

Religion and the arts

THE VISUAL ARTS

Aniconism as a governing principle of art

The Arabs before Islām had hardly any art except poetry, which had been developed to full maturity and in which they took great pride. As with other forms of culture, the Muslim Arabs borrowed their art from Persia and Byzantium. Whatever elements the Arabs borrowed, however, they Islāmized in a manner that fused them into a homogeneous spiritual-aesthetic complex. The most important principle governing art was aniconism; *i.e.,* the religious prohibition of figurization and representation of living creatures. Underlying this prohibition is the assumption that God is the sole author of life and that a person who produces a likeness of a living being seeks to rival God. The tradition ascribed to the Prophet that a person who makes a picture of a living thing will be asked on the Day of Judgment to infuse life into it, whether historically genuine or not, doubtless represents the original attitude of Islām. In the Qurʾān (3:49, 5:113), reflecting an account in a New Testament apocryphal work, it is counted among the miracles of Jesus that he made likenesses of birds from clay "by God's order," and, when he breathed into them, they became real birds, again, "by God's order."

Hence, in Islāmic aniconism two considerations are fused together: (1) rejection of such images that might become idols (these may be images of anything) and (2) rejection of figures of living things. Plato and Plotinus, Greek philosophers, had also dismissed representative art as an "imitation of nature"; *i.e.,* as something removed from reality. The Islāmic attitude is more or less the same, with the added element of attributing to the artist a violation of the sanctity of the principle of life. The same explanation holds for the Qurʾānic criticism of a certain kind of poetry, namely, free indulgence in extravagant image mongering: "They [poets] recklessly wander in every valley" (26:225).

This basic principle has, however, undergone modifications. First, pictures were tolerated if they were confined to private apartments and harems of palaces. This was the case with some members of the Umayyad and ʿAbbāsid dynasties, Turks, and Persians—in particular with the Shīʿah, who have produced an abundance of pictorial representations of the holy family and of the Prophet himself. Second, in the field of pictorial representation, animal and human figures are combined with other ornamental designs such as fillets and arabesques—stressing their ornamental nature rather than representative function. Third, for the same reason, in plastic art they appear in low relief. In other regions of the Muslim world—in North Africa, Egypt, and India (except for Mughal palaces)—representational art was strictly forbidden. Even in paintings, the figures have little representational value and are mostly decorative and sometimes symbolic. This explains why plastic art is one of the most limited areas of Islāmic art. The only fullfledged plastic figures are those of animals and a few human figures that the Seljuqs brought from eastern Turkistan.

Miniature painting

Much more important than plastic art were paintings, particularly frescoes and later Persian and Perso-Indian miniatures. Frescoes are found in the Umayyad and ʿAbbāsid palaces and in Spain, Iran, and in the harem quarters of the Mughal palaces in India. Miniature paintings, introduced in Persia, assumed much greater importance in the later period in Mughal India and Turkey. Miniature painting was closely associated with the art of book illumination, and this technique of decorating the pages of the books was patronized by princes and other patrons from the upper classes. (Miniature painting is also discussed below; see *Illustration of myth and legend.*)

MUSIC

Instrumental music was forbidden by the orthodox in the formative stages of Islām. As for vocal music, its place was largely taken by a sophisticated and artistic form of the recitation of the Qurʾān known as *tajwīd.* Nevertheless, the Muslim princely courts generously patronized and cultivated music. Arab music was influenced by Persian and Greek music. Al-Fārābī, a 10th-century philosopher, is credited with having constructed a musical instrument called the *arghanūn* (organ). In India, Amīr Khosrow, a 14th-century poet and mystic, produced a synthesis of Indian and Persian music and influenced the development of later Indian music.

Among the religious circles, the Ṣūfīs introduced both vocal and instrumental music as part of their spiritual practices. The *samāʿ,* as this music was called, was opposed by the orthodox at the beginning, but the Ṣūfīs persisted in this practice, which slowly won general recognition. The great Ṣūfī poet Jalāl ad-Dīn ar-Rūmī (died 1273)—revered equally by the orthodox and the Ṣūfīs—heard the divine voice in his stringed musical instrument when he said "Its head, its veins (strings) and its skin are all dry and dead; whence comes to me the voice of the Friend?"

LITERATURE

In literature, drama and pure fiction were not allowed—drama because it was a representational art and fiction because it was akin to lying. Similar constraints operated against the elaboration of mythology (see below *Islāmic myth and legend*). Story literature was tolerated, and the great story works of Indian origin—*The Thousand and One Nights* and *Kalīlah wa Dimnah*—were translated from the Persian, introducing secular prose into Arabic. Didactic and pious stories were used and even invented by popular preachers. Much of this folklore found its way back into enlarged editions of *The Thousand and One Nights* and, through it, has even influenced later history writing. Because of the ban on fictional literature, there grew a strong tendency in later literary compositions—in both poetry and prose—toward hyperbole (*mubālaghah*), a literary device to satisfy the need of getting away from what is starkly real without committing literal falsehood, thus often resulting in the caricature and the grotesque. Poetry lent itself particularly well to this device, which was freely used in panegyrics, satires, and lyrics. As a form of effective expression, poetry is eminently characteristic of the East. The Arab genius is almost natively poetical with its strong and vivid imagination not easily amenable to the rigorous order that reason imposes upon the mind.

Poetry as a characteristic of the East

This borderline attitude between the real and the unreal was particularly favourable to the development, in all medieval Islāmic literatures of the Middle East, of the lyric and panegyric forms of poetry wherein every line is a self-contained unit. Much more importantly, it afforded a specially suitable vehicle for a type of mystic poetry in which it is sometimes impossible to determine whether the poet is talking of earthly love or spiritual love. For the same reason, poetry proved an effective haven for thinly veiled deviations from and even attacks on the literalist religion of the orthodox.

ARCHITECTURE

Architecture is by far the most important expression of Islāmic art, particularly the architecture of mosques. It illustrates both the diversity of cultures that participated in the Islāmic civilization and the unifying force of Islāmic monotheism represented by the spacious expanse of the mosque—a veritable externalization of the all-enveloping divine unity, heightened by the sense of infinity of the arabesque design. The arabesque, though ornately decorative, spiritually represents the infinite vastness of God.

Among the earliest monuments are the mosque of ʿAmr built in Egypt in 641–642 and the famous Dome of the Rock of Jerusalem (finished in 691), which, however, is not a mosque but a monument, a concentric-circular structure consisting of a wooden dome set on a high drum and resting on four tiers and 12 columns. The Umayyad ruler al-Walīd (died 715) built the great mosque at Damascus and al-Aqṣā Mosque at Jerusalem with two tiers of arcades in order to heighten the ceiling. The early Syro-Egyptian mosque is a heavily columned structure with a prayer niche (*miḥrāb*) oriented toward the Kaʿbah sanctuary at Mecca.

In Spanish and North African architecture these features are combined with Roman-Byzantine characteristics, the masterpieces of Spanish architecture being the famous Alhambra Palace at Granada and the Great Mosque of Córdoba. In the famous Persian mosques, the characteristic Persian elements are the tapered brick pillars, the arches (each supported by several pillars), the huge arcades, and the four sides called *eyvān*s. With the advent of the Seljuqs in the 11th century, faience decoration (glazed earthenware) of an exquisite beauty was introduced, and it gained further prominence under the Timurids (14th–16th centuries).

In the number and greatness of mosques, Turkey has the pride of first place in the Muslim world. Turkey began with a Persian influence and then later Syrian in the 13th and 14th centuries, but Turkey developed its own cupola domes and monumental entrances. The Turkish architects accomplished symmetry by means of one large dome, four semidomes, and four small domes among them. In the Indo-Pakistan subcontinent, Muslim architecture first employed Hindu architectural features (*e.g.,* horizontal rather than arcuate, or bowlike, arches and Hindu ornamentation), but later the Persian style predominated.

(F.R./Ed.)

Islāmic myth and legend

The strict monotheism of Islām does not allow for much mythological embellishment, and only reluctantly were the scriptural revelations of the Qurʾan elaborated and enlarged by commentators and popular preachers. Thus, in the first three centuries, a number of ideas from the ancient Near East, from Hellenistic and especially from Judeo-Christian traditions were absorbed into Islām and given at least partial sanction by the theologians. At the same time, legends were woven around the Prophet Muḥammad and the members of his family. Though inconsistent with historical reality, these legends formed for the masses the main sources of inspiration about the famous figures of the past.

Role of storytellers

Since early times Islāmic theologians have sought to disregard the Qurʾānic interpretation of both storytellers and mystics. The *quṣṣāṣ,* or storytellers, made the Qurʾānic revelation more understandable to the masses by filling in the short texts with detailed descriptions that were not found in scripture. Though the mystics tried to maintain the purity of the divine word, they also attempted a spiritualization of both the Qurʾān and the popular legends that developed around it. Their way of giving to the Qurʾānic words a deeper meaning, however, and discovering layer after layer of meaning in them, sometimes led to new quasi-mythological forms. Later Islāmic mystical thinkers built up closed systems that can be called almost mythological (*e.g.,* the angelology—theory of angels—of Suhrawardī al-Maqtūl, executed 1191). An interesting development is visible in poetry, especially in the Persian-speaking areas, where mythological figures and pious legend often were turned into secular images that might awaken in the reader a reminiscence of their religious origin. Such images contribute to the iridescent and ambiguous character of Persian poetry.

SOURCES AND VARIATIONS

The Qurʾān and non-Islāmic influences. The sources of Islāmic mythology are first of all the Qurʾānic revelations. Since, for the Muslims, the Qurʾān is the uncreated word of God (the text revealed to Muḥammad considered an earthly manifestation of the eternal and uncreated original in heaven), it contains every truth, and whatever is said in it has been the object of meditation and explanation through the centuries. Thus, since the 9th century, commentators on the Qurʾān have been by far the most important witnesses for Islāmic "mythology." They wove into their explanations various strands of Persian and ancient oriental lore and relied heavily on Jewish tradition. For example, the Jewish convert, Kaʿb al-Aḥbār brought much of the *Isrāʾīliyāt* (things Jewish) into Islāmic tradition. Later on, the mystics' commentaries expressed some gnostic (a dualistic viewpoint in which spirit is viewed as good and matter as evil) and Hellenistic concepts, of which the Hellenistic idea of the Perfect Man—personified in Muḥammad—was to gain greatest prominence. Commentaries written in the border areas of Islāmic countries now and then accepted a few popular traditions from their respective areas; however, the formative period was finished quite early. Traditions about the life and sayings of the Prophet grew larger and larger and are interesting for the study of the adoption of foreign mythological material. A valuable source for Islāmic legends are the *qiṣaṣ al-anbiyāʾ*—stories of the prophets, such as those by Thaʿālibī (born 1035) and Kisāʾī (11th century)—traditions concerning the prophets of yore in which a large number of pre-Islāmic and non-Islāmic ideas were incorporated.

While the classical mythology of Islām, as far as it can be properly called so, is spread over the whole area of Islām, the miracles and legends around a particular Muslim saint are found chiefly in the area of his special influence (especially where his order is most popular). Even if the names of the saints differ, the legends woven around them are very similar to each other and almost interchangeable. In the area where Persian was read—from Ottoman Turkey to India—the mythological concepts of Ferdowsī's *Shāh-nāmeh* are found side by side with the legends taken from ʿAṭṭār's and Rūmī's works.

Trans-cultural proliferation

The mystics. From the 11th century onward, the biographies of the mystics often show interesting migrations of legendary motifs from one culture to another. For the Persian-speaking countries the *Tazkerat ol-Owlīyāʾ* ("Memoirs of the Saints") of Farīd od-Dīn ʿAṭṭār (died *c.* 1220) has become the storehouse of legendary material about the early Ṣūfī mystics. ʿAṭṭār's Persian epics (especially his *Manṭeq ot-ṭeyr,* the "Birds' Conversation") also contain much material that was used by almost every writer after him. *The Masnavī* (a sort of poetic encyclopaedia of mystical thought in 26,000 couplets) of Jalāl od-Dīn Rūmī (died 1273) is another important source for legends of saints and prophets. For the Iranian world view, Ferdowsī's (died *c.* 1020) *Shāh-nāmeh* ("Book of Kings") gave a poetical account of the mythology of old Iran, and its heroes became models for many poets and writers. The whole mythological and legendary heritage is condensed in allusions found in lyrical and panegyrical poetry. The Persian poet Ebrāhim ebn ʿAlī Khāqānī's (*c.* 1121–*c.* 1199) works, *qaṣīdahs* ("Odes"), are typical. The close connection of the Ṣūfī orders with the artisans' lodges and guilds was instrumental in the dissemination of legendary material, especially about the alleged founder, or patron, of the guild (such as Ḥallāj as patron of cottoncarders and Idrīs as patron of the tailors).

Muslim historians interested in world history often began their works with mythological tales; central Asian traditions were added in Iran during the Il-Khanid period (AD 1256–1335). Folk poetry, in the different languages spoken by Muslims, provides a popular representation of traditional material, be it in Arabic, Persian, Turkish, the Indian and Pakistani languages (Urdu, Bengali, Sindhi, Panjabi, Baluchi, etc.), or the African languages; in all of them allusions to myth and legend are found down to the level of riddles and lullabies. Typical of the legendary tradition of the Shī'ah are the *ta'ziyas* ("passion-plays") in Iran, commemorating the death of Husayn ibn 'Alī in Karbalā' (680) and the *marsīyehs* (threnodies or elegies for the dead), which form an important branch of the Urdu poetry of India and Pakistan. A proper study of the distribution of most aspects of mythology in the various Muslim areas has not been undertaken, since much of the popular material is rarely available in print or is written in less-known languages—a good example is the extremely rich collections of legends and popular pious works in the Pakistani language, Sindhi.

TYPES OF MYTH AND LEGEND

Cosmogony and eschatology. The world was created by God's word *kun* ("Be") out of nothing; after the creation of the angelic beings from light, Adam was formed from clay and destined to be God's vicegerent, *khalīfah.* All of the angels obeyed God's order to prostrate themselves before Adam, except Iblīs (Satan), who refused and was cursed; due to Iblīs' instigation Adam ate the forbidden fruit (or grain) and was driven out of paradise. Questions of original sin or of Eve's role do not arise in the Muslim version of creation. Satan's disobedience has been explained by the mystics as actually an expression of his obedience to the divine will that does not allow worship of any but the Lord and that conflicted with the order that Satan prostrate himself before Adam.

By courtesy of the Bibliothèque Nationale, Paris

Satan's refusal to worship Adam, depicting the rebellious angel Iblīs, or Satan, as the human figure on a prayer rug (right). From a 17th-century manuscript of Majāles ol-'Oshshāq. In the Bibliothèque Nationale, Paris (supplément Persan 1559).

Before the creation, God addressed the posterity of Adam: "Am I not your Lord," *alastu birabbikum,* and they answered "Yes" (Qur'ān, *sūrah* 7:172). This pre-eternal covenant is the favourite topic of mystical poetry, especially in the Persian-speaking areas for expressing pre-eternal love between God and man, or the unchangeable fate that was accepted that very day, the Yesterday as contrasted to the Tomorrow of resurrection. Angels and jinns (genies) are living powers that become visible in human life; they are accepted as fully real.

Every destiny is written on the "well-preserved tablet," and now "the pen has dried up"—a change in destiny is not possible. Later mystics have relied on an extra-Qur'ānic revelation in which God attests: "I was a hidden treasure" and have seen the reason for creation in God's yearning to be known and loved. For them, creation is the projection of divine names and qualities onto the world of matter.

Centrality of death and resurrection

The central event of Islām is death and resurrection. The dead will be questioned by two terrible angels (that is why the profession of faith is recited to the dying); only the souls of martyrs go straight to heaven where they remain in the crops of green birds around the divine throne (green is always connected with heavenly bliss). The end of the world will be announced by the coming of the mahdī (literally, "the directed or guided one")—a messianic figure who will appear in the last days and is not found in the Qur'ān but developed out of Shī'ah speculations and sometimes identified with Jesus. The mahdī will slay the Dajjāl, the one-eyed evil spirit, and combat the dangerous enemies, Yājūj and Mājūj, who will come from the north of the earth. The trumpet of Isrāfīl, one of the four archangels, will awaken the dead for the day of resurrection, which is many thousands of years long and the name of which has come to designate a state of complete confusion and turmoil.

The eschatological inventory as described in the Qur'ān was elaborated by the commentators: the scales on which the books or deeds are weighed (an old Egyptian idea), the book in which the two recording angels have noted down man's deeds, and the narrow bridge that is said to be sharper than a sword and thinner than a hair and leads over hell (an Iranian idea). The dreadful angels of hell and the horrors of that place are as thoroughly described by theologians as the pleasures of paradise, with its waters and gardens and the houris who are permanent virgins. Pious tradition promises space in heavenly mansions, filled with everything beautiful, to those who repeat certain prayer formulas a certain number of times, or for similar rewarding deeds, whereas the mystic longs not "for houris some thousand years old" but for the vision of God, who will be visible like the full moon. In the concept of the *sidrah* tree as the noblest place in paradise a remnant may be found of the old tree of life. God's throne is on the waters (Qur'ān, *sūrah* 11:9) in the highest world, surrounded by worshipping angels. The created world, the earth, is surrounded by the mountain Qāf and enclosed by two oceans that are separated by a barrier. Mecca is the navel of the earth, created 2,000 years before everything else, and the deluge did not reach to proto-*Ka'bah.* Often the world is conceived as a succession of seven heavens and seven earths, and a popular tradition says that the earth is on water, on a rock, on the back of a bull, on a *kamkam* (meaning unknown), on a fish, on water, on wind, on the veil of darkness—hence the Persian expression *az māh tā māhī,* from the moon to the fish; *i.e.,* throughout the whole world.

Tales and legends concerning religious figures. The majority of popular legends concern the leading personalities of Islām.

Muhammad. Muḥammad, whose only miracle, according to his own words, was the bringing of the Qur'ān, is credited with innumerable miracles and associated with a variety of miraculous occurrences: his finger split the moon, the cooked poisoned meat warned him not to touch it, the palm trunk sighed, the gazelle spoke for him; he cast no shadow; from his perspiration the rose was created, etc. His ascension to heaven (*mi'rāj*) is still celebrated: he rode the winged horse Burāq in the company of the angel Gabriel through the seven spheres, meeting the other prophets there, until he reached the divine presence, alone, even without the angel of inspiration. Muḥammad-mysticism proper was developed in the late 9th century; he is shown as the one who precedes creation, his light is pre-eternal, and he is the reason for and goal of creation. He becomes the perfect man, uniting the divine and the human sphere as dawn is between night and day. His birth was surrounded by miracles, and his birthday (12. Rabīʿ I) became a popular holiday on which numerous poems were written to praise his achievements. The hope for him who has been sent as "mercy for the worlds" and will intercede for his community on Doomsday is extremely strong, especially among the masses, where these legends have completely overshadowed his historical figure.

Muḥammad-mysticism

Other Qurʾānic figures. In addition to Muḥammad himself, his cousin and son-in-law ʿAlī, the Shīʿah hero, has been surrounded by legends concerning his bravery, his miraculous sword, Dhūa'l-fiqār, and his wisdom. ʿAlī's son, Ḥusayn, is the subject of innumerable poems that concern the day of his final fight in Karbalāʾ.

Almost every figure mentioned in the Qurʾān has become the centre of a circle of legends, be it Yūsuf, the symbol of overwhelming beauty, or Jesus with the lifegiving breath, the model of poverty and asceticism. Of special interest is Khiḍr, identified with the unnamed companion of Moses (Qurʾān, *sūrah* 20). He is the patron saint of the wayfarers, connected with green, the colour of heavenly bliss, appearing whenever a pious person is in need, and immortal since he drank from the fountain of life, which is hidden in the darkness. In many respects, he is the Islāmic counterpart of Elijah. Strong influences of the Alexander romances (a widely distributed literary genre dealing with the adventures of Alexander the Great) are visible in his figure.

The mystics as miracle workers

Mystics and other later figures. The great religious personalities have become legendary, especially the martyr-mystic Ḥallāj (executed in Bagdad, 922). His word *anā al-Ḥaqq,* "I am the Creative Truth," became the motto of many later mystics. His death on the gallows is the model for the suffering of lovers, and allusions to his fate are frequent in Islāmic literature. An earlier mystic, Abū Yazīd al-Bisṭāmī (died 874), was the first to speak about the ascension of the mystic to heaven, which is a metaphor for higher unitive, mystical experience. A variation of the Buddha legend has been transferred onto the person of the first Ṣūfī (mystic) who practiced absolute poverty and trust in God, the Central Asian Ibrāhīm ibn Adham (died *c.* 780). The founders of mystical orders were credited by their followers with a variety of miracles, such as riding on lions, healing the sick, walking on water, being present at two places at the same time, and cardiognosia (which is the knowledge of what is in another's heart, or thought reading). ʿAbd al-Qādir al-Jīlānī (died 1166), the founder of the widespread Qādirīyah order of mystics, and many others have attracted upon themselves a large number of popular stories that formerly had been told about pre-Islāmic saints or about some divinities, and these motifs can easily be transferred from one person to the other. In this sphere the survival of pre-Islāmic customs and legends is most visible. The idea of the hierarchy of saints, culminating in the *quṭb*, the pole or axis, thanks to whose activities the world keeps going, belongs to the mythology of Ṣūfism (Islāmic mysticism).

Mythologization of secular tales. A feature of Islāmic mythology is the transformation of unreligious stories into vehicles of religious experience. The old hero of romantic love in Arabic literature, Majnūn, "the demented one," became a symbol of the soul longing for identification with God, and in the Indus Valley the tales of Sassui or Sohnī, the girls who perish for their love, and other romantic figures, have been understood as symbols of the soul longing for union with God through suffering and death.

Tales and beliefs about numbers and letters. Many Muslim tales, legends, and traditional sayings are built upon the mystical value of numbers, such as the threefold or sevenfold repetition of a certain rite. This is largely explained by examples from the life of a saintly or pious person, often the Prophet himself, who used to repeat this or that formula so and so many times. The number 40, found in the Qurʾān (as also in the Bible) as the length of a period of repentance, suffering, preparation, and steadfastness, plays the same role in Islām where it is connected, for example, with the 40 days' preparation and meditation, or fasting, of the novice in the mystical brotherhood. To each number, as well as to each day of the week, special qualities are attributed through the authority of both actual and alleged statements of the Prophet. Many pre-Islāmic customs were thus justified. The importance given to the letters of the Arabic alphabet is peculiar to Muslim pious thought. Letters of the alphabet were assigned numerical values: the straight *alif* (numerical value one), the first letter of the alphabet, becomes a symbol of the uniqueness and unity of Allāh; the *b* (numerical value two), the first letter of the Qurʾān, represents to many mystics the creative power by which everything came into existence; the *h* (numerical value five) is the symbol of *huwa,* He, the formula for God's absolute transcendence; the *m* (numerical value 40) is the "shawl of humanity" by which God, the One (al-Aḥad), is separated from *Aḥmad* (*Muḥammad*). M is the letter of human nature and hints at the 40 degrees between man and God. The sect of the Ḥurūfīs developed these cabalistic interpretations of letters, but they are quite common in the whole Islāmic world and form almost a substitute for mythology.

Number symbolism

ILLUSTRATION OF MYTH AND LEGEND

Persian, Turkish, and Indian Mughal miniature painting

Since the art of representation is opposed in Islām, illustrations of mythological and legendary subjects are rarely found. Miniature painting developed only in the Persian and, later on, in the Turkish and Indo-Muslim areas. Books such as Zakarīyāʾ ebn Moḥammad al-Qazvīnī's *Cosmography* contain in some manuscripts a few pictures of angels, like Isrāfīl with the trumpet, and histories of the world or histories of the prophets, written in Iran or Turkey, also contain in rare manuscripts representations of angels or of scenes as told in the Qurʾān, especially the story of Yūsuf and Zalīkhā, which inspired many poems. The *Shāh-nāmeh* has been fairly frequently illustrated. When the Prophet of Islām is shown at all, his face is usually covered and in several cases his companions or his family members are also shown with veiled faces.

The only subject from the legends surrounding Muḥammad that has been treated by miniaturists several times is his ascension to heaven. There are a number of splendid Persian miniatures depicting this. In poetical manuscripts that contain allusions to legends of the saints, these topics were also sometimes illustrated (*e.g.,* Jonah and the great fish or scenes from the wanderings of Khiḍr). Several miniatures deal with the execution of the mystic al-Ḥallāj. Mythological themes proper are found almost exclusively in the paintings of Mughal India; especially in the period of Jahāngīr, in which the eschatological peace of lion and lamb lying together is illustrated as well as the myth of the earth resting on the bull, on the fish, etc. But by that time European influence was also already visible in Mughal art.

SIGNIFICANCE AND MODERN INTERPRETATIONS

Mythology proper has only a very small place in official Islām and is mostly an expression of popular traditions through which pre-Islāmic influences seeped into Islām. Reformers tried to purge Islām of all non-Qurʾānic ideas and picturesque elaborations of the texts, whereas the mystics tried to spiritualize them as far as possible. Modern Muslim exegesis attempts to interpret many of the mythological strands of the Qurʾān in the light of modern science, as psychological factors, like Muḥammad's ascension to heaven, and especially deprives the eschatological parts of the Qurʾān of their religious significance. Cosmic events are interpreted as predictions of modern scientific research. To some interpreters, jinns and angels are spiritual forces; to others, jinns are microbes or the like. Thus the religious text is confused with a textbook of science. Popular legends surrounding the Prophet and the saints are still found among the masses but are tending to disappear under the influence of historical research, though many of them have formed models for the behaviour and spiritual life of the Muslim believer. (An.Sc.)

BIBLIOGRAPHY

Islām. *General works: Cambridge History of Islam,* (1970, vol. 2, pt. 8; reissued 1977, vol. 2B), an excellent survey; MARSHALL G.S. HODGSON, *The Venture of Islam,* 3 vol. (1974), a major and influencial study of the religion and civilization; R.M. SAVORY (ed.), *Introduction to Islamic Civilization* (1976), a collection of scholarly articles on Islāmic history, religion, literature, language, and other topics; BERNARD LEWIS (ed.), *The World of Islam: Faith, People, Culture* (U.S. title, *Islam and the Arab World,* 1976), a collection of articles on various aspects of Islāmic culture, and *Islam: From the Prophet Muhammad to the Capture of Constantinople,* 2 vol. (1974), a history composed of translations of original sources; W. MONTGOMERY WATT, *The Majesty That Was Islam: The Islamic World, 661–1100* (1974), a concise introductory history of the rise and decline of the Islāmic Empire; HAMILTON A.R. GIBB, *Mohammedanism,* 2nd ed. (1953, reissued with revisions 1969), a penetrating and concise account of the development of Islām; LOUIS GARDET, *Mohammedanism,* trans. by WILLIAM BURRIDGE (1961), a systematic presentation of Islām, with religious insight; FAZLUR RAHMAN, *Islam,* 2nd ed. (1979), a historical and systematic interpretation of Islām, and *Islamic Methodology in History* (1965), a critical appraisal of the development of *sunnah, ijmāʿ,* and *ijtihād;* REUBIN LEVY, *An Introduction to the Sociology of Islam* (1930–), useful account of the development of Islāmic society and institutions. JOHN W. BAGNOLE, *Cultures of the Islamic Middle East* (1978), an annotated guide to 402 English-language readings for the nonspecialist.

Education: ARTHUR S. TRITTON, *Materials on Muslim Education in the Middle Ages* (1957), an informative, useful compilation; BAYARD DODGE, *Muslim Education in Medieval Times* (1962), a useful sketch.

Political theory and institutions: ERWIN I.J. ROSENTHAL, *Political Thought in Medieval Islam* (1958), a good general survey of the subject.

Islāmic arts: In view of the wealth of descriptive treatments, rather than theory, it is difficult to point to a single source. K.A.C. CRESWELL, *A Bibliography of the Architecture, Arts and Crafts of Islam to 1st Jan. 1960* (1961), and *Supplement, Jan. 1960 to Jan. 1972* (1973), contain all the necessary references. See also his *Early Muslim Architecture,* 2nd ed. (1969), and AMERICAN UNIVERSITY AT CAIRO, CENTER FOR ARABIC STUDIES, *Studies in Islamic Art and Architecture in Honor of Professor K.A.C. Creswell* (1965); HAMILTON A.R. GIBB, *Arabic Literature: An Introduction,* 2nd ed. (1974), a probing survey of 1,500 years of literature; SALIH J. ALTOMA, *Modern Arabic Literature* (1975), a bibliography of 850 general and scholarly works covering 1800–1970.

Muḥammad. *Biographical works:* W. MONTGOMERY WATT, *Muhammad at Mecca* (1953) and *Muhammad at Medina* (1956, reprinted 1977), a full-scale treatment, summarized in *Muhammad: Prophet and Statesman* (1961, reissued 1974), by the same author; FRANTS BUHL, . . . *Das Leben Muhammeds,* 3rd ed. (1961), a German translation of a work published in Danish in 1903 that is still considered reliable; TOR ANDRAE, *Mohammed: The Man and His Faith,* trans. by THEOPHIL MENZEL (1936, reprinted 1971), chiefly concerned with religious aspects; IBN HISHĀM, *The Life of Muhammad: A Translation of Isḥāq's Sīrat rasūl Allāh* (1955, reissued 1967), the primary Arabic biography; NABIA ABBOTT, *Aishah: The Beloved of Mohammed* (1942, reprinted 1973), a scholarly work; JOHN BAGOT GLUBB, *The Life and Times of Muhammad* (1970), a popular account based on the author's familiarity with Arab life.

Primary sources: The Qurʾān, of course, contains basic contemporary materials on Muḥammad, but it is difficult to assess them without broader historical knowledge. The vast collections of Traditions (Ḥadīth), or anecdotes about Muḥammad's words and deeds, are historically disputable and, besides, seldom tell us anything significant about Muḥammad's career. The main sources of historical value are the early biographies (8th–9th century), especially the *Sīrah* of IBN ISḤĀQ, as adapted by IBN HISHĀM, and the *Maghāzi* ("Expeditions") of AL-WAQIDI, together with the supplementary materials recorded by his associate IBN SAʿD. The last item contains much material about the Companions (persons in contact with Muḥammad) and thus about the Prophet's relation to and work with them. Some contemporary documents are preserved in the early biographical works, the most important being the so-called "Constitution of Medina." The latter is in Guillaume's translation of Ibn Isḥāq cited above; other documents are in Watt's *Muhammad at Medina.*

The Qurʾān. The basic work is T. NÖLDEKE, *Geschichte des Qorāns* (1860, 2nd ed. by FRIEDRICH SCHWALLY 1909–38, reprinted 1970). Less comprehensive but more modern are RICHARD BELL, *Introduction to the Qurʾān,* new ed., rev. and enl. by W. MONTGOMERY WATT (1970); and RÉGIS BLACHÈRE, *Introduction au Coran,* 2nd ed. (1977). The history of Qurʾānic interpretation is set forth in IGNÁC GOLDZIHER, *Die Richtungen der islamischen Koranauslegung* (1920, reprinted 1970). It should be supplemented by JOHANNES M.S. BALJON, *Modern Muslim Koran Interpretation, 1880–1960* (1961, reprinted 1968). ARTHUR JEFFERY, *The Qurʾān as Scripture* (1952, reprinted 1980), deals with the Qurʾān's view of its own function. KENNETH CRAGG, *The Event of the Qurʾān* (1971), deals with the meaning of the Qurʾān in Islāmic life. English translations include *The Meaning of the Glorious Qurʾan: Text and Explanatory Translation,* by MUHAMMAD M. PICKTHALL (1938, reissued 1977); *The Koran Interpreted* by ARTHUR J. ARBERRY (1964), which is well known for its literary qualities; and HELMUT GÄTJE, *The Qurʾān and Its Exegesis: Selected Texts with Classical and Modern Muslim Interpretations,* trans. and ed. by ALFORD T. WELCH (1977).

Ḥadīth. J. ROBSON, "Hadīth," in *The Encyclopaedia of Islam,* new ed., vol. 3, pp. 23–28 (1971), and TH.W. JUYNBOLL, "Hadīth," in *The Encyclopaedia of Islam,* vol. 2, pp. 189–194 (1927), two important summaries with extensive bibliographies; ALFRED GUILLAUME, *The Traditions of Islam* (1924, reprinted 1980), a general introduction serviceable for a first study; IGNÁC GOLDZIHER, *Études sur la tradition islamique,* ed. by LÉON BERCHER (1952), a French trans. of the major part of vol. 2 of *Muhammedanische Studien,* 2 vol. (1888–90, Eng. trans. by C.R. BARBER and S.M. STERN, *Muslim Studies,* ed. by S.M. STERN, 1967–71), a classic work on the early development of Ḥadīth, reflecting the early history of religious ideas in Islām; MAULANA MUHAMMAD ALI, *A Manual of Hadith* (1944), a general selection, mainly from al-Bukhārī, in Arabic and English; A.J. WENSINCK, *A Handbook of Early Muhammadan Tradition* (1927, reprinted 1971), an alphabetical arrangement by a great Dutch scholar; MUHAMMAD Z. SIDDIQI, *Ḥādīth Literature: Its Origin, Development, Special Features and Criticism* (1961), an Asian Muslim's presentation.

Theology and philosophy. FRANZ ROSENTHAL (ed.), *The Classical Heritage in Islam,* trans. from the German by EMILE MARMORSTEIN and JENNY MARMORSTEIN (1975), and RICHARD WALZER, *Greek into Arabic: Essays on Islamic Philosophy* (1962, reissued 1970), deal with the Greek and Hellenistic background and its appropriation. PARVIZ MOREWEDGE (ed.), *Islamic Philosophical Theology* (1979), is a major contribution by nine internationally known authorities written for advanced students; W. MONTGOMERY WATT, *The Formative Period of Islamic Thought* (1973), is a study of the evolution of theological thought in the 300 years after Muḥammad's death, and his *Free Will and Predestination in Early Islam* (1948, reissued 1972), is an excellent treatment of the formative period of Islāmic theology; ASAF A.A. FYZEE (ed. and trans.), . . . *a Shīʿite Creed* (1942), is an annotated translation of a standard Shīʿite creed by Ibn Bābawayh; LOUIS GARDET and M.-M. ANAWATI, *Introduction à la théologie musulmane,* 2nd ed. (1970), is a comprehensive handbook on Sunnī theology; and A.J. WENSINCK, *The Muslim Creed* (1932, reprinted 1965), discusses the background and development of Sunnī doctrines. The theology of the Shīʿah is given a prominent place in HENRY CORBIN, *Histoire de la philosophie islamique* (1964–); and its early development is discussed by WILFERD MADELUNG in both *Der Imam al-Qāsim ibn Ibrāhīm und die Glaubenslehre der Zaiditen* (1965), and "Imamism and Muʿtazilite Theology," in *Le Shīʿisme imâmite,* pp. 13–30 (1970). M.M. SHARIF (ed.), *A History of Muslim Philosophy,* 2 vol. (1963–66), is a comprehensive collective work on the history of Islāmic philosophy and related subjects; it is especially useful for the later medieval and modern periods. MAJID FAKHRY, *A History of Islamic Philosophy* (1970), is a general history. FAZLUR RAHMAN discusses the development of the later synthesis between mysticism and philosophy in "Dream, Imagination and ʿĀlam al-Mithāl," *Islamic Studies,* vol. 3, no. 2, pp. 167–180 (June 1964) in the introduction to *Selected Letters of Shaikh Aḥmad Sirhindī* (1968), and in "The Eternity of the World and the Heavenly Bodies in Post-Avicennan Philosophy," in GEORGE F. HOURANI (ed.), *Essays on Islamic Philosophy and Science* (1975), a collection representing recent trends in interpreting Islāmic philosophy.

Islāmic mysticism. *Introductory works:* ARTHUR H. PALMER (comp.), *Oriental Mysticism: A Treatise on Sufiistic and Unitarian Theosophy of the Persians,* 2nd ed. by ARTHUR J. ARBERRY (1938, reprinted 1974), an exposition of later mystical ideas; ANNEMARIE SCHIMMEL, *Mystical Dimensions of Islam* (1975), a multifaceted, introductory study of Ṣūfism; REYNOLD A. NICHOLSON, *The Mystics of Islam* (1914, reprinted 1975), a very readable introduction to classical Ṣūfism and Ṣūfī poetry; ARTHUR J. ARBERRY, *Sufism: An Account of the Mystics of Islam* (1950), a historical survey of classical Ṣūfism; G.-C. ANAWATI and LOUIS GARDET, *Mystique musulmane,* 3rd ed. (1976), an excellent study of the major trends and leading personalities in classical Ṣūfism; ROBERT C. ZAEHNER, *Hindu and Muslim Mysticism* (1960, reissued 1969), a thought-provoking study of the possible relations between Indian and early Muslim mysticism.

History: MARGARET SMITH, *Rābi'a the Mystic & Her Fellow-Saints in Islam: Being the Life and Teachings of Rābi'a al-'Adawiyya al-Qaysiyya of Basra, Together with Some Account of the Place of the Women Saints in Islam* (1928, reprinted 1977), the first study of the herald of mystical love in Islām; JOSEPH VAN ESS, *Die Gedankenwelt des Ḥārit al-Muḥāsibī anhand von übersetzungen aus seinen Schriften dargestellt und erläutert* (1961), an excellent introduction to the theology and psychology of early mystical thought in Islām; LOUIS MASSIGNON, *La Passion de Husayn ibn Mansûr Hallâj: martyr mystique de l'Islam,* new ed. 4 vol. (1975), an indispensable source book for the history of Ṣūfism in the classical period; ANNEMARIE SCHIMMEL, *Al-Halladsch, Märtyrer der Gottesliebe* (1968), a German translation of parts of Ḥallāj's poetry and prose, and a study of his influence on the literatures of the different Islāmic peoples; SERGE DE LAUGIER DE BEAURECUEIL, *Khwādja 'Abdullāh Anṣārī (396–481 H./1006–1089): Mystique Hanbalite* (1965), a biography of the author of the beautiful Persian *munājāt* (prayers) and other mystical books; A.J. WENSINCK, *La Pensée de Ghazzālī* (1940), a short and reliable introduction to Ghazālī's thought; JOHN A. SUBHAN, *Sufism: Its Saints and Shrines* (1938, reissued 1978), a useful survey of the later development of Islāmic mysticism.

Ṣūfī literature: HELMUT RITTER, *Das Meer der Seele* (1955, reissued 1978), an exhaustive work on Farīd ud-Dīn 'Aṭṭār's thought as reflected in his mystical poetry; JALĀLU'DDIN RŪMĪ, *The Mathnawí,* ed. with critical notes, translation, and commentary by REYNOLD A. NICHOLSON, 8 vol. (1925–40), the encyclopaedia of mystical thought in the 13th century in masterly translation; H.T. SORLEY, *Shah Abdul Latīf of Bhit* (1940; reprinted 1966), a study of the greatest mystical poet of Sind.

Ṣūfī thought and practice: BENEDIKT REINERT, *Die Lehre vom Tawakkul in der klassischen Sufik* (1968), the first fundamental study of a single concept central to early Islāmic mysticism, built upon a critical analysis of all available sources; ARTHUR J. ARBERRY, *The Doctrine of the Ṣūfīs* (1935, reprinted 1977), a useful translation of Kalābādhi's *Kitāb at-ta'arruf,* one of the early treatises on Ṣūfī thought; *The Kashf al-Maḥjub: The Oldest Persian Treatise on Sufism* by ALI BIN UTHMAN al-HUJWIRI, trans. by REYNOLD A. NICHOLSON (1911, reprinted 1976), a masterly translation of the voluminous 11th-century account of Ṣūfī thought; G.-H. BOUSQUET (ed.), *Ih'yâ 'ouloûm ed-dîn; ou Vivification des sciences de la foi* (1955), an analytical index of the most widely read work on moderate mystical thought, prepared with the assistance of numerous scholars; CONSTANCE E. PADWICK, *Muslim Devotions* (1961), the only account of the popular mystically tinged piety of the Muslims as reflected in their prayer books; LALEH BAKHTIAR, *Sufi: Expressions of the Mystic Quest* (1976), discusses and shows through illustrations the Ṣūfī experience and its expression in the arts.

Theosophical Ṣūfism: A.E. AFFIFI, *The Mystical Philosophy of Muhyid Dín-Ibnul 'Arabí* (1939, reissued 1974), the first attempt, in a Western language, to systematize the pantheistic system of the 13th-century theosophist; HENRY CORBIN, *Creative Imagination in the Sūfism of Ibn 'Arabī,* trans. by RALPH MANHEIM (1970); REYNOLD A. NICHOLSON, *Studies in Islamic Mysticism* (1921, reissued 1978), a study of Abū Sa'īd and a discussion of Jīlī's Perfect Man and of Ibn al-Fāriḍ, with a superb translation of most of his odes.

Ṣūfī orders: OCTAVE DEPONT and XAVIER CAPPOLANI, *Les Confréries religieuses musulmanes* (1897), a comprehensive account of Ṣūfī brotherhoods; HANS J. KISSLING, "Die Wunder der Derwische," ZDMG (*Zeitschrift der deutschen morgenländischen Gesellschaft),* vol. 107, no. 2, pp. 348–361 (August 1957), a fully documented account of the kinds of miracles performed by dervishes; KHALIQ A. NIZAMI, *The Life and Times of Shaikh Faridud-din Ganj-i-Shakar* (1955, reprinted 1973), a good survey of the life of one of the leading Chishtī saints in India; RENÉ BRUNEL, *Le Monachisme errant dans l'Islam: Sīdi Heddi et les Heddāwa* (1955), a penetrating study of a little known fraternity of dervishes in North Africa; JAMIL M. ABUN-NASR, *The Tijaniyya: A Sufi Order in the Modern World* (1965), a study of the development of political activities of this 19th-century order in the northern and western parts of Africa; J. SPENCER TRIMINGHAM, *The Sufi Orders in Islam* (1971), the first attempt to give a survey of all orders in Islām, and, as such, quite useful.

Islāmic law. A general survey of the Islāmic legal system, covering its historical development, jurisprudential theory, and the most important spheres of the substantive law, is contained in JOSEPH SCHACHT, *An Introduction to Islamic Law* (1964); JAMES N.D. ANDERSON, *Islamic Law in the Modern World* (1959, reprinted 1975); and NOEL J. COULSON, *History of Islamic Law* (1964, reprinted 1971). The reader is referred to the bibliographies of these books, particularly for the numerous articles written by James N.D. Anderson on developments in the law. JOSEPH SCHACHT, *Origins of Muhammadan Jurisprudence* (1950, reissued with corrections and additions, 1967), is a fundamental work of modern research on the early development of legal theory written by the pioneer scholar of this subject. A sound analysis of traditional legal theory is presented in ABDUR RAHIM, *Muhammadan Jurisprudence* (1911, reprinted 1981); and in *The Philosophy of Jurisprudence in Islam,* (1961), an Eng. trans. by FARMAT J. ZIADEH of the Arabic text of an outstanding Muslim jurist, Subhi Mahmassani. MAJID KHADDURI and HERBERT J. LIEBESNY (eds.), *Law in the Middle East* (1955), includes chapters by Muslim scholars and Western Orientalists on the various spheres of substantive Islāmic law, traditional and modern. ASAF A.A. FYZEE, *Outlines of Muhammadan Law,* 4th ed. (1974), is a standard text dealing with Islāmic law as it is applied in India and Pakistan. NORMAN ANDERSON, *Law Reform in the Muslim World* (1976), is a comparative study of the history, philosophy, and achievements of legal reform. *The Encyclopaedia of Islam* (1913–42; new ed., 1960–), contains numerous articles on individual legal topics.

Islāmic myth and legend. TOR ANDRAE, *Die Person Muhammeds in Lehre und Glauben seiner Gemeinde* (1917), on the development of Muḥammad-mysticism; ISRAEL FRIEDLÄNDER, *Die Chadhirlegende und der Alexander-Roman* (1913), on the relation between the Alexander romance and the figure of Khiḍr; MAX J.H. HORTEN, *Die religiöse Gedankenwelt der gebildeten Muslime in heutigen Islam* (1916), an account of popular Islām, and *Die religiöse Gedankenwelt des Volkes im heutigen Islam,* 2 pt. (1917–18), an account of the ideas of educated people in Islām; A.J. WENSINCK, "The Ocean in the Literature of the Western Semites," *Verhandelingen der Koninklijke Akademie van Wetenschappen,* vol. 19, no. 2 (1918), and "The Ideas of the Western Semites Concerning the Navel of the Earth," *ibid.,* vol. 17, no. 1 (1916); SEYYED H. NASR, *Three Muslim Sages* (1964, reissued 1976), an account of the theories of Suhrawardī al-Maqtūl and Ibn 'Arabī; JOSEPH HOROWITZ, "The Growth of the Mohammed Legend," *Moslem World,* vol. 10, no. 1, pp. 49–58 (January 1920), stresses the haggadic influences; WALTHER EICKMANN, *Angelologie und Dämonologie des Korans . . .* (1908), a study of the Qur'ānic concepts of angels and demons; ERNST A. ZBINDEN, *Die Djinn des Islam und der altorientalische Geisterglaube* (1953), a study of the different types of spirits in Islāmic folklore and tradition; RUDOLF KRISS and HUBERT KRISS-HEINRICH, *Volksglaube im Bereich des Islam,* 2 vol. (1960–62), useful studies in Islāmic folklore, with extensive bibliographies; TAUFIC CANAAN, *Mohammedan Saints and Sanctuaries in Palestine* (1927), on Palestinian folklore; articles in the *Shorter Encyclopaedia of Islam* (1953), an authoritative collection of information, each article furnished with an extensive bibliography.

Islāmic Arts

The vast populations of the Middle East and elsewhere that adopted the Islāmic faith from the 7th century onward have created such an immense variety of literatures, performing arts, visual arts, and music that it virtually defies any comprehensive definition. In the narrowest sense, the arts of the Islāmic peoples might be said to include only those arising directly from the practice of Islām; more commonly, however, the term is extended to include all of the arts produced by Muslim peoples, whether connected with their religion or not. In this article, the subject includes the arts created in pre-Islāmic times by Arabs and other peoples in Asia Minor and North Africa who eventually adopted the Islāmic faith. On the other hand, arts produced in cultural areas that were only partially Muslim, such as South Asia, Southeast Asia, and Central Asia, are discussed primarily in articles on arts of those regions.

This article is divided into the following sections:

General considerations

It is difficult to establish a common denominator for all of the artistic expressions of the Islāmic peoples. Such a common denominator would have to be meaningful for miniature painting and historiography, for a musical mode and the form of a poem. The relationship between the art of the Islāmic peoples and its religious basis is anything but direct.

The prohibition on representation

Like most prophetic religions, Islām is not conducive to fine arts. Representation of living beings is prohibited—not in the Qurʾān but in the prophetic tradition. Thus, the centre of the Islāmic artistic tradition lies in calligraphy, a distinguishing feature of this culture, in which the word as the medium of divine revelation plays such an important role. Representational art was found, however, in some early palaces and "at the doors of the bathhouses," according to later Persian poetry. After the 13th century a highly refined art of miniature developed, primarily in the non-Arab countries; it dwells, however, only rarely upon religious subjects. The typical expression of Muslim art is the arabesque, both in its geometric and in its vegetabilic form—one leaf, one flower growing out of the other, without beginning and end and capable of almost innumerable variations—only gradually detected by the eye—which never lose their charm. An aversion to empty spaces distinguishes that art; neither the tile-covered walls of a mosque nor the rich imagery of a poem allows an unembellished area; and the decoration of a carpet can be extended almost without limit.

The centre of Islāmic religion is the clean place for prayer, enlarged into the mosque, which comprises the community and all its needs. The essential structure is similar throughout the Muslim world. There are, of course, period and regional differences—large, wide court mosques of early times; court mosques, with big halls, of Iran and adjacent countries; central buildings with the wonderfully shaped domes of the Ottoman Empire. The implements, however, are the same: a niche (*miḥrāb*)—pointing to Mecca—made of wood, marble, mosaic, stone, tiles; a small pulpit for the Friday sermon; minarets, locally differently shaped but always rising like the call to prayer that is uttered from their tops; the wooden carved stands for the Qur'ān, which is to be written in the most perfect form; sometimes highly artistic lamps (made in Syria and proverbially mentioned all over the Muslim world); perhaps bronze candlesticks, with inlaid ornaments; and rich variations of the prayer mats. If any decoration was needed, it was the words of God, beautifully written or carved in the walls or around the domes. At first connected with the mosques and later independent of them are schools, mausoleums, rooms for the students, and cells for the religious masters.

The poetry of the Arabs consisted in the beginning of praise and satirical poems thought to be full of magic qualities. The strict rules of the outward form of the poems (monorhyme, complicated metre) even in pre-Islāmic times led to a certain formalism and encouraged imitation. Goethe's statement that the stories of *The Thousand and One Nights* have no goal in themselves shows his understanding of the character of Arabic belles lettres, contrasting them with the Islāmic religion, which aims at "collecting and uniting people in order to achieve one high goal." Poets, on the other hand, rove around without any ethical purpose, according to the Qur'ān. For many pious Muslims, poetry was something suspect, opposed to the divine law, especially since it sang mostly of forbidden wine and of free love. The combination of music and poetry, as practiced in court circles and among the mystics, has always aroused the wrath of the lawyer divines who wielded so much authority in Islāmic communities. This opposition may partly explain why Islāmic poetry and fine arts took refuge in a kind of unreal world, using fixed images that could be correctly interpreted only by those who were knowledgeable in the art.

The ambiguity of Persian poetry, which oscillates between the worldly, the divine, and often the political level, is typical of Islāmic writings. Especially in Iran and the countries under its cultural influence, this kind of poetry formed the most important part of literature. Epic poetry of all kinds developed exclusively outside the Arabic-speaking countries; Western readers look in vain for an epical structure in such long poems (as in the case of the prose-romances of the Arabs) and find, instead, a rather aimless representation of facts and fictions. A similar characteristic even conditions innumerable historical works in Arabic, Persian, and Turkish, which, especially in classical times, contain much valuable information, put together without being shaped into a real work of art; only rarely does the historian or philosopher reach a comprehensive view. The first attempt at a philosophy of history, Ibn Khaldūn's *Muqaddimah,* in the 14th century, was rarely studied by his Arab compatriots.

The accumulation of large amounts of material, which is carefully organized up to the present, seems typical of all branches of Islāmic scholarship, from theology to natural sciences. There are many minute observations and descriptions but rarely a full view of the whole process. Later, especially in the Persian, Turkish, and Indo-Muslim areas, a tendency to overstress the decorative elements of prose is evident; and the contents even of official chronicles are hidden behind a network of rhymed prose, which is difficult to disentangle.

The characteristic lack of structure

This tendency is illustrated in all branches of Islāmic art: the lack of "architectural" formation. Instead, there is a kind of carpet-like pattern; the Arabic and Persian poem is, in general, judged not as a closed unity but rather according to the perfection of its individual verses. Its main object is not to convey a deep personal feeling but to perfect to the utmost the traditional rules and inherited metaphors, to which a new image may sometimes be added; thus the personality of the poet becomes visible only through the minimal changes of expression and rhythm and the application of certain preferred metaphors, just as the personality of the miniature painter can be detected by a careful observation of details, of his way of colouring a rock or deepening the shade of a turban. The same holds true for the arabesques, which were developed according to a strict ritual to a mathematical pattern and were refined until they reached a perfection of geometrical complicated figures, as in the dome of the Karatay Medrese in Konya (1251); it corresponds both to the most intricate lacelike Kūfic inscriptions around this dome and to the poetical style of Jalāl ad-Dīn ar-Rūmī, who wrote in that very place and during those years. His immortal mystical poems comprise thousands of variations on the central theme of love. Although such a perfect congruency of poetry and fine arts is not frequently found, the precept about Persian art that "its wings are too heavy with beauty" can also be applied to Persian poetry. Thus, the tile work of a Persian mosque, which combines different levels of arabesque work with different styles of writing, is reminiscent of the way Persian poetry combines at least two levels of reality. And a perfect harmony is reached in some of the miniature manuscripts of Iran, Muslim India, or Ottoman Turkey, which, in their lucid colours and fine details of execution, recall both the perfection of the calligraphy that surrounds them on delicate paper and the subtlety of the stories or poems that they accompany or illustrate.

Those accustomed to the Western ideals of plasticity or form in the fine arts and literature or to the polyphonic interweaving of melodic lines in music have some difficulties in appreciating this art. The palaces seem to be without a fixed architectural plan; rooms and gardens are simply laid out according to daily needs. The historian offers an astounding amount of detailed reports and facts but with no unifying concept. The Muslim writer prefers this carpet-like form; he adds colour to colour, motif to motif, so that the reader only understands the meaning and end of the whole web from a certain distance. Music, differentiated as it may be in the countries between Morocco and India, follows the same model: variations of highest subtlety on a comparatively simple given subject or theme.

Man, a puppet; God, the only actor

Drama and opera in the Western sense did not develop in the Islāmic countries until the 19th century; and the art of the novel is a very recent development. There was no reason for drama: in the Muslim perception God is the only actor who can do whatever he pleases, whose will is inscrutable. Man is, at best, a puppet on a string, behind whose movement those with insight detect the hand of the play master; neither is the problem of personal guilt and absolution posed as it is in the West, nor is a catharsis, or purging of emotion, needed through drama. The atomist theory, widely accepted in Islām since the 10th century, leaves no room for a "dramatic" movement; it teaches that God creates everything anew in every moment, and what is called a "law of nature" is nothing but God's custom, which he can interrupt whenever he pleases.

It is true that certain other forms are found in the more folkloristic arts of Islām. Every region has produced poetry, in regional languages, that is more lively and more realistic than the classical court poetry; but such poetry tends to become restricted to certain fixed forms that can be easily imitated. Attempts at drama in Islām come from these more popular spheres in Iran (and, rarely, in Lebanon and Iraq), where the tragic events of the murder of Ḥusayn (680) at Karbalā' were dramatized in strange forms, using the vocabulary of traditional Persian poetry and theology. Thus, strangely hybrid forms emerge in the Islāmic arts, highly interesting for the historian of religion and the student of literature but not typical of the classical Islāmic ideals. Popular illustrations of tales and legends and those of some of the Shī'ah heroes are similarly interesting but atypical. In modern times, of course, there have been imitations of all forms of Western literary and visual arts: paintings in the Impressionist or Cubist style, the use of free verse instead of the stern classical forms; and novels, dramas, motion pictures, and music combining Western and Eastern modes. Belief in the Qur'ānic dictum "Whatever is on earth will perish save His face" discouraged artistic endeavour on a large scale; but the Prophetic tradition "Verily God is beautiful and loves beauty" has inspired numberless artists and artisans, writers and poets, musicians, and mystics to develop their arts and crafts as a reflection of that divine beauty. A theory of aesthetics comprising the various artistic expressions of the Muslim peoples has yet to be written. Although there have been a number of studies in literary criticism, the formal indebtedness of some of the best modern poets and painters to the Islāmic heritage has never been studied in full.

It is notable that the arts of the Islāmic peoples have had relatively little impact on other cultures, certainly far less than their artistic merit would appear to warrant.

Islāmic art in the West

Europe has known art objects of Islāmic origin since the early Middle Ages, when they were brought home by the crusaders or manufactured by the Arabs in Sicily and Spain. Much admired and even imitated, they formed part of the material culture in those times, so much so that even the coronation robes of the German emperor were decorated with an Arabic inscription. At the same time, Islāmic motives wandered into the belles lettres of Europe, and Islāmic scientific books formed a basis for the development of Western science. Islāmic culture as such, however, was rather an object of hatred than of admiration; a more objective appreciation of both the works of art and of literature did not start until the mid-17th century, when travelers told of the magnificent buildings in Iran and Mughal India, and the first works from Persian literature were translated, influencing German classical literature. Indian miniatures inspired Rembrandt, just as European paintings were imitated by Islāmic, especially Mughal, artists. Persian carpets were among the most coveted gifts for princes and princesses.

A bias against the cultures of the East persisted, however, until after the 18th-century Age of Enlightenment; the indefatigable work of the British scholars at Fort William at Calcutta brought new literary treasures to Europe, where they were studied carefully by specialists in the emerging field of Islāmic studies. Poets such as Goethe in Germany in the early 19th century paved the way for a deeper understanding of Islāmic poetry. Islāmic literatures, however, continue to be known to the larger Western public almost exclusively by *The Thousand and One Nights,* or *The Arabian Nights' Entertainment* (translated first in the early 18th century), Omar Khayyam's *robāʿīyāt,* and the lyrics of Ḥāfeẓ. Even experts who are aware of the immense wealth of the literatures in the different Islāmic languages (such as Arabic, Persian, Turkish, and Urdu) until now have rarely appreciated the literatures from an aesthetic viewpoint; rather, they have used them as a source for lexicography and for philological and historical research. The situation in Islāmic fine arts and architecture is similar. Although the beauty of the Alhambra, for example, had already inspired European scholars and artists in the early 19th century, a thorough study of Islāmic art as an independent field began only in the 20th century. There was even less interest in the music of Islāmic peoples, the arabesque-like uniformity of which seems strange to Western ideals of harmony.

Islāmic literatures

NATURE AND SCOPE

It would be almost impossible to make an exhaustive survey of Islāmic literatures. There are so many works, of which hundreds of thousands are available only in manuscript, that even a very large team of scholars could scarcely master a single branch of the subject. Islāmic literatures, moreover, exist over a vast geographical and linguistic area, for they were produced wherever the Muslims went, pushing out from their heartland in Arabia through the countries of the Near and Middle East as far as Spain, North Africa, and, eventually, West Africa. Iran (Persia) is a major centre of Islām, along with the neighbouring areas that came under Persian influence, including Turkey and the Turkic-speaking peoples of Central Asia. Many Indian vernaculars contain almost exclusively Islāmic literary subjects; there is an Islāmic content in the literature of Malaysia and in that of some East African languages, including Swahili. In many cases, however, the Islāmic content proper is restricted to religious works—mystical treatises, books on Islāmic law and its implementation, historical works praising the heroic deeds and miraculous adventures of earlier Muslim rulers and saints, or devotional works in honour of the prophet Muḥammad.

The vast majority of Arabic writings are scholarly—the same, indeed, is true of the other languages under discussion. There are superb, historically important translations made by medieval scholars from Greek into Arabic; historical works, both general and particular; a range of religiously inspired works; books on grammar and on stylistics, on ethics and on philosophy. All have helped to shape the spirit of Islāmic literature in general, and it is often difficult to draw a line between such works of "scholarship" and works of "literature" in the narrower sense of that term. Even a strictly theological commentary can bring about a deeper understanding of some problem of aesthetics. A work of history composed in florid and "artistic" language would certainly be regarded by its author as a work of art as well as of scholarship, whereas the grammarian would be equally sure that his keen insights into the structure of Arabic grammar were of the utmost importance in preserving that literary beauty in which Arabs and non-Arabs alike took pride.

Pride in literary beauty

In this treatment of Islāmic literatures, however, the definition of "literature" is restricted to poetry and belles lettres, whether popular or courtly in inspiration. Other categories of writing will be dealt with briefly if these shed light on some peculiar problem of literature.

The range of Islāmic literatures. Although Islāmic literatures appear in such a wide range of languages and in so many different cultural environments, their unity

> is safeguarded by the identity of the basic existential experience, by the identity of the fundamental intellectual interests, by the authoritativeness of certain principles of form and presentation, not to mention the kindred political and social organization within which those peoples aspire to live.

Arabic: language of the Qurʾān. The area of Islāmic culture extends from western Africa to Malaysia, Indonesia, and the Philippines; but its heartland is Arabia, and the prime importance and special authority of the Arabic language was to remain largely unquestioned after the spread of Islām. The Arabic poetry of pre-Islāmic Arabia was regarded for centuries afterward as the standard model for all Islāmic poetic achievement, and it directly influenced literary forms in many non-Arab literatures. The Qurʾān, Islām's sacred scripture, was accepted by pious Muslims as God's uncreated word and was considered to be the highest manifestation of literary beauty. A whole literature defended its inimitability (*iʿjāz*) and unsurpassable beauty. Because it was God's own word, the Qurʾān could not legitimately be translated into any other language; the study of at least some Arabic was therefore required of every Muslim. Arabic script was used by all those peoples who followed Islām, however much their own languages might differ in structure from Arabic. The Qurʾān became the textbook of the Muslims' entire philosophy of life; theology, lexicography, geography, historiography, and mysticism all grew out of a deep study of its form and content; and even in the most secular works there can be found allusions to the holy book. Its imagery not unexpectedly permeates all Islāmic poetry and prose.

Allusions to the Qurʾān in all literary genres

Between the coming of Islām in the 7th century and the 11th, a great deal of poetry and prose in Arabic was produced. One branch of literature in Spain and North Africa matured in perfect harmony with the classical ideals of the Muslim East although its masters, during the 11th and 12th centuries, invented a few strophic forms unknown to classical Arabic poetry. In modern times, North African Muslim literature—mainly from Algeria and Morocco—often uses French as a means of expression, since the tradition of Arabic writing was interrupted by the French occupation in the 19th century and has had to be built up afresh.

Persian. In 641 the Muslims entered Iran, and Persian influence on literary taste becomes apparent in Arabic literature from the mid-8th century onward. Many stories and tales were transmitted from, or through, Iran to the Arab world and often from there to western Europe. Soon Iran could boast a large literature in its own tongue. Persian literature was more varied in its forms and content than that written in classical Arabic. Although Persian adopted many of the formal rules of the Arabic language (including prosody and rhyme patterns), new genres, including epic poetry, were introduced from Iran. The lyric, elegant and supple, also reached its finest expression in the Persian language.

South Asian. Persian culture was by no means restricted to Iran itself. Northwestern India and what is now Pakistan became a centre of Islāmic literature as early as the 11th century, with Delhi and Agra being of special importance. It was to remain a stronghold of Muslim cultural life, which soon also extended to the east (Bengal) and south (Deccan). Persian remained the official language of Muslim India until 1835, and not only its poetry but even its historiography was written in the high-flown manner that exemplified the Persian concept of fine style. Muslim India can further boast a fine heritage of Arabic poetry and prose (theological, philosophical, and mystical works).

At various times in its history the Indian subcontinent was ruled by princes of Turkish origin (indeed, the words *Turk* and *Muslim* became synonymous in some Indian languages). The princes surrounded themselves with a military aristocracy of mainly Turkish extraction, and thus a few poetical and prose works in Turkish were written at some Indian courts. In various regions of the subcontinent an extremely pleasing folk literature has flourished throughout the ages: Sindhi in the lower Indus Valley, for example, and Punjabi in the Punjab are languages rich in an emotional poetry that uses popular metres and forms. At the Indo-Iranian border the oldest fragments of the powerful Pashto poetry date from the Middle Ages. The neighbouring Baluchi poetry consists largely of ballads and religious folksongs. All the peoples in this area have interpreted Islāmic mysticism in their own simple, touching imagery. In the east of the subcontinent, Bengali Muslims possess a large Islāmic literary heritage, including religious epics from the 14th and 15th centuries and some lovely religious folksongs. The achievements of modern novelists and lyric poets from Bangladesh are impressive. To the north, where Islām came in the 14th century, a number of classical themes in Islāmic lore were elaborated in Kashmiri lyric and epic poetry. To the south, an occasional piece of Islāmic religious poetry can be found even in Tamil and Malayalam. Some fine Muslim short stories have been produced in modern Malayalam.

Islāmic mystical thought in folk literature

Urdu, now the chief literary language of Muslim India and Pakistan, borrowed heavily from Persian literature during its classical period in the 18th century. In many writings only the verbs are in Urdu, the rest consisting of Persian constructions and vocabulary; and the themes of traditional Urdu literature were often adapted from Persian. Modern Urdu prose, however, has freed itself almost completely from the past, whereas in poetry promising steps have been taken toward modernization of both forms and content (see SOUTH ASIAN ARTS).

Turkish. An elaborate "classical" style developed in Turkish after the 14th century, reaching its peak in the 17th. Like classical Urdu, it was heavily influenced by Persian in metrics and vocabulary. Many exponents of this "high" style came from the Balkan provinces of the Ottoman Empire. On the other hand, a rich and moving folk poetry in popular syllable-counting metres has always flourished among the Turkish population of Anatolia and Rumelia. The mystical songs of their poet Yunus Emre (died *c.* 1321) contributed greatly toward shaping this body of literature, which was preserved in the religious centres of the Ṣūfī orders of Islām. From this folk tradition, as well as from Western literature, modern Turkish literature has derived a great deal of its inspiration.

Turkic languages. A great deal of the Muslim literature of Central Asia is written in Turkic languages, which include Uzbek, Tatar, and Kirgiz. Its main cultural centres (Samarkand, Bukhara, Fergana) became part of the Muslim empire after 711. Central Asia was an important centre of Islāmic learning until the Tsarist invasions in the 1870s, and the peoples of this region have produced a classical literature in Arabic. Many of the most famous Arabic and Persian scholars and poets writing in the heyday of Muslim influence were Central Asians by birth. Central Asians also possess a considerable literature of their own, consisting in large part of epics, folktales, and mystical "words of wisdom." The rules of prosody which hold for Arabic and Persian languages have been deliberately imposed on the Turkic languages on several occasions, notably by ʿAlī Shīr Navāʾī (died 1501), a master of Chagatai poetry and prose in Herāt, and by Bābur (died 1530), the first Mughal emperor in India. Tadzhik literature is basically Persian, both as it is written today in the Tadzhik Soviet Socialist Republic and as it existed in earlier forms, when it was indistinguishable from classical Persian. After the Russification of the country, and especially after the 1917 Revolution, a new literature emerged that is part and parcel of the Soviet Union's literature. The same can be said, by and large, about the literatures of other Muslim Turkic peoples of Central Asia.

Other languages. Smaller fragments of Islāmic literature, in Chinese, are found in China (which has quite a large Muslim population) and in the Philippines. The literary traditions of Indonesia and of Malaysia, where the religion of Islām arrived long ago, are also worth noting. Historical and semimythical tales about Islāmic heroes are a feature of the literature in these areas, a fact of immense interest to folklorists.

Contact with Islām and its "written" culture also helped to preserve national idioms in many regions. Often such idioms were enriched by Arabic vocabulary and Islāmic concepts. The leaders of the Muslims in such areas in northern Nigeria, for example, preferred to write poetry and chronicles in Arabic, while using their mother tongue for more popular forms of literature (see AFRICAN ARTS). Of particular interest in this connection is Kurdish literature, which has preserved in an Iranian language several important, popular heterodox texts and epics.

Islāmic literatures and the West. Small fragments of Arabic literature have long been known in the West. There were cultural interrelations between Muslim Spain (which, like the Indus Valley, became part of the Muslim empire after 711) and its Christian neighbours, and this meant that many philosophical and scientific works filtered through to western Europe. It is also likely that the poetry of Muslim Spain influenced the growth of certain forms of Spanish and French troubadour poetry and provided an element, however distorted, for medieval Western romances and heroic tales.

First Western studies

Investigation of Oriental literatures by Western scholars did not begin until the 16th century in the Netherlands and England. First attempts toward an aesthetic understanding of Arabic and Persian poetry came even later: they were made by the British Orientalists of Fort William, Calcutta, and by German pre-Romantics of the late 18th century. In the first half of the 19th century the publication of numerous translations of Oriental poetry, especially into German, began to interest some Europeans. The poetical translations from Arabic, Persian, and Sanskrit made by the German Orientalist and poet Friedrich Rückert can scarcely be surpassed, either in accuracy or in poetical mastery. The Persian poet Ḥāfeẓ became well known in German-speaking countries, thanks to Johann Wolfgang von Goethe's enchanting poems, *West-östlicher Divan* (1819), a collection which was the first response to Persian poetry and the first aesthetic appreciation of the character of Oriental poetry by an acknowledged giant of European literature. An "Orientalizing style," which employed Arabo-Persian literary forms such as the *ghazal* (a short, graceful poem with monorhyme), became fashionable at times in Germany. Later, Edward FitzGerald aroused new interest in Persian poetry with his free adaptations of Omar Khayyam's *robāʿīyāt* (*The Rubáiyát of Omar Khayyám,* 1859). The fairy tales known as *The Thousand and One Nights,* first translated in 1704, provided abundant raw material for many a Western writer's play, novel, story, or poem about the Islāmic East.

EXTERNAL CHARACTERISTICS

In order to understand and enjoy Oriental literature, the external characteristics of it have to be studied most carefully. The literatures of the Islāmic peoples are "intellectual"; in neither poetry nor prose are there many examples of subjective lyricism, as it is understood in the West. The principal genres, forms, and rules were inherited from pre-Islāmic Arabic poetry but were substantially elaborated afterward, especially by the Persians.

Rhyme and metre. Arabic poetry is built upon the principle of monorhyme, and the single rhyme, usually consisting in one letter, is employed throughout every poem, long or short. The structure of Arabic permits such monorhymes to be achieved with comparative ease. The Persians and their imitators often extended the rhyming part over two or more syllables (*radīf*) or groups of words, which are repeated after the dominant rhyming consonant. The metres are quantitative, counting long and short syllables (*ʿarūḍ*). Classical Arabic has 16 basic metres in five groupings; they can undergo certain variations, but the poet is not allowed to change the metre in the course of his poem. Syllable-counting metres, as well as strophic forms, are used in popular, or "low," poetry; only in post-classical Arabic were some strophic forms introduced into "high" poetry. Many modern Islāmic poets, from Pakistan to Turkey and North Africa, have discarded the classical system of prosody altogether. In part they have substituted verse forms imitating Western models such as strophic poems with or without rhyme; since about 1950 free verse has almost become the rule, although a certain tendency toward rhyming or to the use of alliterative quasi-rhymes can be observed.

Genres. The chief poetic genres, as they emerged according to traditional rules, are the *qaṣīdah,* the *ghazal,* and the *qiṭʿah;* in Iran and its adjacent countries there are, further, the *robāʿī* and the *mas̄navī.*

Qaṣīdah. The *qaṣīdah* (literally "purpose poem"), a genre whose form was invented by pre-Islāmic Arabs, has from 20 to more than 100 verses and usually contains an account of the poet's journey. In the classic pattern, the parts followed a fixed sequence, beginning with a love-poem prologue (*nasīb*), followed by a description of the journey itself, and finally reaching its real goal by flattering the poet's patron, sharply attacking some adversaries of his tribe, or else indulging in measureless self-praise. Everywhere in the Muslim world the *qaṣīdah* became the characteristic form for panegyric. It could serve for religious purposes as well: solemn praise of God, eulogies of the Prophet, and songs of praise and lament for the martyr heroes of Shīʿah Islām were all expressed in this form. Later, the introductory part of the *qaṣīdah* often was taken up by a description of nature or given over to some words of wisdom; or the poet took the opportunity to demonstrate his skill in handling extravagant language and to show off his learning. Such exhibitions were made all the more difficult because, though it varied according to the rank of the person to whom it was addressed, the vocabulary of each type of *qaṣīdah* was controlled by rigid conventions. This type of poetry, however, could obviously lend itself easily to empty verbosity or to pedantry.

The love poem

Ghazal. The *ghazal* possibly originated as an independent elaboration of the *qaṣīdah*'s introductory section, and it usually embodies a love poem. Ideally, its length varies between five and 12 verses. It can be used either for religious or secular expression, the two often being blended indistinguishably. Its diction is light and graceful, its effect comparable to that of chamber music, whereas the *qaṣīdah*-writer employs, so to speak, the full orchestral resources.

Qiṭʿah. Monorhyme is used in both the *qaṣīdah* and *ghazal.* But while these two forms begin with two rhyming hemistiches (half-lines of a verse), in the *qiṭʿah* ("section") the first hemistich does not rhyme, and the effect is as though the poem had been "cut out" of a longer one (hence its name). The *qiṭʿah* is a less serious literary form that was used to deal with aspects of everyday life; it served mainly for occasional poems, satire, jokes, word games, and chronograms.

Robāʿī. The form of the *robāʿī,* which is a quatrain in fixed metre with a rhyme scheme of *a a b a,* seems to go back to pre-Islāmic Persian poetical tradition. It has supplied the Persian poets with a flexible vehicle for ingenious aphorisms and similarly concise expressions of thought for religious, erotic, or skeptical purposes. The peoples who came under Persian cultural influence happily adopted this form.

Mas̄navī. Epic poetry was unknown to the Arabs, who were averse to fiction, whether it was expressed in poetry or in prose. The development of epic poetry was thus hindered, just as was the creation of novels or short stories. Nevertheless, *mas̄navī*—which means literally "the doubled one," or rhyming couplet, and by extension a poem consisting of a series of such couplets—became a favourite poetical form of the Persians and those cultures they influenced. The *mas̄navī* enabled the poet to develop the thread of a tale through thousands of verses. Yet even in such poetry, only a restricted number of metres was employed, and no metre allowed more than 11 syllables in a hemistich. Metre and diction were prescribed in accordance with the topic; a didactic *mas̄navī* required a style and metre different from a heroic or romantic one. The *mas̄navī* usually begins with a praise of God, and this strikes the keynote of the poem.

Other poetic forms. There is a variety of other forms that are more or less restricted to folk poetry, such as the *sīḥarfī* ("golden alphabet"), in which each line or each stanza begins with succeeding letters of the Arabic alphabet. In Muslim India the *bārāmāsa* ("12 months") is a sort of lovers' calendar in which the poet, assuming the role of a young woman of longing, expresses the lover's feelings in accord with the seasons of the year. Apart from these, later writers tried to develop strophic forms. Sometimes *ghazal*s with the same metre were bound together as "stanzas" to form a longer unit through the use of a linking verse. When the linking verse was recurrent, the poem was called a *tarjīʿ-band* (literally "return-tie"); when the linking verse was varied, the poem was called a *tarkīb-band* (literally "composite-tie"). True stanzas of varying lengths were also invented. Among these, mainly in Urdu and Turkish, a six-line stanza known as *musaddas* became the form used for the *mars̄īyeh* (dirge for the martyrs of Karbalāʾ). Because it had come to be associated with lofty feeling and serious thought, *musaddas* later was used for the first reformist modern poems.

Rhyming prose

The Arabs inherited a love for rhymed prose from pre-Islāmic Arabia. Although the extent of prose literature, even in the field of belles lettres, is very large, the novel and novella were introduced only after contact with European literatures.

Maqāmah. The most typical expression of the Arabic—and Islāmic—spirit in prose is the *maqāmah* (gathering, assembly), which tells basically simple stories in an extremely and marvelously complicated style (abounding in word plays, logographs, double entendre, and the like) and which comes closest to the Western concept of the short story.

The versatility and erudition of the classical *maqāmah* authors is dazzling, but the fables and parables that, during the first centuries of Islām, had been told in a comparatively easy flowing style, later became subject to a growing trend toward artificiality, as did almost every other literary genre, including expository prose. Persian historiographers and Turkish biographers, Indo-Muslim writers on mysticism and even on science all indulged in a style in which rhyme and rhetoric often completely obscured the meaning. It is only since the late 19th century that a matter-of-fact style has slowly become acceptable in literary circles; the influence of translations from European languages, the role of journalism, and the growing pride in a pure language freed from the cobwebs of the past worked together to make Islāmic languages more pliable and less artificial.

Imagery. In all forms of poetry and in most types of prose, writers shared a common fund of imagery that was gradually refined and enlarged in the course of time. The main source of imagery was the Qurʾān, its figures and utterances often divested of their sacred significance. Thus, the beautiful Joseph (*sūrah* 12) is a fitting symbol for the handsome beloved; the nightingale may sing the psalms of David (*sūrah* 21:79 *a.o*); the rose sits on Solomon's wind-borne throne (*sūrah* 21:81 *a.o*), and its opening petals can be compared to Joseph's shirt rent by Potiphar's wife (*sūrah* 12:25 ff.), its scent to that of Joseph's shirt, which cured blind Jacob (*sūrah* 12:94). The tulip reminds the poet of the burning bush before which Moses stood (*sūrah* 20:9 ff.), and the coy beloved refuses the lover's demands by answering, like God to Moses, "Thou shalt not see me" (*sūrah* 7:143); but her (or his) kiss gives the dying lover new life, like the breath of Jesus (*sūrah* 3:49). Classical Persian poetry often mentions knights and kings from Iran's history alongside those from Arabic heroic tales. The cup of wine offered by the "old man of the Magians" is comparable to the miraculous cup owned by the Iranian mythical king Jamshīd or to Alexander's mirror, which showed the marvels of the world; the nightingale may sing "Zoroastrian tunes" when it contemplates the "fire temple of the rose." Central scenes from the great Persian *mas̄navīs* contributed to the imagery of later writers in Persian-, Turkish-, and Urdu-speaking areas. Social and political conditions are reflected in a favourite literary equation between the "beautiful and cruel beloved" and "the Turk": since in Iran and India the military caste was usually of Turkish origin, and since the Turk was always considered "white" and handsome, in literary imagery he stood as the "ruler of hearts." Minute arabesque-like descriptions of nature, particularly of garden scenes, are frequent: the rose and the nightingale have almost become substitutes for mythological figures. The versatile writer was expected to introduce elegant allusions to classical Arabic and Persian literature and to folklore and to know enough about astrology, alchemy, and medicine to use the relevant technical terms accurately. Images inspired by the pastimes of the grandees—chess, polo, hunting, and the like—were as necessary for a good poem as were those referring to music, painting, and calligraphy. Similarly, allusions in poetic imagery to the Arabic letters—often thought to be endowed with mystical significance or magical properties—were very common in all Islāmic literatures. The poet had to follow strict rules laid down by the masters of rhetoric, rigidly observing the harmonious selection of similes thought proper to any one given sphere (four allusions to Qur'ānic figures, for example; or three garden images all given in a single verse). The poet was expected to invent new fantastic etiologies (*ḥosn-e ta‘līl*): he had to describe natural phenomena in some elegant and surprising metaphor. Thus, "The narcissus has strewn silver in the way of the bride rose . . ." means simply "The narcissus has withered"—for when the rose (dressed in red, like an Oriental bride) appears in late spring it is time for the narcissus to shed its white petals, just as people would shed silver coins in the way of a bridal procession.

Qualities expected of the versatile writer

Skills required of the writer. The writer was also expected to use puns and to play with words of two or more meanings. He might write verses that could provide an intelligible meaning even when read backward. He had to be able to handle chronograms, codes based on the numerical values of a phrase or verse, which, when understood, gave the date of some relevant event. Later writers sometimes supplied the date of a book's compilation by hiding a chronogram in its title. A favourite device in poetry was the "question and answer" form, employed in the whole poem, or only in chosen sections.

One was expected to show his talent at both improvisation and elaboration on any theme if he wished to attract the interest of a generous patron. His poetry was judged according to the perfection of its individual verses. Only in rare cases was the poem appreciated as a whole: the lack of coherent argument, which often puzzles the Western reader in *ghazal* poetry, is in fact deliberate.

It would be idle to look for the sincere expression of personal emotion in Arabic, Turkish, or Persian poetry. The conventions are so rigid that the reader is allowed only a rare glimpse into the poet's feelings. Indeed, such feelings were put through the sieve of intellect, and personal experiences were thereby transformed into arabesque-like work of artistry, if not art. In the hands of mediocre versifiers and prose writers, however, literature became mannered and completely artificial. The reader soon tires of the constantly recurring moon faces, hyacinth curls, ruby lips, and cypress statures (that is, tall and slender). Yet the great masters of poetry and rhetoric (who all have their favourite imagery, rhymes, and rhythmical patterns) will sometimes allow the patient reader a glimpse into their hearts by a slight rhythmical change or by a new way of expressing a conventional thought.

These are, of course, quite crude generalizations. Folk poetry, for instance, has to be judged by different standards, though even here conventional forms and inherited imagery make it, on the whole, more standardized than might be wished. Only in the 20th century has a complete break with classical ideals been made—sincerity instead of monotonous imitation, political and social commitment instead of empty panegyric, realism instead of escapism: these are the characteristic features of modern literatures of the Muslim countries.

HISTORICAL DEVELOPMENTS: PRE-ISLĀMIC LITERATURE

The first known poetic compositions of the Arabs are of such perfect beauty and, at the same time, are so conventionalized, that they raise the question as to how far back an actual poetic tradition does stretch. A great number of pre-Islāmic poems, dating from the mid-6th century, were preserved by oral tradition. The seven most famous pieces are *al-Mu‘allaqāt* ("The Suspended Ones," known as *The Seven Odes*), and these are discussed more fully below. The term *mu‘allaqāt* is not fully understood: later legend asserts that the seven poems had been hung in the most important Arab religious sanctuary, the Ka‘bah in Mecca, because of their eloquence and beauty and had brought victory to their authors in the poetical contests traditionally held during the season of pilgrimage. Apart from these seven, quite a number of shorter poems were preserved by later scholars. An independent genre in pre-Islāmic poetry was the elegy, often composed by a woman, usually a deceased hero's sister. Some of these poems, especially those by the poetess al-Khansā' (died after 630) are notable for their compact expressiveness.

The first poetry

Poetry. The poet (called a *shā‘ir,* a wizard endowed with magic powers) was thought to be inspired by a spirit (*jinn, shayṭān*). The poet defended the honour of his tribe and perpetuated their deeds. Religious expression was rare in pre-Islāmic poetry. In the main it reflects the sense of fatalism that was probably needed if the harsh circumstances of Bedouin life in the desert were to be endured.

The most striking feature of pre-Islāmic poetry is the uniformity and refinement of its language. Although the various tribes, constantly feuding with one another, all spoke their own dialects, they shared a common language for poetry whether they were Bedouins or inhabitants of the small capitals of al-Ḥīrah and Ghassān (where the influence of Aramaic culture was also in evidence).

Arabic was even then a virile and expressive language, with dozens of synonyms for the horse, the camel, the lion, and so forth; and it possessed a rich stock of descriptive adjectives. Because of these features, it is difficult for foreigners and modern Arabs alike to appreciate fully the artistic qualities of early Arabic poetry. Imagery is precise, and descriptions of natural phenomena are detailed. The sense of universal applicability is lacking, however, and the comparatively simple literary techniques of simile and metaphor predominate. The imaginative power that was later to be the hallmark of Arabic poetry under Persian influence had not yet become evident.

The virility of Arabic

The strikingly rich vocabulary of classical Arabic, as well as its sophisticated structure, is matched by highly elaborate metrical schemes, based on quantity. The rhythmical structures were analyzed by the grammarian Khalīl of Basra (died *c.* 791), who distinguished 16 metres. Each was capable of variation by shortening the foot or part of it; but the basic structure was rigidly preserved. One and the same rhyme letter had to be maintained throughout the poem. (The rules of rhyming are detailed and very complicated but were followed quite strictly from the 6th to the early 20th century.)

As well as rules governing the outward form of poetry, a system of poetic imagery already existed by this early period. The sequence of a poem, moreover, followed a fixed pattern (such as that for the *qaṣīdah*). Pre-Islāmic poetry was not written down but recited; and therefore sound and rhythm played an important part in its formation, and the *rāwīs* (reciters) were equally vital to its preservation. A *rāwī* was associated with some famous bard and, having learned his master's techniques, might afterward become a poet himself. This kind of apprenticeship to a master whose poetic style was thus continued became a common practice in the Muslim world (especially in Muslim India) right up to the 19th century.

The seven authors of *The Seven Odes*

From pre-Islāmic times the seven authors of *The Seven Odes,* already described, are usually singled out for special praise. Their poems and miscellaneous verses were collected during the 8th century and ever since have been the subject of numerous commentaries in the East. They have been studied in Europe since the early 19th century.

The poet Imru' al-Qays (died *c.* AD 550), of the tribe of Kindah, was foremost both in time and in poetic merit. He was a master of love poetry; his frank descriptions of dalliance with his mistresses are considered so seductive that (as orthodox Puritanism claims) the Prophet Muḥammad called him "the leader of poets on the way to Hell." His style is supple and picturesque. It grips the attention whether his poems sing of his love adventures or describe a seemingly endless rainy night. Of all classical Arabic poets he is probably the one who appeals most to modern taste. At the other extreme stands Zuhayr, praising the chiefs of the rival tribes of ʿAbs and Dhubyān for ending a long feud. He is chiefly remembered for his serious *qaṣīdah* in which, old, wise, and experienced, he meditates upon the terrible escalation of war. Various aspects of Bedouin life, as well as the attitude of the Arabs to the rulers of the small kingdom of al-Ḥīrah on the Euphrates, are reflected in the poems of an-Nābighah adh-Dhubyānī, ʿAmr, and Ṭarafah. The boastful pride of the self-centred Arab warrior can be observed best in the poems of al-Ḥārith, who became proverbial for his arrogance. ʿAntarah, son of a black slave girl, won such fame on the battlefield and for his poetry that he later became the hero of an Arabic folk romance.

Two other masters can stand beside these seven. Exciting for their savagery and beauty are some poems by Ta'abbaṭa Sharran and Shanfarā, both outlaw warriors. Their verses reveal the wildness of Bedouin life, with its ideals of bravery, revenge, and hospitality. Ta'abbaṭa Sharran is the author of a widely translated "Song of Revenge" (for his uncle), composed in a short, sharp metre. Shanfarā's *lāmīyah* (literally "poem rhyming in l") vividly, succinctly, and with a wealth of detail tells of the experiences to be had from life in the desert. This latter poem has sometimes been considered a forgery, created by a learned grammarian. The suggestion highlights the question, often posed, of how much pre-Islāmic poetry is genuine and how much is the product of later scholars. Some modern critics—without proper justification—would dismiss the entire *corpus* as counterfeit.

Prose. While poetry forms the most important part of early Arabic literature and is an effective historical preservation of the Arabs' past glory, there is also a quantity of prose. Of special interest is the rhymed prose (*sajʿ*) peculiar to soothsayers, which developed into an important form of ornate prose writing in every Islāmic country. Tales about the adventures and battle days of the various tribes (*ayyām al-ʿArab,* or "The Days of the Arabs") were told and handed down from generation to generation, usually interspersed with pieces of poetry. Proverbs and proverbial sayings were as common as in most cultures at a comparable level of development. The "literary" genre most typical of Bedouin life is the *musāmarah,* or "nighttime conversation," in which the central subject is elaborated not by plot but by carrying the listener's mind from topic to topic through verbal associations. Thus, the language as language played a most important role. The *musāmarah* form inspired the later *maqāmah* literature.

The *musāmarah:* "nighttime conversation"

It has been said—and this certainly holds true for the *musāmarah*—that Arabic literature demands attention from its listeners only in short bursts; for listeners are carried from verse to verse, from anecdote to anecdote, from pun to pun, along a theme whose broad outline is entirely familiar. Western Orientalists have for this reason spoken of the "molecular," or "atomic," structure both of classical Arabic literature and of traditional Islāmic thought. An audience listening to one of the ancient bards—or to a modern poet or orator in the Muslim world—would be able to listen without tiring. The sheer emotive power of the Arabic language to enrapture and bewitch its listeners by sound alone should be kept in mind when considering any piece of Arabic literature. Only a people endowed with peculiar sensibility to the word could properly appreciate the refinement of pre-Islāmic poetry and be ready to accept the concept of divine revelation appearing through the word in the Qur'ān.

EARLY ISLĀMIC LITERATURE

With the coming of Islām the attitude of the Arabs toward poetry seems to have changed. The new Muslims, despite their long-standing admiration for powerful language, often shunned poetry as reminiscent of pagan ideals now overthrown. For the Qur'ān, in *sūrah* 26:225 ff., condemned the poets "who err in every valley, and say what they do not do. Only the perverse follow them!" The Qur'ān, as the uncreated word of God, was now considered the supreme manifestation of literary beauty. It became the basis and touchstone of almost every cultural and literary activity and attained a unique position in Arabic literature.

Age of the caliphs. It might be expected that a new and vigorous religion would stimulate a new religious literature to sing of its greatness and glory. This, however, was not the case. Maybe the once boastful poets felt, at least for a while, that they were nothing but humble servants of Allāh. At any rate, no major poet was inspired by the birth and astonishingly rapid expansion of Islām. Only much later did poets claim that their work was the "heritage of prophecy" or draw upon a tradition that calls the tongues of the poets "the keys of the treasures beneath the Divine Throne." The old, traditional literary models were still faithfully followed: a famous ode by Kaʿb, the son of Zuhayr, is different from pre-Islāmic poetry only insofar as it ends in praise of the Prophet, imploring his forgiveness, instead of eulogizing some Bedouin leader. Muḥammad's rather mediocre eulogist, Ḥassān ibn Thābit (died *c.* 659), also slavishly repeated the traditional patterns (even including the praise of wine that had been such a common feature of pre-Islāmic poetry at the court of al-Ḥīrah, despite the fact that wine had been by then religiously prohibited).

Influence of pre-Islāmic poetry

Religious themes are to be found in the *khuṭbah*s, or Friday sermons, which were delivered by governors of the provinces. In these *khuṭbah*s, however, political considerations frequently overshadow the religious and literary aspects. The *quṣṣāṣ* (storytellers), who interpreted verses from the Qur'ān, attracted large audiences and may be regarded as the inventors of a popular religious prose. Their interpretations were highly fanciful, however, and hardly squared with the theologian's orthodoxy.

The desire to preserve words of wisdom is best reflected in the sayings attributed to ʿAlī, the fourth caliph (died 661). These, however, were written down, in superbly concise diction, only in the 10th century under the title *Nahj al-balāghah* ("The Road of Eloquence"), a work that is a masterpiece of the finest Arabic prose and that has inspired numerous commentaries and poetical variations in the various Islāmic languages.

Umayyad dynasty. The time of the "Four Righteous Caliphs," as it is called, ended with ʿAlī's assassination in 661. The Umayyad dynasty then gained the throne, and a new impetus in poetry soon became perceptible. The Umayyads were by no means a pious dynasty, much enjoying the pleasures of life in their residence in Damascus and in their luxurious castles in the Syrian desert. One of their last rulers, the profligate al-Walīd ibn Yazīd (died 744), has become famous not so much as a conqueror (although in 711 the Muslims reached the lower Indus basin, Transoxania, and Spain) but as a poet who excelled in frivolous love verses and poetry in praise of wine. He was fond of short, light metres to match his subjects and rejected the heavier metres preferred by *qaṣīdah* writers. His verses convey a sense of ease and gracious living. Al-Walīd was not, however, the first to attempt this kind of poetry: a remarkable poet from Mecca, ʿUmar ibn Abī Rabīʿah (died *c.* 712 or 720), had contributed in large measure to the separate development of the love poem (*ghazal*) from its subordinate place as the opening section of the *qaṣīdah.* Gentle and charming, in attractive and lively rhythms, his poems sing of amorous adventures with the ladies who came to Mecca on pilgrimage. His gay, melodious poems still appeal to modern readers.

The love poets of Medina

In Medina, on the other hand, idealized love poetry was the vogue; its invention is attributed to Jamīl (died 701), of the tribe ʿUdhrah, "whose members die when they love." The names of some of these "martyrs of love," together with the names of their beloved, were preserved and eventually became proverbial expressions of the tremendous force of true love. Such was Qays, who went mad because of his passion for Laylā and was afterward known as Majnūn (the "Demented One"). His story is cherished by later Persian, Turkish, and Urdu poets; as a symbol of complete surrender to the force of love, he is dear both to religious mystics and to secular poets.

Notwithstanding such new developments, the traditional *qaṣīdah* form of poetry was by no means neglected during the Umayyad period. Moreover, as the satirists of Iraq rose to fame, the *naqāʾiḍ* ("polemic poetry matches") between Jarīr (died *c.* 729) and al-Farazdaq (died *c.* 728 or 730) excited and delighted tribesmen of the rival settlements of Basra and Kūfah (places that later also became rival centres of philological and theological schools). The work of these two poets has furnished critics and historians with rich material for a study of the political and social situation in the early 9th century. The wealth of al-Farazdaq's vocabulary led one of the old Arabic critics to declare: "If Farazdaq's poetry did not exist, one-third of the Arabic language would be lost." Philologists, eager to preserve as much of the classical linguistic heritage as possible, have also paid a great deal of attention to the largely satirical poetry of al-Ḥuṭayʾah (died 674). The fact that Christians as well as Muslims were involved in composing classical Arabic poetry is proved by the case of al-Akhṭal (died *c.* 710), whose work preserves the pre-Islāmic tradition of al-Ḥīrah in authentic form. He is particularly noted for his wine songs. Christians and Jews had been included among the pre-Islāmic poets.

Prose literature was still restricted to religious writing. The traditions of the Prophet began to be compiled, and, after careful sifting, those regarded as trustworthy were preserved in six great collections during the late 9th century. Two of these—that of al-Bukhārī and that of Muslim ibn al-Ḥajjāj—were considered second only to the Qurʾan in religious importance. The first studies of religious law and legal problems, closely connected with the study of the Qurʾān, also belong to that period.

The ʿAbbāsids. It was not until the ʿAbbāsids assumed power in 750, settling in Baghdad, that the golden age of Arabic literature began. The influx of foreign elements added new colour to cultural and literary life. Hellenistic thought and the influence of the ancient cultures of the Near East, for example, contributed to the rapid intellectual growth of the Muslim community. Its members, seized with insatiable intellectual curiosity, began to adapt elements from all the earlier high cultures and to incorporate them into their own. They thus created the wonderful fabric of Islāmic culture that was so much admired in the Middle Ages by western Europe. Indian and Iranian threads were also woven into this fabric, and a new sensitivity to beauty in the field of poetry and the fine arts was cultivated.

The classical Bedouin style was still predominant in literature and was the major preoccupation of grammarians. These men were, as the modern critic Sir Hamilton Gibb has emphasized, the true humanists of Islām. Their efforts helped to standardize "High Arabic," giving it an unchangeable structure once and for all. By now the inhabitants of the growing towns in Iraq and Syria were beginning to express their love, hatred, religious fervour, and frivolity in a style more appealing to their fellow townsmen. Poets no longer belonged exclusively to what had been the Bedouin aristocracy. Artisans and freed slaves, of non-Arab origin, were included among their number. Bashshār ibn Burd (died *c.* 784), the son of a Persian slave, was the first representative of the new style. This ugly, blind workman excelled as a seductive love poet and also as a biting satirist—"Nobody could be secure from the itch of his tongue," it was later said—and he added a new degree of expressiveness to the old forms. The category of *zuhdīyāt* (didactic-ascetic poems) was invented by the poet Abū al-ʿAtāhīyah (died 825 or 826) from Basra, the centre of early ascetic movements. His pessimistic thoughts on the transitory nature of this world were uttered in an unpretentious kind of verse that rejected all current notions of style and technical finesse. He had turned to ascetic poetry after efforts at composing love songs.

The works of Abū Nuwās

The same is said of Abū Nuwās (died *c.* 813), the most outstanding of the ʿAbbāsid poets. His witty and cynical verses are addressed mainly to handsome boys; best known are his scintillating drinking songs. His line "Accumulate as many sins as you can" seems to have been his motto; and compared with some of his more lascivious lines, even the most daring passages of pre-Islāmic poetry sound chaste. Abū Nuwās had such an incomparable command over the language, however, that he came to be regarded as one of the greatest Arabic poets of all time. Nevertheless, orthodox Muslims would quote of him and of his imitators the Prophet's alleged saying that "poetry is what Satan has spit out," since he not only described subjects prohibited by religious law but praised them with carefree lightheartedness.

Tendency toward mannered writing

The "new" style. The new approach to poetry that developed during the 9th century was first accorded scholarly discussion in the *Kitāb al-badīʿ* ("Book of the Novel and Strange") by Ibn al-Muʿtazz (died 908), caliph for one day, who laid down rules for the use of metaphors, similes, and verbal puns. The ideal of these "modern" poets was the richest possible embellishment of verses by the use of tropes, brilliant figures of speech, and farfetched poetic conceits. Many later handbooks of poetics discussed these rules in minute detail, and eventually the increasing use of rhetorical devices no longer produced art but artificiality. (Ibn al-Muʿtazz was himself a fine poet whose descriptions of courtly life and nature are lovely; he even tried to compose a tiny epic poem, a genre otherwise unknown to the Arabs.) The "modern" poets, sensitive to colours, sounds, and shapes, also were fond of writing short poems on unlikely subjects: a well-bred hunting dog or an inkpot; delicious sweetmeats or jaundice; the ascetic who constantly weeps when he remembers his sins; the luxurious garden parties of the rich; an elegy for a cat; or a description of a green ewer. Their amusing approach, however, was sooner or later bound to lead to mannered compositions. The growing use of colour images may be credited to the increasing Persian influence upon ʿAbbāsid poetry; for the Persian poets were, as has been often observed, on the whole more disposed to visual than to acoustic imagery.

New attitudes toward love, too, were being gradually developed in poetry. Eventually, what was to become a classic theme, that of *ḥubb ʿudhrī* ("ʿUdhrah love")—the lover would rather die than achieve union with his beloved—was expounded by the Ẓāhirī theologian Ibn Dāʾūd (died 910) in his poetic anthology *Kitāb az-zahrah* ("Book of the Flower"). This theme was central to the *ghazal* poetry of the following centuries. Although at first completely secular, it was later taken over as a major concept in mystical love poetry. (The first examples of this adoption, in Iraq and Egypt, took place in Ibn Dāʾūd's lifetime.) The wish to die on the path that leads to the beloved became commonplace in Persian, Turkish, and Urdu poetry; and most romances in these languages end tragically. Ibn Dāʾūd's influence also spread to the western Islāmic world. A century after his death, the theologian Ibn Ḥazm (died 1064), drawing upon personal experiences, composed in Spain his famous work on "pure love" called *Ṭawq al-ḥamāmah* (*The Ring of the Dove*). Its lucid prose, interspersed with poetry, has many times been translated into Western languages.

The conflict between the traditional ideals of poetry and the "modern" school of the early ʿAbbāsid period also led to the growth of a literary criticism, the criteria of which were largely derived from the study of Greek philosophy.

Traditional poetry, meanwhile, was not neglected. But its style was somewhat modified in accordance with the new ideas. Two famous anthologies of Bedouin poetry, both called *Ḥamāsah* ("Poems of Bravery"), were collected by the Syrian Abū Tammām (died 845 or 846) and his disciple al-Buḥturī (died 897), both good classical poets in their own right. They provide an excellent survey of those poems from the stock of early Arabic poetry that were considered worth preserving. A century later Abū al-Faraj al-Iṣbahānī (died 967), in a multivolume work entitled *Kitāb al-aghānī* ("Book of Songs"), collected a great number of poems and biographical notes on poets and musicians. This material gives a colourful and valuable panorama of literary life in the first four centuries of Islām.

Mutanabbī's poetry

In the mid-10th century a new cultural centre emerged at the small court of the Ḥamdānids in Aleppo. Here the Central Asian scholar al-Fārābī (died 950) wrote his fundamental works on philosophy and musical theory. Here, too, for a while, lived Abū aṭ-Ṭayyib al-Mutanabbī (died 965), who is in the mainstream of classical *qaṣīdah* writers but who surpasses them all in the extravagance of what has been called his "reckless audacity of imagination." He combined some elements of Iraqi and Syrian stylistics with classical ingredients. His compositions—panegyrics of rulers and succinct verses (which are still quoted)—have never ceased to intoxicate the Arabs by their daring hyperbole, their marvelous sound effects, and their formal perfection. The Western reader is unlikely to derive as much aesthetic pleasure from Mutanabbī's poetry as does one whose mother tongue is Arabic. He will probably prefer the delicate verses about gardens and flowers by Mutanabbī's colleague in Aleppo, aṣ-Ṣanawbarī (died 945), a classic exponent of the descriptive style. This style in time reached Spain, where the superb garden and landscape poetry of Ibn Khafājah (died 1139) displayed an even higher degree of elegance and sensitivity than that of his Eastern predecessors.

Before turning to the development of prose, it is necessary to mention a figure unique among those writing in Arabic. This was Abū al-ʿAlāʾ al-Maʿarrī (died 1057), a blind poet of Syria, whose verses have appealed greatly to young Arabs of the present because of the poems' sincerity and humanity. But al-Maʿarrī's vocabulary is so difficult, his verses, with their double rhymes, are so compressed in meaning, that even his contemporaries, flocking to his lectures, had to ask him to interpret their significance. His outlook is deeply pessimistic and skeptical. Although his poems display a mastery of the Arabic traditional stylistic devices, they run counter to the conventional ideals of Arab heroism by speaking of bitter disappointment and emphasizing asceticism, compassion, and avoidance of procreation.

> Taking reason for his guide he judges men and things with a freedom which must have seemed scandalous to the rulers and privileged classes of the day. Among his meditations on the human tragedy a fierce hatred of injustice, hypocrisy, and superstition blazes out. Vice and folly are laid bare in order that virtue and wisdom may be sought . . .

says Reynold A. Nicholson, al-Maʿarrī's foremost interpreter in the West, who has also translated his *Risālat al-ghufrān* ("Epistle of Pardon"), which describes a visit to the Otherworld. Maʿarrī's extremely erudite book also contains sarcastic criticism of Arabic literature. His *Al-Fuṣūl wa al-ghāyāt* ("Paragraphs and Periods") is an ironic commentary on man and nature but is presented as a sequence of pious exhortations in rhymed prose. It has scandalized the pious, some of whom see it as a parody of the Qurʾān. Maʿarrī's true intention in this book, which came to light only recently, is unknown.

Development of literary prose. During the ʿAbbāsid period, literary prose also began to develop. Ibn al-Muqaffaʿ (died *c.* 756), of Persian origin, translated the fables of Bidpai into Arabic under the title *Kalīlah wa Dimnah.* These fables provided Islāmic culture with a seemingly inexhaustible treasure of tales and parables, which are to be found in different guises throughout the whole of Muslim literature. He also introduced into Arabic the fictitious chronicles of the Persian *Khwatāy-nāmak* ("Book of Kings"). This was the source of a kind of pre-Islāmic mythology that the literati preferred above the somewhat meagre historical accounts of the Arab pagan past otherwise available to them. These activities demanded a smooth prose style, and Ibn al-Muqaffaʿ has therefore rightly been regarded as the inaugurator of what is called "secretarial literature" (that produced by secretaries in the official chancelleries). He also translated writings on ethics and the conduct of government, which helped to determine the rules of etiquette (*adab*). His works are the prototype of the "Mirror for Princes" literature, which flourished during the late Middle Ages both in Iran and in the West. In this literature, a legendary Persian counselor, Bozorgmehr, was presented as a paragon of wise conduct. Later, stories were invented that combined Qurʾānic heroes with historical characters from the Iranian past.

The work of al-Jāḥiẓ

A growing interest in things outside the limits of Bedouin life was reflected in a quantity of didactic yet entertaining prose by such masters as the broadminded and immensely learned al-Jāḥiẓ (died 869). In response to the wide-ranging curiosity of urban society, the list of his subjects includes treatises on theology, on misers, on donkeys, and on thieves. His masterpiece is *Kitāb al-Ḥayawān* ("Book of Animals"), which has little to do with zoology but is a mine of information about Arab proverbs, traditions, superstitions, and the like. Al-Jāḥiẓ's style is vigorous, loquacious, and uninhibited. His work, however, is not well constructed, and it lacks the clear sobriety of the "secretarial style." Yet the glimpses it affords into the life of various strata of society during the 9th century have rightly attracted the special interest of Western scholars. Less impressive, but almost as multifaceted, are the treatises of Ibn Abī ad-Dunyā (died 894).

The concept of *adab* was soon enlarged to include not only educational prose dealing with etiquette for all classes of people but belles lettres in general. The classic example of Arabic style for prose writers in this field, accepted as such for almost a millennium, is the writing of the Persian Ibn Qutaybah (died 889). His *ʿUyūn al-akhbār* ("Fountains of Stories"), in 10 books, each dealing with a given subject, provided a model to which numberless essayists in the Muslim world conformed. In his book on poetry and poets, Ibn Qutaybah dared, for the first time, to doubt openly that pre-Islāmic poetry was incomparable. The most vigorous prose style was achieved by Abū Ḥayyān at-Tawḥīdī (died 1023), who portrayed the weaknesses of the two leading viziers, both notorious for their literary ambitions, ". . . with such bitterness," as Gibb remarks, "that the book was reputed to bring misfortune upon all who possessed a copy." This work, like others by Tawḥīdī that have quite recently been discovered, reveals the author's sagacity and striking eloquence. His correspondence on problems of philosophy with Miskawayh (died 1030), the author of a widely circulated book on ethics and of a general history, helps to complete the picture of this extraordinary writer.

Some time about 800 the Arabs had learned the art of papermaking from the Chinese. Henceforth, cheap writing material was available, and literary output was prodigious. The *Fihrist* ("Index"), compiled by the bookseller Ibn an-Nadīm in 988, gave a full account of the Arabic literature extant in the 10th century. This Index covered all kinds of literature, from philology to alchemy; but most of these works unfortunately have been lost. In those years manuals of composition (*inshāʾ*) were written elaborating the technique of secretarial correspondence, and they grew into an accepted genre in Arabic as well as in Persian and Turkish literature. The devices thought indispensable for elegance in modern poetry were applied to prose. The products were mannered, full of puns, verbal tricks, riddles, and the like. The new style, which was also to affect the historian's art in later times, makes a good deal of this post-classical Arabic prose look very different from the terse and direct expression characteristic of the early specimens. Rhymed prose, which at one time had been reserved for such religious occasions as the Friday sermons, was now regarded as an essential part of elegant style.

Rhetorical artistry in the *maqāmah*

This rhetorical artistry found its most superb expression in the *maqāmah,* a form invented by Badīʿ az-Zamān al-Hamadhānī (died 1008). Its master, however, was al-Ḥarīrī (died 1122), postmaster (head of the intelligence service) at Basra and an accomplished writer on grammatical subjects. His 50 *maqāmah*s, which tell the adventures of Abū Zayd as-Sarūjī, with a wealth of language and learning, come closer to the Western concept of short story than anything else in classical Arabic literature. They abound in verbal conceits, ambivalence, assonance, alliteration, palindromes; they change abruptly from earnest to jest, from the crude to the most sublime, as the modern scholar G.E. von Grunebaum has pointed out in his evaluation of this form, which he regards as the most typical literary reflection of the Islāmic spirit. The work of al-Ḥarīrī has certainly been widely admired in the East; it has been imitated in Syriac and in Hebrew and has formed part of the syllabus in Muslim high schools of India. The pleasure to be derived from the brilliant artifice and ingenuity behind such compositions has led to their being imitated in other literary fields: quite often, in later Persian literature, one finds poems—sometimes whole books—composed of letters without diacritical marks (which distinguish otherwise similar-looking letters) or even made up entirely of unconnected letters. Even a commentary on the Qurʾān, in undotted letters, has been written in India (by Fayẓī, died 1595).

Writers on music and philology

Achievements in the western Muslim world. The Arabic literature of Moorish Spain and of the whole Maghrib developed parallel with that of the eastern countries but came to full flower somewhat later. Córdoba, the seat of the Umayyad rulers, was the centre of cultural life. Its wonderful mosque has inspired Muslim poets right up to the 20th century (such as Sir Muḥammad Iqbāl, whose Urdu ode, "The Mosque of Córdoba," was written in 1935). Moorish Spain was a favourite topic for reformist novelists of 19th-century Muslim India, who contrasted their own country's troubled state with the glory of classical Islāmic civilization. Moorish Spain reached its cultural, political, and literary heyday under ʿAbd ar-Raḥmān III (912–961). Literary stylistic changes, as noted in Iraq and Syria, spread to the west: there the old Bedouin style had always been rare and soon gave way to descriptive and love poetry. Ibn Hāniʾ (died 973) of Seville has been praised as the Western counterpart of al-Mutanabbī, largely because of his eulogies of the Fāṭimid caliph al-Muʿizz, who at that time still resided in North Africa. The entertaining prose style of Ibn ʿAbd Rabbihi (died 940) in his *al-ʿIqd al-farīd* ("The Unique Necklace") is similar to that of his elder contemporary Ibn Qutaybah, and his book in fact became more famous than that of his predecessor. Writers on music and philology also flourished in Spain; literary criticism was practiced by Ibn Rashīq (died 1064) and, later, by al-Qarṭājannī (died 1285) in Tunis. Ibn Ḥazm (died 1064), theologian and accomplished writer on pure love, has already been mentioned.

Philosophy: Averroës and Avicenna. Philosophy, medicine, and theology, all of which flourished in the ʿAbbāsid East, were also of importance in the Maghrib; and from there strong influences reached medieval Europe. The influences often came through the mediation of the Jews, who, along with numerous Christians, were largely Arabized in their cultural and literary outlook. The eastern Muslim countries could boast of the first systematic writers in the field of philosophy, including al-Kindī (died *c.* 870), al-Fārābī (died 950), and especially Avicenna (Ibn Sīnā, died 1037). Avicenna's work in philosophy, science, and medicine was outstanding and was appreciated as such in Europe. He also composed religious treatises and tales with a mystical slant. One of his romances was reworked by the Maghribi philosopher Ibn Ṭufayl (died 1185) in his book *Ḥayy ibn Yaqẓān* ("Alive Son of Awake"), or *Philosophus Autodidactus* (the title of its first Latin translation, made in 1671). It is the story of a self-taught man who lived on a lonely island and who, in his maturity, attained the full knowledge taught by philosophers and prophets. This theme was elaborated often in later European literature.

The dominating figure in the kingdom of the Almohads, however, was the philosopher Averroës (Ibn Rushd, died 1198), court physician of the Berber kings in Marrākush (Marrakech) and famous as the great Arab commentator on Aristotle. The importance of his frequently misinterpreted philosophy in the formation of medieval Christian thought is well known. Among his many other writings, especially notable is his merciless reply to an attack on philosophy made by Ghazālī (died 1111). Ghazālī had called his attack *Tahāfut al-falāsifah* (*The Incoherence of the Philosophers*), while Averroës' equally famous reply was entitled *Tahāfut at-tahāfut* (*The Incoherence of the Incoherence*). The Persian-born Ghazālī had, after giving up a splendid scholarly career, become the most influential representative of moderate Ṣūfism. His chief work, *Iḥyāʾ ʿulūm ad-dīn* ("The Revival of the Religious Sciences"), was based on personal religious experiences and is a perfect introduction to the pious Muslim's way to God. It inspired much later religious poetry and prose. The numerous writings by mystics, who often expressed their wisdom in rather cryptic language (thereby contributing to the profundity of Arabic vocabulary), and the handbooks of religious teaching produced in eastern Arab and Persian areas (Sarrāj, Kalābādhī, Qushayrī, and, in Muslim India, al-Hujwīrī) are generally superior to those produced in western Muslim countries. Yet the greatest Islāmic theosophist of all, Ibn al-ʿArabī (died 1240), was Spanish in origin and was educated in the Spanish tradition. His writings, in both poetry and prose, shaped large parts of Islāmic thought during the following centuries. Much of the later literature of eastern Islām, particularly Persian and Indo-Persian mystical writings, indeed, can be understood only in the light of his teachings. Ibn al-ʿArabī's lyrics are typical *ghazal*s, sweet and flowing. From the late 9th century, Arabic-speaking mystics had been composing verses often meant to be sung in their meetings. At first a purely religious vocabulary was employed, but soon the expressions began to oscillate between worldly and heavenly love. The ambiguity thus achieved eventually became a characteristic feature of Persian and Turkish lyrics.

The pious Muslim's way to God

Among the Arabs, religious poetry mainly followed the classical *qaṣīdah* models, and the poets lavishly decorated their panegyrics to the Prophet Muḥammad with every conceivable rhetorical embellishment. Examples of this trend include *al-Burdah* ("The Mantle") of al-Buṣīrī (died 1298), upon which dozens of commentaries have been written (and which has been translated into most of the languages of Muslims because of the power to bless attributed to it). More sophisticated but less well known is an ode on the Prophet by the Iraqi poet Ṣafī ad-Dīn al-Ḥilli (died 1350), which contains 151 rhetorical figures. The "letters of spiritual guidance" developed by the mystics are worth mentioning as a literary genre. They have been popular everywhere; from the western Islāmic world the letters of Ibn ʿAbbād (died 1390) of Ronda (in Spain) are outstanding examples of this category, being written clearly and lucidly.

Geographical literature. The Maghrib also made a substantial contribution to geographical literature, a field eagerly cultivated by Arab scholars since the 9th century. The Sicilian geographer ash-Sharīf al-Idrīsī produced a famous map of the world and accompanied it with a detailed description in his *Kitāb nuzhat al-mushtāq fī ikhtirāq al-āfāq* ("The Delight of Him Who Wishes to Traverse the Regions of the World," 1154), which he dedicated to his patron, Roger II. The Spanish traveler Ibn Jubayr (died 1217), while on pilgrimage to Mecca, kept notes of his experiences and adventures. The resulting book became a model for the later pilgrims' manuals that are found everywhere in the Muslim world. The Maghribi explorer Ibn Baṭṭūṭah (died 1368/69 or 1377) described his extensive travels to the Far East, India, and the region of the Niger in a book filled with information about the cultural state of the Muslim world at that time. The value of his narrative is enhanced by the simple and pleasing style in which it is written.

Spanish mastery of poetical forms

Poetry. In the field of poetry, Spain, which produced a considerable number of masters in the established poetical forms, also began to popularize strophic poetry, possibly deriving from indigenous models. The *muwashshaḥ* ("girdled") poem, written in the classical short metres and arranged in four- to six-line stanzas, was elaborated, enriched by internal rhymes, and, embodying some popular expressions in the poem's final section, soon achieved a standardized form. The theme is almost always love. Among the greatest lyric poets of Spain was Ibn Zaydūn of Córdoba (died 1071), who was of noble birth. After composing some charming love songs dedicated to the Umayyad princess Wallādah, he turned his hand to poetic epistles. He is the author of a beautiful *muwashshaḥ* about his hometown, which many later poets imitated. When the *muwashshaḥ* was transplanted to the eastern Arabic countries, however, it lost its original spontaneity and became as stereotyped as every other lyric form of expression during the later Middle Ages. Another strophic form developed in Spain is the songlike *zajal* (melody), interesting for its embodiment of dialect phrases and the use of occasional words from Romance languages. Its master was Ibn Quzmān of Córdoba (died 1160), whose life-style was similar to that of Western troubadours. His approach to life as expressed in these melodious poems, together with their mixed idiom, suggests an interrelationship with the vernacular troubadour poetry of Spain and France.

Historiography: Ibn Khaldūn. Any survey of western Muslim literary achievements would be incomplete if it did not mention the most profound historiographer of the Islāmic world, the Tunisian Ibn Khaldūn (died 1406). History has been called the characteristic science of the Muslims because of the Qurʾānic admonition to discover signs of the divine in the fate of past peoples. Islāmic historiography has produced histories of the Muslim conquests, world histories, histories of dynasties, court annals, and biographical works classified by occupation—scholars, poets, and theologians. Yet, notwithstanding their learning, none of the earlier writers had attempted to produce a comprehensive view of history. Ibn Khaldūn, in the famous *Muqaddimah* or introduction to a projected general history, *Kitāb al-ʿibar,* sought to explain the basic factors in the historical development of the Islāmic countries. His own experiences, gained on a variety of political missions in North Africa, proved useful in establishing general principles that he could apply to the manifestations of Islāmic civilization. He created, in fact, the first "sociological" study of history, free from bias. Yet his book was little appreciated by his fellow historians, who still clung to the method of accumulating facts without shaping them properly into a well-structured whole. Ibn Khaldūn's work eventually attracted the interest of Western Orientalists, historians, and sociologists alike; and some of his analyses are still held in great esteem.

Significance of Ibn Khaldūn

Decline of the Arabic language. Ibn Khaldūn, who had served in his youth as ambassador to Pedro I the Cruel, of Castile, and in his old age as emissary to Timur, died in Cairo. After the fall of Baghdad in 1258, this city had become the centre of Muslim learning. Historians there recorded every detail of the daily life and the policies of the Mamlūk sultans; theologians and philologists worked under the patronage of Turkish and Circassian rulers who often did not speak a word of Arabic. The amusing, semicolloquial style of the historian Ibn Iyās (died after 1521) is an interesting example of the deterioration of the Arabic language. While classical Arabic was still the ideal of every literate man, it had become exclusively a "learned" language. Even some copyists who transcribed classical works showed a deplorable lack of grammatical knowledge. It is hardly surprising that poetry composed under such circumstances should be restricted to insipid versification and the repetition of well-worn clichés.

MIDDLE PERIOD: THE RISE OF PERSIAN AND TURKISH POETRY

The new Persian style. During the ʿAbbāsid period, the Persian influence upon the Arabic had grown considerably: at the same time, a distinct Modern Persian literature came into existence in northeastern Iran, where the house of the Sāmānids of Bukhara and Samarkand had revived the memory of Sāsānian glories.

The first famous representative of this new literature was the poet Rūdakī (died 940/941), of whose *qaṣīdah*s only a few have survived. He also worked on a Persian version of *Kalīlah wa Dimnah,* however, and on a version of the *Sendbād-nāmeh.* Rūdakī's poetry, modeled on the Arabic rules of prosody that without exception had been applied to Persian, already points ahead to many of the characteristic features of later Persian poetry. The imagery in particular is sophisticated, although when compared with the mannered writing of subsequent times his verse was considered sadly simple. From the 10th century onward, Persian poems were written at almost every court in the Iranian areas, sometimes in dialectical variants (for example, in Ṭabarestāni dialect at the Zeyārid court). In many cases the poets were bilingual, excelling in both Arabic and Persian (a gift shared by many non-Arab writers up to the 19th century).

Influence of Maḥmūd of Ghazna. The first important centre of Persian literature existed at Ghazna (present-day Ghaznī, Afg.), at the court of Maḥmūd of Ghazna (died 1030) and his successors, who eventually extended their empire to northwestern India. Himself an orthodox warrior, Maḥmūd in later love poetry was transformed into a symbol of "a slave of his slave" because of his love for a Turkmen officer, Ayāz. Under the Ghaznavids, lyric and epic poetry both developed, as did the panegyric. Classical Iranian topics became the themes of poetry, resulting in such diverse works as the love story of Vāmeq and ʿAzrā (possibly of Greek origin) and the *Shāh-nāmeh* ("Book of Kings"). A number of gifted poets praised Maḥmūd, his successors, and his ministers. Among them was Farrokhī of Seistan (died 1037), who was the author of a powerful elegy on Maḥmūd's death, one of the finest compositions of Persian court poetry.

Ferdowsī's literary achievement

Epic and romance. The main literary achievement of the Ghaznavid period, however, was that of Ferdowsī (died 1020). He compiled the inherited tales and legends about the Persian kings in one grand epic, the *Shāh-nāmeh,* which contains between 35,000 and 60,000 verses in short rhyming couplets. It deals with the history of Iran from its beginnings—that is, from the "time" of the mythical kings—passing on to historical events, giving information about the acceptance of the Zoroastrian faith, Alexander's invasion, and, eventually, the conquest of the country by the Arabs. A large part of the work centres on tales of the hero Rostam. These stories are essentially part of a different culture, thus revealing something about the Indo-European sources of Iranian mythology. The struggle between Iran and Tūrān (the central Asian steppes from which new waves of nomadic conquerors distributed Iran's urban culture) forms the central theme of the book; and the importance of the legitimate succession of kings, who are endowed with royal charisma, is reflected throughout the composition. The poem contains very few Arabic words and is often considered the masterpiece of Persian national literature, although it lacks proper historical perspective. Its episodes have been the inspiration of miniaturists since the 14th century. Numerous attempts have been made to emulate it in Iran, India, and Turkey.

The epics of Neẓāmī

Other epic poems, on a variety of subjects, were composed during the 11th century. The first example is Asadī's (died *c.* 1072) didactic *Garshāsb-nāmeh* ("Book of Garshāsb"), whose hero is very similar to Rostam. The tales of Alexander and his journeys through foreign lands were another favourite topic. Poetical romances were also being written at this time; they include the tale of *Varqeh o-Golshāh* by ʿEyyūqī (11th century) and *Vīs o-Rāmīn* by Fakhr od-Dīn Gorgānī (died after 1055), which has parallels with the Tristan story of medieval romance. These were soon superseded, however, by the great romantic epics of Neẓāmī of Ganja (died *c.* 1209), in Caucasia. The latter are known as the *Khamseh* ("Quintet") and, though the names of Vīs or Vāmeq continued for some time to serve as symbols of the longing lover, it was the poetical work of Neẓāmī that supplied subsequent writers with a rich store of images, similes, and stories to draw upon. The first work of his *Khamseh, Makhzan ol-asrār* ("Treasury of Mysteries"), is didactic in intention; the subjects of the following three poems are traditional love stories. The first is the Arabic romance of Majnūn, who went mad with love for Laylā. Second is the Persian historical tale of Shīrīn, a Christian princess, loved by both the Sāsānian ruler Khosrow II Parvīz and the stonecutter Farhād. The third story, *Haft peykar* ("Seven Beauties"), deals with the adventures of Bahrām Gūr, a Sāsānian prince, and seven princesses, each connected with one day of the week, one particular star, one colour, one perfume, and so on. The last part of the *Khamseh* is *Eskandar-nāmeh,* which relates the adventures of Alexander III the Great in Africa and Asia, as well as his discussions with the wise philosophers. It thus follows the traditions about Alexander and his tutor, Aristotle, emphasizing the importance of a counselor-philosopher in the service of a mighty emperor. Neẕāmī's ability to present a picture of life through highly refined language and a wholly apt choice of images is quite extraordinary. Human feelings, as he describes them, are fully believable; and his characters are drawn with a keen insight into human nature. Not surprisingly, Neẕāmī's work inspired countless poets' imitations in different languages—including Turkish, Kurdish, and Urdu—while painters constantly illustrated his stories for centuries afterward.

Other poetic forms. In addition to epic poetry, the lesser forms, such as the *qaṣīdah* and *ghazal,* developed during the 11th and 12th centuries. Many poets wrote at the courts of the Seljuqs and also at the Ghaznavid court in Lahore, where the poet Mas'ūd-e Sa'd-e Salmān (died 1121) composed a number of heartfelt *qaṣīdahs* during his political imprisonment. They are outstanding examples of the category of *ḥabsīyah* (prison poem), which usually reveals more of the author's personal feelings than other literary forms. Other famous examples of *ḥabsīyahs* include those written by the Arab knight Abū Firās (died 968) in a Byzantine prison; those by Muḥammad II al-Mu'tamid of Seville (died 1095) in the dungeons of the Almohads; those by the 12th-century Persian Khāqānī; those by the Urdu poets Ghālib, in the 19th, and Faiz, in the 20th century; and by the contemporary Turkish poet Nazim Hikmet (died 1963).

The most complicated forms were mastered by poets of the very early period, the limits of artificiality being reached in Azerbaijani *qaṣīdahs* by the poet Qaṭrān (died 1072), whose work displays virtuosity for virtuosity's sake. The court poets tried to top one another in the accumulation of complex metaphors and paradoxes, each hoping to win the coveted title "Prince of Poets." Anvarī (died *c.* 1189), whose patrons were the Seljuqs, is considered the most accomplished writer of panegyrics in the Persian tongue. His verses contain little descriptive material but abound in learned allusions. His "Tears of Khorāsān," mourning the passing of Seljuq glory, is among the best known of Persian *qaṣīdahs*. In the west of Iran, Anvarī's contemporary Khāqānī (died *c.* 1190), who wrote mainly at the court of the Shīrvān-Shāhs of Transcaucasia, is the outstanding master of the hyperbolic style. His mother was a Christian, and his imagery has more than the usual amount of allusions to Christian themes. His vocabulary seems inexhaustible; he uses uncommon rhetorical devices and very strong language. His poems, with their long chains of oath-formulae (*sowgandnāmeh*), are as impressive as his poignant antithetic formulations. Khāqānī's verses on the ruined Ṭāq Kisrā at Ctesiphon on the Tigris have become proverbial. His *qaṣīdahs* on the pilgrimage to Mecca, which also inspired his *masnavī, Tuḥfat al-'Irāqayn ol-'Erāqeyn* ("Gift of the Two Iraqs"), translate most eloquently the feelings of a Muslim at the festive occasion. In the hand of lesser poets, however, *qaṣīdah* writing became more and more conventionalized, repeating outworn clichés and employing inflated terms entirely devoid of feeling.

Decline of the *qaṣīdah* form

Scholarship: al-Bīrūnī. The Ghaznavid and Seljuq periods produced first-rate scholars such as al-Bīrūnī (died 1048) who, writing in Arabic, investigated Hinduism and gave the first unprejudiced account of India—indeed, of any non-Islāmic culture. He also wrote notable books on chronology and history. In his search for pure knowledge he is undoubtedly one of the greatest minds in Islāmic history. Interest in philosophy is represented by Nāṣer-e Khosrow (died 1087/88) who acted for a time as a missionary for the Ismā'īlī branch of Shī'ah Islām. His book about his journey to Egypt, entitled *Safar-nāmeh,* is a pleasing example of simple, clearly expressed, early Persian prose. His poetical works in the main seek to combine Greek wisdom and Islāmic thought: the gnostic Ismā'īlī interpretation of Islām seemed, to him, an ideal vehicle for a renaissance of the basic Islāmic truths.

Robā'īyāt: Omar Khayyam. The work done in mathematics by early Arabic scholars and by al-Bīrūnī was continued by Omar Khayyam (died 1122), to whom the Seljuq empire in fact owes the reform of its calendar. But Omar has become famous in the West through the free adaptations by Edward FitzGerald of his *robā'īyāt.* These quatrains have been translated into almost every known language and are largely responsible for colouring European ideas about Persian poetry. The authenticity of these verses has often been questioned. The quatrain is an easy form to use—many have been scribbled on Persian pottery of the 13th century—and the same verse has been attributed to many different authors. The latest research into the question of the *robā'īyāt* has established that a certain number of the quatrains can, indeed, be traced back to the great scientist who condensed in them his feelings and thoughts, his skepticism and love, in such an enthralling way that they appeal to every reader. The imagery he uses, however, is entirely inherited; none of it is original. (One of the most noted, and notorious, writers of this genre was the poetess Mahsaṭī [first half of the 12th century], who frequently addressed members of different professions in rather frivolous lines.) The quatrain was also popular as a means of embodying pieces of mystical wisdom. One has to do away with the old theory that the first author of such mystical *robā'īyāt* was Abū Sa'īd ibn Abū al-Khayr (died 1049). A number of his contemporaries, however, including Bābā Ṭāher 'Oryān (died after 1055), used simpler forms of the quatrain, sometimes in order to express their mystical concepts.

The mystical poem. Whereas the mystical thought stemming from Iran had formerly been written in Arabic, writers from the 11th century onward turned to Persian. Along with works of pious edification and theoretical discussions, what was to be one of the most common types of Persian literature came into existence: the mystical poem. Khwajah 'Abd Allāh al-Anṣārī of Herāt (died 1088), a prolific writer on religious topics in both Arabic and Persian, first popularized the literary "prayer," or mystical contemplation, written in Persian in rhyming prose interspersed with verses. Sanā'ī (died 1131?), at one time a court poet of the Ghaznavids, composed the first mystical epic, the didactic *Ḥadīqat al-ḥaqīqat wa sharī'at aṭ-ṭariqah* ("The Garden of Truth and the Law of the Path"), which has some 10,000 verses. In this lengthy and rather dry poem, the pattern for all later mystical *masnavīs* is established: wisdom is embodied in stories and anecdotes; parables and proverbs are woven into the texture of the story, eventually leading back to the main subject, although the argument is without thread and the narration puzzling to follow. Among Sanā'ī's smaller *masnavīs, Sayr al-'ibād ilā al-ma'ād* ("The Journey of the Servants to the Place of Return") deserves special mention. Its theme is the journey of the spirit through the spheres, a subject dear to the mystics and still employed in modern times as, for example, by Iqbāl in his Persian *Jāvīd-nāmeh* (1932). Sanā'ī's epic endeavours were continued by one of the most prolific writers in the Persian tongue, Farīd od-Dīn 'Aṭṭār (died *c.* 1220). He was a born storyteller, a fact that emerges from his lyrics but even more so from his works of edification. The most famous among his *masnavīs* is the *Manṭiq uṭ-ṭayr* (*The Conversation of the Birds*), modeled after some Arabic allegories. It is the story of 30 birds, who, in search of their spiritual king, journey through seven valleys. The poem is full of tales, some of which have been translated even into the most remote Islāmic languages. (The story of the pious Sheykh Ṣan'ān, who fell in love with a Christian maiden, is found, for example, in Kashmiri.) 'Aṭṭār's symbolism of the soul-bird was perfectly in accord with the existing body of imagery beloved of Persian poetry, but it was he who added a scene in which the birds eventually realize their own identity with God (because they, being *sī morgh,* or "30 birds," are identified with the mystical Sēmorgh, who represents God). Also notable are his *Elāhī-nāmeh,* an allegory of a king and his six sons, and his profound *Moṣībat-nāmeh* ("Book of Affliction"), which closes with its hero's being immersed in the ocean of his soul after wandering through the 40 stages of his search for God. The epic exteriorizes the mystic's experiences in the 40 days of seclusion.

Importance of Mawlānā Jalāl ad-Dīn ar-Rūmī. The most famous of the Persian mystical *masnavīs* is by Mawlānā ("Our Lord") Jalāl ad-Dīn ar-Rūmī (died 1273) and

Most famous Persian mystical *mas̄navī*

is known simply as the *Mas̄navī*. It comprises some 26,000 verses and is a complete—though quite disorganized—encyclopaedia of all the mystical thought, theories, and images known in the 13th century. It is regarded by most of the Persian-reading orders of Ṣūfis as second in importance only to the Qurʾān. Its translation into many Islāmic languages and the countless commentaries written on it up to the present day indicate its importance in the formation of Islāmic poetry and religious thought. Jalāl ad-Dīn, who hailed from Balkh and settled in Konya, the capital of the Rūm, or Anatolian Seljuqs (and hence was surnamed "Rūmī"), was also the author of love lyrics whose beauty surpasses even that of the tales in the *Mas̄navī*. Mystical love poetry had been written since the days of Sanāʾī, and theories of love had been explained in the most subtle prose and sensitive verses by the Ṣūfis of the early 12th century. Yet Rūmī's experience of mystical love for the wandering mystic, Shams ad-Dīn of Tabriz, was so ardent and enraptured him to such an extent that he identified himself completely with Shams, going so far as to use the beloved's name as his own pen name. His dithyrambic lyrics, numbering more than 30,000 verses altogether, are not at all abstract or romantic. On the contrary, their vocabulary and imagery are taken directly from everyday life, so that they are vivid, fresh, and convincing. Often their rhythm invites the reader to partake in the mystical dance practiced by Rūmī's followers, the Mawlawīyah. His verses sometimes approach the form of popular folk poetry; indeed, Rūmī is reputed to have written mostly under inspiration; and despite his remarkable poetical technique, the sincerity of his love and longing is never overshadowed, nor is his personality veiled. In these respects he is unique in Persian literature.

Zenith of Islāmic literature. During the 13th century, the Islāmic lands were exposed, on the political plane, to the onslaught of the Mongols and the abolition of the ʿAbbāsid caliphate, while vast areas were laid to waste. Yet this was in fact the period in which Islāmic literatures reached their zenith. Apart from Rūmī's superb poetry, written in the comparative safety of Konya, there was also the work of the Egyptian Ibn al-Fāriḍ (died 1235), who composed some magnificent, delicately written mystical poems in *qaṣīdah* style, and that of Ibn al-ʿArabī, who composed love lyrics and numerous theosophical works that were to become standard. In Iran, one of the greatest literati, Moṣleḥ od-Dīn Saʿdī (died 1292), returned in about 1256 to his birthplace, Shīrāz, after years of journeying; his *Būstān* ("The Orchard") and *Golestān* ("Rose Garden") have been popular ever since. The *Būstān* is a didactic poem telling wise and uplifting moral tales, written in polished, easy-flowing style and a simple metre; the *Golestān,* completed one year later, in 1258, has been judged ". . . the finest flower that could blossom in a Sultan's garden" (Herder). Its eight chapters deal with different aspects of human life and behaviour. At first sight, its prose and poetical fragments appear to be simple and unassuming; but not a word could be changed without destroying the perfect harmony of the sound, imagery, and content. Saʿdī's *Golestān* is thus essential in discovering the nature of the finest Persian literary style. Since the mid-17th century, its moralizing stories have been translated into many Western languages. Saʿdī was likewise the author of some spirited *ghazals*; he may have been the first writer in Iran to compose the sort of love poetry that is now thought of as characteristic of the *ghazal.* A few of his *qaṣīdahs* are also of note, although he is at his best in shorter forms. His elegant aphoristic poems, words of wisdom, and sensible advice all display what has been called the philosophy of common sense—how to act in any given situation so as to make the best of it both for oneself and others, basing one's conduct on the virtues of gentleness, elegance, modesty, and polite behaviour.

Saʿdī's "philosophy of common sense"

The influence of mysticism, on the one hand, and of the elaborate Persian poetical tradition, on the other, is apparent during the later decades of the 13th century, both in Anatolia and in Muslim India. The Persian mystic, Fakhr-ud-Dīn ʿIrāqī (died 1289), a master of delightful love lyrics, lived for almost 25 years in Multān (in present-day Pakistan), where his lively *ghazals* are still sung. His short treatises, in a mixture of poetry and prose (and written under Ibn al-ʿArabī's influence), have been imitated often. While in Multān he may have met the young Amīr Khosrow of Delhi (died 1325), who was one of the most versatile authors to write in Persian, not only in India but in the entire realm of Persian culture. Amīr Khosrow, son of a Turkish officer, but whose mother was Indian, is often styled, because of the sweetness of his speech, "the parrot of India." (In Persian, it should be noted, parrots are always "sugar-talking"; they are, moreover, connected with Paradise and are thought of as wise birds—thus models of the sweet-voiced sage.) He wrote panegyrics of seven successive kings of Delhi and was also a pioneer of Indian Muslim music. Imitating Neẓāmī's *Khamseh,* Khosrow introduced a novelistic strain into the *mas̄navī* by recounting certain events of his own time in poetical form, some parts of which are lyrics. His style of lyrical poetry has been described as "powdered"; and his *ghazals* contain many of the elements that in the 16th and 17th centuries were to become characteristic of the "Indian" style. Khosrow's poetry surprises the reader in its use of unexpected forms and unusual images, complicated constructions and verbal plays, all handled fluently and presented in technically perfect language. His books on the art of letter writing prove his mastery of high-flown Persian prose. Khosrow's younger contemporary, Ḥasan of Delhi (died 1328), is less well known and had a more simple style. He nevertheless surpassed Khosrow in warmth and charm, qualities that have earned him the title of "the Saʿdī of Hindustan."

Turkish literature. As for the literary developments in Turkey around 1300, the mystical singer Yunus Emre is the first and most important in a long line of popular poets. Little is known about his life, which he probably spent not far from the Sakarya River of Asia Minor. Before him, in Central Asia, the religious leader Ahmed Yesevi (died 1166) had written some rather dry verses on wisdom in Turkish. Yunus, in Anatolia, however, was the first known poet to have caught something of Rūmī's fervour and translated it into a provincial setting, creating ". . . a Turkish vernacular poetry that was to be the model for all subsequent literary productions of popular religion." Sometimes he used the inherited Arabo-Persian prosody, but his best poems are those written in four-line verses using syllable-counting metres. Yunus drew heavily on the reservoir of imagery that had been collected by the great Persian writing mystics, notably Rūmī; but his classical technique did not hinder the expression of his own unself-conscious simplicity, which led him to introduce new images taken from everyday life in Anatolian villages. His *ilahis* (hymns), probably written to be sung at the meetings of the Ṣūfis in the centres of their orders, are still loved by the Turks and memorized by their children.

Use of images from everyday life

Influence of Yunus Emre. The Turkish people rightly claim Yunus as the founder of Turkish literature proper. His poetry is considered the chief pillar of poetry of the Bektāshīyah Ṣūfī order, and many poets of this and other orders have imitated his style (though without reaching the same level of poetic truth and human warmth). Among the later poets claimed by the Bektāshīs may be mentioned Kaygusuz Abdal (15th century), who probably came from the European provinces of the Ottoman Empire. His verses are full of burlesque and even coarse images: in their odd mixture of worldliness and religious expression they are often as amusing as they are puzzling. In the 16th century, Pir Sultan Abdal (executed *c.* 1560) is noted for a few poems of austere melancholy. He was executed for collaboration with the Ṣafavids, the archenemies of the Ottomans; and in this connection it is worth remembering that the founder of the Iranian Ṣafavid dynasty, Shāh Esmāʿīl I (died 1524), wrote Turkish poetry under the pen name Khaṭāʾī and is counted among the Bektāshī poets.

Religious poetry. Mystically tinged poetry has always been very popular in Turkey, both in cities and rural areas. The best loved religious poem of all was, and still is, Süleyman Çelebi's (died 1419) *Mevlûd,* a quite short *mas̄navī* in honour of the Prophet Muḥammad's birth. This type of poetry has been known in the Islāmic countries since at least the 12th century and was soon adopted wherever Islām spread. There are a great number of *mevlûd* written in Turkish, but it was Süleyman Çelebi's unpretentious description of the great religious event that captured the hearts of the Turks; and it is still sung on many occasions (on the anniversary of a death, for example). The poem makes an excellent introduction to an understanding of the deep love for the Prophet felt by the pious Muslim.

Persian literature: 1300–1500. In the Iran of the Middle Ages, a vast number of poets flourished at the numerous courts. Not only professional poets but even the kings and princes contributed more or less successfully to the body of Persian poetry. Epics, panegyrics, and mystico-didactical poetry had all reached their finest hour by the end of the 13th century; the one genre to attain perfection slightly later was the *ghazal,* of which Moḥammad Shams od-Dīn Ḥāfeẓ (died 1389/90) is the incontestable master.

Lyric poetry: Moḥammad Shams od-Dīn Ḥāfeẓ. Ḥāfeẓ lived in Shīrāz; his pen name—"Who Knows the Qur'ān by Heart"—indicates his wide religious education, but little is known about the details of his life. The same is true of many Persian lyrical poets, since their products rarely contain much trustworthy biographical material. Ḥāfeẓ's comparatively small collection of work—his *Dīvān* contains about 400 *ghazal*s—was soon acclaimed as the finest lyrical poetry ever written in Persian. The discussion of whether or not to interpret its wine and love songs on a mystical plane has continued for centuries. Yet this discussion seems sterile since Ḥāfeẓ, whose verbal images shine like jewels, is an outstanding exponent of the ambiguous and oscillating style that makes Persian poetry so attractive and so difficult to translate. The different levels of experience are all expressed through the same images and symbols: the beloved is always cruel, whether a chaste virgin (a rare case in Persian poetry!) or a professional courtesan, or, as in most cases, a handsome young boy, or God himself, mysterious and unattainable—or even, on the political plane, the remote despot, the wisdom of whose schemes must never be questioned by his subjects. Since mystical interpretation of the world order had become almost second nature to Persians during the 13th century, the human beloved could effortlessly be regarded as God's manifestation; the rose became a symbol of highest divine beauty and glory; the nightingale represented the yearning and complaining soul; wine, cup, and cupbearer became the embodiment of enrapturing divine love. The poets' multicoloured images were not merely decorative embroidery but were a structural part of their thought. One must not expect Ḥāfeẓ (or any other poet) to unveil his personal feelings in a lyrical poem of experience. But no other Persian poet has used such complex imagery on so many different levels with such harmonious and well-balanced lucidity as did Ḥāfeẓ. His true greatness lies in this rather than in the content of his poetry. It must be stressed again that, according to the traditional view, each verse of a *ghazal* should be unique, precious for its own sake, and that the apparent lack of logic behind the sequence of verses was considered a virtue rather than a defect. (It may help to think of the glass pieces in a kaleidoscope, which appear in different patterns from moment to moment, yet themselves form no logical pattern.) To what extent an "inner rhythm" and a "contrapuntal harmony" can be detected in Ḥāfeẓ's poetry is still a matter for discussion; but that he perfected the *ghazal* form is indisputable. Whether he is praised as a very human love poet, as an interpreter of esoteric lore, or, as has been recently suggested, as a political critic, his verses have a continuing appeal to all lovers of art and artistry.

Use of a standard set of images and symbols

Parodies of classic forms. Ḥāfeẓ's contemporary in Shīrāz was the satirist ʿObeyd-e Zākānī (died 1371), noted for his obscene verses (even the most moralistic and mystical poets sometimes produced surprisingly coarse and licentious lines) and for his short *maṣnavī* called *Mūsh o-gorbeh* ("Mouse and Cat"), an amusing political satire. Since few new forms or means of expression were open to them, ʿObeyd and other poets began ridiculing the classic models of literature: thus, Bosḥāq (died *c.* 1426) composed odes and *ghazal*s exclusively on the subject of food.

Poetry of the Timurid period

The Timurid period in Iran produced only moderately good poetry, despite the rulers' interest in art. Allegorical *maṣnavī*s were much in vogue, such as the *Shabestān-e khayāl* ("Bedchamber of Fantasy") by the prolific writer Fattāḥī of Nīshāpūr (died 1448) and *Gūy o-chowgān* ("Ball and Polo-stick") by ʿĀrefī (died 1449); the latter work is an elaboration of the cliché that the lover is helpless before the will of his beloved, just as the ball is subject to the will of the polo-stick (". . . the head of the lover in the polo-stick of the beloved's tresses").

Eclecticism of ʿAbd or-Raḥmān Jāmī. The last great centre of Islāmic art in the region of Iran was the Timurid court of Herāt, where Dowlatshāh (died 1494) composed his much-quoted biographical work on Persian poets. The leading figure in this circle was ʿAbd or-Raḥmān Jāmī (died 1492), who is sometimes considered the last and most comprehensive of the "seven masters" in Persian literature, since he was a master of every literary genre and did not specialize in one form only, as Anvarī and Ḥāfeẓ, among others, had done. Jāmī wrote an excellent imitation of Neẓāmī's *Khamseh,* enlarging it by the addition of two mystical *maṣnavī*s into a septet called *Haft owrang* ("The Seven Thrones," or "Ursa Major"). His interest in Ṣūfism—he was initiated into the Naqshbandīyah order—is clear from his famous biographies of the Ṣūfī saints (which were an elaboration of a similar work by the 11th-century ʿAbd Allāh al-Anṣārī). In imitation of Saʿdī, Jāmī also composed the *Bahārestān* ("Orchard of Spring"), written in prose interspersed with verses. He left no less than three large divans, which contain work of high quality and demonstrate his gift for inventing picturesque images. Although his work abounds in lavishly ornamented verses, his style on the whole lacks the perfect beauty of Ḥāfeẓ's lyrics and is already tending toward the heavier, more opaque "Indian" style. Jāmī also wrote treatises about literary riddles and various kinds of intellectual games, of which Muslim society in the late 15th century was very fond and which remain a feature of erudite Persian and Turkish poetry. His influence on the work of later poets, especially in Ottoman Turkey, was very powerful.

An interesting aspect of the Timurid court in Herāt was the attention given to Chagatai Turkish, which was spoken in the eastern regions of Islām. ʿAlī Shīr Navāʾī, minister at the court (and a close friend of Jāmī), emphasized the beauties of his Turkic mother tongue as compared with Persian in his *Muḥākamat al-lughatayn* ("Judgment of the Two Languages"). He composed most of his lyrics and epics in Chagatai, which previously had been used by some members of the Timurid family and their courtiers for poetry but which became, thanks to him, an established literary medium. Even the arts-loving ruler of Herāt, Ḥusayn Bayqara (died 1506), wrote poetry in Turkic, following in every respect conventional literary taste.

Prose works: the "Mirror for Princes." During the first five centuries of Modern Persian literary life, a multitude of prose works were written. Among them, the "Mirror for Princes" deserves special mention. This genre, introduced from Persian into Arabic as early as the 8th century, flourished once more in Iran during the late 11th century. One important example is the *Qābūs-nāmeh* by the Zeyārid prince ʿOnṣor ol-Maʿālī Keykāvūs (died 1098), which presents "a miscellany of Islāmic culture in pre-Mongol times." At the same time, Niẓām al-Mulk (died 1092), the grand vizier of the Seljuqs, composed his *Seyāsat-nāmeh* ("Book of Government"), a good introduction to the statesman's craft according to medieval Islāmic standards. The *Seyāsat-nāmeh* was heavily influenced by pre-Islāmic Persian tradition. In the same period and environment, even a mystic like al-Ghazālī felt disposed to write a *Naṣīḥat al-mulūk* (Counsel for Kings), although the idealized relationship he makes between religious theory and practical statesmanship was not very realistic. A later mystic to compose a similar work was Sayyid ʿAlī Hamadhānī (died 1385), who had settled in Kashmir and initiated its Sūfi poetry. Others, especially in India, exhorted rulers in their writings.

The "Book of Government"

Belles lettres. Belles lettres proper found a fertile soil in Iran. The fables of *Kalīlah wa Dimnah,* for example, were retold several times in Persian. The most famous version, though a rather turgid one, is called *Anvār-e soheylī* ("Lights of Canopus") and was composed by a famous mystic, Ḥoseyn Wāʿeẓ-e Kāshefī of Herāt (died 1504). The "cyclic story" form (in which several unconnected tales are held together by a common framework or narrator device), inherited from India, became as popular in Iran as it had been in the Arabic-speaking countries. The *Sendbād-nāmeh* and the *Ṭūṭī-nāmeh* ("Parrot Book"), which is based on Indian tales, are both good examples of the popular method whereby a variety of instructive stories are skillfully strung together within a basic "running" story. The first comprehensive collection of entertaining prose is *Jawāmiʿ al-ḥikayat* ("Collections of Stories"), a veritable storehouse of tales and anecdotes, by ʿOwfī (died *c.* 1230). Anecdotes were an important feature of the biographical literature that became popular in Iran and Muslim India. Biographies of the poets of a certain age or of a specified area were collected together. They provide the reader with few concrete facts about the subjects concerned; but they abound in anecdotes, sayings, and verses attributed to the subjects, thus preserving material that otherwise might have been lost. Many of these biographical manuals, such as ʿOwfī's *Lubāb al-albāb* ("Quintessence of the Hearts") or Dowlatshāh's *Tazkirat ash-shuʿarā* ("Biography of the Poets"), make agreeable reading. The authors concerned wished to demonstrate their own erudition and rhetorical technique as much as to immortalize their subjects; consequently, their books are important equally as stylistic documents and as historical sources. One of the most remarkable works in this field is *Chahār maqāleh* ("Four Treatises") by Neẓāmī-ye ʿArūẓī, a writer from eastern Iran. Written in about 1156, this little book is an excellent introduction to the ideals of Persian literature and its writers, discussing in detail what is required to make a perfect poet, giving a number of instances of the sort of poetic craftsmanship thought especially admirable, and allowing glimpses into the various arts in which the literary man was expected to excel.

"Anecdotal" writing

This tendency toward "anecdotal" writing, which is also manifest in the work of a number of Arab historians, can be observed in the cosmographical books and in some of the historical books produced in medieval Iran. Ḥamdollāh Mostowfī's (died after 1340) cosmography, *Nuzhat al-qulūb* ("Pleasure of the Hearts"), like many earlier works of this genre, underlined the mysterious aspects of the marvels of creation and was the most famous of several instructive collections of mixed folkloristic and scientific material. Early miniaturists, too, loved to illustrate the most unlikely tales and pieces of information given in such works. Historical writing proper had been begun by the Persians as early as the late 10th century, when Balʿamī's abridged translation of aṭ-Ṭabarī's (died 923) vast Arabic chronicle first acquainted them with this outstanding piece of early Arabic historical literature. The heyday of historiography in Iran, however, was the Il-Khanid period (mid-13th to mid-14th century). Iran was then ruled by the successors of Genghis Khan, and scholars began to extend their interest back to the history of pre-Islāmic Central Asia, whence the rulers had come. *Tārīkh-e jehān-goshāy* ("History of the World Conqueror") by ʿAṭā Malek-e Joveynī (died 1283) and *Jāmiʿ at-tawārīkh* ("Collector of Chronicles") by the physician and vizier Rashīd ad-Dīn (executed 1318) are both outstanding examples of histories filled with valuable information. Although the writing of history became a firmly established art in Iran and the adjacent Muslim countries, the facts were unfortunately all too often concealed in a bombastic style and a labyrinth of cumbersome, long-winded sentences. A history written by Vaṣṣāf (died 1323) is the most notorious example of turgidity, but even his style was surpassed by some later writers. These stylistic tendencies deeply influenced Turkish prose writing: 17th-century Turkish historical works, such as those of Peçevi (died *c.* 1650) and Naima (died 1716), for this reason almost defy translation. Later Persian prose in India suffered from the same defects. This development in Persian and Turkish prose is also reflected in the handbooks on style and letter writing that were written during the 14th and 15th centuries and afterward. They urged the practice of all the artificial tricks of rhetoric by this time considered essential for an elegant piece of prose.

Popular literature. Islāmic literatures, however, should not be thought to consist only of erudite and witty court poetry, of frivolous or melancholy love lyrics full of literary conceits, or of works deeply mystical in content. Such works are counterbalanced by a great quantity of popular literature, of which the most famous expression is *Alf laylah wa laylah* (*The Thousand and One Nights,* also known as *The Arabian Nights' Entertainment*). The tales collected under this title come from different cultural areas; their nucleus is of Indian origin, first translated into Persian as *Hazār afsānak* ("Thousand Tales") and then into Arabic. These fanciful fairy tales were later expanded with stories and anecdotes from Baghdad. Subsequently, some tales—mainly from the lower strata of society—about rogues, tricksters, and vagabonds were added in Egypt. Independent series of stories, such as that of Sindbad the Sailor, were also included. The entire collection is very important as a reflection of several aspects of Oriental folklore and allows, now and then, glimpses into the court life of the various dynasties. Since its first translation into French (1704), it has inspired many Western readers' dreams about the "romantic" East.

The Thousand and One Nights

From pre-Islāmic times the Arabs had recounted tales of the *ayyām al-ʿArab* ("Days of the Arabs"), which were stories of their tribal wars, and had dwelt upon tales of the heroic deeds of certain of their brave warriors, such as ʿAntarah. Modern research, however, suggests that his story in its present setting belongs to the period of the Crusades. The Egyptian queen, Shajar ad-Durr (died 1250), and the first brave Mamlūk ruler, Baybars I (died 1277), as well as the adventures of the Bedouin tribe Banū Hilāl on its way to Tunisia, are all the subjects of lengthy popular tales.

In Iran, many of the historical legends and myths had been borrowed and turned into high literature by Ferdowsī. Accounts of the glorious adventures of heroes from early Islāmic times were afterward retold throughout Iran, India, and Turkey. Thus, the *Dāstān-e Amīr Ḥamzeh,* a story of Muḥammad's uncle Ḥamzah ibn ʿAbd al-Muṭṭalib, was slowly enlarged by the addition of more and more fantastic details. This form of *dāstān,* as such literature is called, to some extent influenced the first attempts at novel writing in Muslim India during the 19th century. The epics of Köroğlu are common to both Iranian and Turkish tradition. He was a noble warrior-robber who became one of the central figures in folk literature from Central Asia to Anatolia.

Some popular epics were composed in the late Middle Ages, having as their basis local traditions. One such epic had as its basis the Turco-Iranian legend of an 8th-century hero, Abū Muslim, another the Turkish tales of the knight Dānishmend. Other epics, such as the traditional Turkish tale of Dede Korkut, were preserved by storytellers who improvised certain parts of their tales (which were noted down only afterward). Also, the role of the Ṣūfī orders and of the artisans' lodges in preserving and transmitting such semihistorical popular epics seems to have been considerable. Apart from heroic figures, the Muslim peoples further share a comic character—basically a type of low-class theologian, called Nasreddin Hoca in Turkish, Juḥā in Arabic, and Mushfiqī in Tadzhik. Anecdotes about this character, which embody the mixture of silliness and shrewdness displayed by this "type," have amused generations of Muslims.

Popularity of romantic love stories

Shortly after the introduction of the printing press, Turkey and Iran began to produce cheap books, sometimes illustrated, containing popular romantic love stories. Large numbers of fairy tales were published in these cheap editions, and still other fairy tales have been collected by European and Muslim folklorists.

A truly popular poetry is everywhere to be found: lullabies sung by Baluchi, Kurdish, and Ibo mothers have obvious similarities; workers sing little rhythmical poems to accompany their work, and nomads remember the adventures of their ancestors in their ballads. Such popular poems often contain dialect expressions, and the metres differ from the classical quantitative system. Some of these simple verses, such as a two-line *lanḍay* in Pashto, are among the most graceful products of Islāmic poetry. Many folksongs—lullabies, wedding songs, and dirges—

have a distinct mystical flavour and reflect the simple Muslim's love for the Prophet and his trust in God's grace even under the most difficult circumstances. Irony and wit are features of the riddle poem, a favourite form among Muslims everywhere. Folk poets were also fond of humorous descriptions of imaginary disputations between two entities—they might compose dialogues between coffee and tobacco (Morocco), between a big and a small mosque (Yemen), between a cat and a dog, or between a boy and a girl. This kind of literature in the semicolloquial or dialectical Arabic poetry of the 17th and 18th centuries in Yemen, Upper Egypt, and central Arabia would bear a thorough study. All the Iranian and Turkic languages, too, possess a rich heritage of popular poetry, which in many cases appeals more immediately to modern tastes than does the rather cerebral high literature of the urban and court cultures.

THE PERIOD FROM 1500 TO 1800

According to Persian tradition, the last classic author in literature was Jāmī, who died in 1492. In that year, Christopher Columbus discovered America, and the Christians reconquered Granada, the last Moorish stronghold of Spain. The beginning of the 16th century was as crucial in the history of the Muslim East as in that of the Western Hemisphere. In 1501, the young Esmāʿīl founded the Ṣafavid rule in Iran, and the Shīʿah persuasion of Islām was declared the state religion. At the same time, the kingdoms of the last Timurid rulers in Central Asia were overthrown by the Uzbeks, who, for a while, tried to continue the cultural tradition in both Persian and Turkic at their courts in Bukhara. In 1526, after long struggles, one member of the Timurid house, Bābur, laid the foundation of the Mughal Empire in India. In the Near East, the Ottoman Turks, having expanded their empire (beginning in the late 13th century) from northwestern Anatolia into the Balkans, conquered crumbling Mamlūk Egypt and adjacent countries, including the sacred places of Mecca and Medina in 1516–17. Thus, three main blocks emerged, and the two strongholds of Sunnī Islām—Ottoman Turkey and Mughal India—were separated by Shīʿah Iran.

Decentralization of Islāmic literatures. Ṣafavid Iran, as it happened, lost most of its artists and poets to the neighbouring countries: there were no great masters of poetry in Iran between the 16th and 18th centuries. And while the Persian Shāh Esmāʿīl wrote Turkish mystical verses, his contemporary and enemy, Sultan Selim I of Turkey (died 1520), composed quite elegant Persian *ghazals*. Bābur (died 1530), in turn, composed his autobiography in Eastern Turkic.

Islāmic autobiographical literature

Bābur's autobiography is a fascinating piece of Turkish prose and at the same time one of the comparatively rare examples of Islāmic autobiographical literature. The classic example in this genre, however, was a lively Arabic autobiography by Usāmah ibn Munqidh (died 1188), which sheds much light upon the life and cultural background of a Syrian knight during the Crusades. A number of mystics, too, had written their spiritual autobiographies in a variety of languages, with varying degrees of artistic success. Bābur's book, however, gives a wonderful insight into the character of this intrepid conqueror. It reveals him as a master of concise, matter-of-fact prose, as a keen observer of daily life, full of pragmatic common sense, and also as a good judge of poetry. Bābur even went so far as to write a treatise in Turkish about versification. Many of his descendants, both male and female, inherited his literary taste and talent for poetry; among them are remarkably good poets in Persian, Turkish, and Urdu, as well as accomplished authors of autobiographies (Jahāngīr) and letters (Aurangzeb). Among the nobility of India, the Turkish language remained in use until the 19th century. Lovely Turkish verses were written, for example, by Akbar's general, Khān-e Khānān ʿAbd-ur-Raḥīm (died 1626), who was a great patron of fine arts and poetry.

In the Arab world, there was hardly a poet or original writer of note during the three centuries that followed the Ottoman conquest, apart from some theologians (ʿAbd al-Wahhāb ash-Shaʿrānī, died 1565; ʿAbd al-Ghanī an-Nābulusī, died 1731) and grammarians. Yet Arabic still remained the language of theology and scholarship throughout the Muslim world; both Turkey and India could boast a large number of scholars who excelled in the sacred language. In Ottoman Turkey, Taşköprüzāde (died 1560) compiled a historical survey of outstanding Turkish intellectuals in Arabic. Although a fine example of Islāmic learning, it does not compare in usefulness with the bibliographical work in Arabic by Hacı Halifa (Kâtib Çelebî; died 1658), which is a valuable source for modern knowledge of literary history.

New importance of Indian literature. India's share in the development of Arabic literature at this time was especially large. In addition to the quantity of theological work written in the language of the Qur'ān, from the conquest of Sind in 711 right up until the 19th century, much philosophical and biographical literature in Arabic was also being written in the subcontinent. Persian taste predominated in the northwest of India, but in the southern provinces there were long-standing commercial and cultural relationships with the Arabs, especially in Yemen and Ḥaḍramawt, and an inclination toward preserving these intact. Thus, much poetry in conventional Arabic style was written during the 16th and 17th centuries, mainly in the kingdom of Golconda. There are even attempts at the epic form. A century after the heyday of Arabic in the Deccan, Āzād Bilgrami (died 1786) composed numerous poetical and biographical works in Persian; but his chief fame was as the "Ḥassān of Hind," since he, like the Prophet Muḥammad's *protégé* Ḥassān ibn Thābit, wrote some powerful Arabic panegyrics in honour of the Prophet of Islām. He even attempted to make a comparison of the characteristics of Arabic and Sanskrit poetry and tried to prove that India was the real homeland of Islām. It should be added that al-Sayyid Murtaḍā az-Zabīd (died 1791), a leading philologist, author of the fundamental work of lexicography *Tāj al-ʿarūs* ("The Bride's Crown"), and commentator on Ghazālī's main work, was of Indian origin. Laudatory poems and belles lettres in Arabic were still popular in the early 19th century at the Shīʿite court of Lucknow, then the chief centre of Urdu poetry.

Poetry in conventional Arabic style

Indian literature in Persian. Nevertheless, the main contribution of Muslim India to high literature was made in the Persian tongue. Persian had been the official language of the country for many centuries. The numerous annals and chronicles that were compiled during the 14th and 15th centuries, as well as the court poetry, had been composed exclusively in this language even by Hindus. During the Mughal period, its importance was enhanced both by Akbar's attempt to have the main works of classical Sanskrit literature translated into Persian and by the constant influx of poets from Iran who came seeking their fortune at the lavish tables of the Indian Muslim grandees. At this time what is known as the "Indian" style of Persian emerged. The translations from Sanskrit enriched the Persian vocabulary, and new stories of Indian origin added to the reservoir of classical imagery. The poets, bound to the inherited genres of *mas̄navī, qaṣīdah,* and *ghazal,* tried to outdo each other in the use of complex rhyme patterns and unfamiliar, often stiff, metres. It became fashionable to conceive a poem according to a given *zamīn* ("ground"), in emulation of a classical model, and then to enrich it with newly invented tropes. The long-held ideal of "harmonious selection of images" was not always met. Difficult, even awkward grammatical constructions and inverted metaphors can be found. At times, pseudo-philosophical utterances in the second hemistich of a verse contrast strangely with semicolloquial expressions elsewhere. Objects recently introduced to India, such as the eyeglass or hourglass, were eagerly adopted as images by the poets, who wanted new-fangled conceits to bolster their tortuous inventiveness. Notwithstanding the colourful descriptive poems written in praise of such subjects as Mughal palaces, marvelously illuminated manuscripts, rare elephants, or court scenes, the general mood of lyric poetry became more gloomy. The transitory nature of the world, also a central theme in classical Persian poetry, was stressed and depicted in bizarre images: "burnt nest," "breakdown," "yawning" (indicating insatiable thirst); these were some of the new "stylish" words.

The works of ʿUrfī

Yet some truly great poets are to be found even in this period. ʿUrfī, who left Shīrāz for India and died in his mid-30s in Lahore (1592), is without doubt one of the few genuine masters of Persian poetry, especially in his *qaṣīdah*s. His verses pile up linguistic difficulties; yet their dark, glowing quality cannot fail to touch the hearts and minds even of critical modern readers—more so than the elegant but rather cerebral verses of his colleague Fayẕī (died 1595), one of Akbar's favourites. Fayẕī's brother Abū-ul-Faẕl ʿAllāmī (died 1602), the author of an important, though biased, historical work, deeply influenced the Emperor's religious ideas. Among 17th-century Mughal court poets, the most outstanding is Abū Ṭālib Kalīm (died 1651), who came from Hamadan. Abounding in descriptive passages of great virtuosity, his poignant and often pessimistic verses have become proverbial, thanks to their compact diction and fluent style. Also of some importance is Ṣā'ib of Tabriz (died 1677), who spent only a few years in India before returning to Iran. Yet, of his immense poetical output (300,000 couplets), the great majority belongs to the stock-in-trade expression of the Persian-speaking world. Other poets described the lives and adventures of members of the royal families, usually in verbose *mašnavī*s (this kind of descriptive historical poetry was practiced throughout Muslim India and also in Ottoman Turkey). Outside the Mughal environment, the lyrics and *mašnavī*s by Ẓuhūrī (died 1615) at the court of Bijāpur are charming and enjoyable. The heir apparent of the Mughal Empire, Dārā Shikōh (executed 1659), also followed Akbar's path. His inclination to mysticism is reflected in both his prose and poetry. The Persian translation of the *Upaniṣad*s, which he sponsored (and in part wrote himself), enriched Persian religious prose and made a deep impression on European idealistic philosophy in the 19th century. A group of interesting poets gathered about him, none of them acceptable to orthodoxy. They included the convert Persian Jew Sarmad (executed 1661), author of mystical *robāʿīyāt,* and the Hindu Brahman (died 1662), whose prose work *Chahār chaman* ("Four Meadows") gives an interesting insight into life at court. With the long rule of Dārā Shikōh's brother, the austere Aurangzeb (died 1707), the heyday of both poetry and historical writing in Muslim India was over. Once more, orthodox religious literature gained preeminence, while poets tried to escape into a fantasy world of dreams. The style of the two leading poets of this age, Nāṣir ʿAlī Sirhindī (died 1697) and Mīrzā Bēdil (died 1721), is convoluted and obscure, prompting the Persian poet Ḥazīn (died 1766), who came to India in the early 18th century, to write ironic comments about its incomprehensibility. Bēdil, however, was a very interesting writer. His lyric poetry is difficult but often rewarding, while his many philosophical *mašnavī*s deserve deep study. His prose work, interspersed with poetry, is called *Chahār ʿunṣur* ("Four Elements") and contains some biographical details. His prose is nearly as difficult as his poetry, and consequently his works rarely have been read west of India. His poetry, however, has had a great influence in Afghanistan and Central Asia. Many Persian-speaking people there consider him the forerunner of Tadzhik literature, since virtually everyone in Bukhara and Transoxania who tried his hand at poetry followed Bēdil's example. His ideas, sometimes astoundingly modern and progressive, have also impressed the 20th-century poet and philosopher Muḥammad Iqbāl in Muslim India.

Bēdil's progressive ideas

With Bēdil, the "Indian summer" of Persian literature comes to an end, even though the output of Persian poetry and prose during the 18th century in the subcontinent was immense. Some of the biographical dictionaries and handbooks of mysticism are valuable for the scholar but are less interesting as part of the general history of literature. The main vehicle of poetry now became Urdu, while mystical poetry flourished in Sindhi and Punjabi.

Pashto poetry: Khushḥāl Khān Khaṭak. From the borderlands of the Persian-speaking zone, culturally under the Mughal rule, one man deserves special attention. The chief of the Pashtun tribe of Khaṭak, Khushḥāl Khān (died 1689), rightly deserves to be called the "father" of Pashto poetry, for he virtually created a literature of his own in his mother tongue. His skill in translating the sophisticated traditions of Persian literature into the not too highly developed idiom of the Pashtuns is astonishing. His lively lyric poems are his finest works, reflecting that passionate love of freedom for which he fought against the Mughals. The poems he wrote from prison in "hell-like hot India" are as dramatic as they are touching in their directness. Many members of his family took to poetry; and during the 18th century original works, both religious and secular, were composed in Pashto, and the classics of Persian literature were translated into that language.

Ottoman Turkey. The development of literature in Ottoman Turkey is almost parallel with that of Iran and India. Yunus Emre had introduced a popular form of mystical poetry; yet the mainstream of secular and religious literature followed Persian models (although it took some time to establish the Persian rules of prosody because of the entirely different structure of the Turkish language). In the religious field, the vigour and boldness expressed in the poems of Nesimî (executed 1417) left their traces in the work of later poets, none of whom, however, reached his loftiness and grandeur of expression. The 14th- and 15th-century representatives of the classical style had displayed great charm in their literary compositions, their verses simple and pleasing. Sultan Cem (Jem; died 1495), son of Mehmed the Conqueror, is an outstanding representative of their number. But soon the high-flown style of post-classical Persian was being imitated by Ottoman authors, rhetoric often being more important to them than poetical content. The work of Bâkî (Bāqī; died 1600) is representative of the entire range of these Baroque products. Yet his breathtaking command of language is undeniable; it is brilliantly displayed in his elegy on Süleyman the Magnificent. In his time, according to a popular saying, one could find "a poet under every stone of Istanbul's pavement." Istanbul was the unique cultural centre of the Near East, praised throughout the ages by all who lived in the imperial city.

Poetry of Fuzûlî of Baghdad. Much greater than most of these minor poets, however, was a writer living outside the capital, Fuzûlî of Baghdad (died 1556), who wrote in Arabic, Persian, and Azeri Turkish. Apart from his lyrics, his Turkish *mašnavī* on the traditional subject of the lovers Majnūn and Laylā is admirable. From earliest times, Turkish poets had emulated the classical Persian romantic *mašnavī*s, sometimes surpassing their models in expressiveness. Fuzûlî's diction is taut, his command of imagery masterly. His style unfortunately defies poetical translation, and his complicated fabric of plain and inverted images, of hidden and overt allusions is well-nigh impossible for all but the initiated Muslim reader to disentangle. Fuzûlî, moreover, like his fellow poets, would blend Arabic, Persian, and Turkish constructions and words to make up a multifaceted unit. The same difficulty is found in Turkish prose literature of the same period. It is a major task to unravel the long trailing sentences of a writer such as Evliya Çelebî (died after 1679), who, in an account of his travels (*Seyahatnâme*), has left extremely valuable information about the cultural climate in different parts of the Ottoman Empire.

Emulation of classical Persian *mašnavī*s

Later developments. Growing interest in the Indo-Persian style, particularly in ʿUrfī's *qaṣīdah*s, led the 17th-century Ottoman poets to a new integrated style and precision of diction. An outstanding representative was Nefʿî, whose bent for merciless satire made him dreaded in the capital and eventually led to his assassination. At the start of the 18th century, a marked but short-lived movement in Turkish art known as the "Tulip Period" was the Ottoman counterpart of European Rococo. The musical poems and smooth *ghazal*s of Nedim (died 1730) reflect the manners and style of the slightly decadent, relaxed, and at times licentious high society of Istanbul and complement the miniatures of his contemporary Levnî. Good Turkish poetry is characterized by an easy grace, to be found even in such mystically tinged poems (thousands of which were written throughout the centuries) as those of Niyazî Misrî (died 1697). The Mevlevî (Mawlawī) poet Gâlib Dede (died 1799) was already standing at the threshold of what can now be recognized as modern poetical expression in some of the lyrical parts of his *mas̄navī,* called *Hüsn u aşk* ("Beauty and Love"), which brought fresh treatment to a well-worn subject of Iran's philosophical and secular literature. His work cannot be properly understood, however, without a thorough knowledge of mystical psychology, expressed in multivalent images.

Folk poetry. One branch of literature, however, was totally neglected by the sophisticated inhabitants of the Ottoman capital. Nobody thought much of the folk poets who wandered through the forgotten villages of Anatolia singing in simple syllable-counting verses of love, longing, and separation. The poems of the mid-17th-century figure Karacaoğlan, one of the few historically datable folk poets, give a vivid picture of village life, of the plight of girls and boys in remote Anatolian settlements. This kind of poetry was rediscovered only after the foundation of the Turkish Republic in 1923 and then became an important influence on modern lyric poetry.

EUROPEAN AND COLONIAL INFLUENCES: EMERGENCE OF WESTERN FORMS

The rise of nationalism. For the Islāmic countries, the 19th century marks the beginning of a new epoch. Napoleon's conquest of Egypt, as well as British colonialism, brought the Muslims into contact with a world whose technology was far in advance of their own. The West had experienced the ages of Renaissance, Reformation, and Enlightenment, whereas the once-flourishing Muslim civilization had for a long while been at a near stagnation point despite its remarkable artistic achievements. The introduction of Muslim intellectuals to Western literature and scholarship—the Egyptian aṭ-Ṭahṭāwī (died 1873), for example, studied in France—ushered in a new literary era the chief characteristic of which was to be "more matter, less art." The literatures from this time onward are far less "Islāmic" than those of the previous 1,000 years, but new intellectual experiences also led to "the liberation of the whole creative impulse within the Islāmic peoples" (Kritzeck). The introduction of the printing press and the expansion of newspapers helped to shape a new literary style, more in line with the requirements of the modern times, when "the patron prince has been replaced by a middle-class reading public" (Badawi). Translations from Western languages provided writers with the model examples of genres previously unknown to them, including the novel, the short story, and dramatic literature. Of those authors whose books were translated, Guy de Maupassant, Sir Walter Scott, and Anton Chekhov have been most influential in the development of the novel and the novella. Important also was the ideological platform derived from Tolstoy, whose criticism of Western Christianity was gratefully adopted by writers from Egypt to Muslim India. Western influences can further be observed in the gradual discarding of the time-hallowed static (and turgid) style of both poetry and prose; in the tendency toward simplification of diction; and in the adaptation of syntax and vocabulary to meet the technical demands of emulating Western models. Contact with the West also encouraged a tendency toward retrospection. Writers concentrated their attention on their own country and particular heritage, such as the "pharaoic myth" of Egypt, the Indo-European roots of Iran, and the Central Asian past of Turkey. In short, there was an emphasis on differentiation, inevitably leading to the rise of nationalism, instead of an emphasis on the unifying spirit and heritage of Islām.

Muslim intellectuals' introduction to Western scholarship

Arab literatures. Characteristically, therefore, given this situation, the heralds of Arab nationalism (as reflected in literature) were Christians. The historical novels of Jurjī Zaydān (died 1914), a Lebanese living in Egypt, made a deep impression on younger writers by glorifying the lion-hearted national heroes of past times. Henceforth, the historical novel was to be a favourite genre in all Islāmic countries, including Muslim India. The inherited tradition of the heroic or romantic epic and folktale was blended with novelistic techniques learned from Sir Walter Scott. Two writers in the front rank of Arab intellectuals were: Amīr Shakīb Arslān (died 1946), of Druze origin, and Muḥammed Kurd ʿAlī (died 1953), the founder of the Arab Academy of Damascus, each of whom, by encouraging a new degree of awareness, made an important contribution to the education of modern historians and men of letters. An inclination toward Romanticism can be detected in prose writing but not, surprisingly, in poetry; thus, the Egyptian al-Manfalūṭī (died 1924) poured out his feelings in a number of novels that touch on Islāmic as well as national issues.

Poetry. It is fair to say of this transition period that the poetry being written was not as interesting as the prose. The *qaṣīdah*s of the "Prince of Poets," Aḥmad Shawqī (died 1932), are for the most part ornate imitations of classical models. Even the "Poet of the Nile," Muḥammad Ḥāfiẓ Ibrahim (died 1932), who was more interested in the real problems of the day, was nonetheless content to follow conventional patterns. In his poems, Khalīl Muṭrān (died 1949) attempted to achieve a unity of structure hitherto almost unknown; and he also adopted a more subjective approach to expressive lyricism. Thus, he can be said to have inaugurated an era of "Romantic" poetry, staunchly defended by those men of letters who had come under English rather than French influence. These included the poet and essayist Ibrāhīm al-Māzinī (died 1949) and the prolific writer of poetry and prose ʿAbbās Maḥmūd al-ʿAqqād (died 1964).

Prose. A major contribution to the development of modern prose in the Arabic language was made by a number of writers born between 1889 and 1902. One of them, the "humanist" Taha Hussein, became well known in the West as a literary critic who attacked the historical authenticity of pre-Islāmic poetry and stressed the importance of Greek and Latin for the literatures of the modern Near East. He is also the author of a successful novel called *The Tree of Misery;* but his best creative writing is in his autobiographical notes, *al-Ayyām* ("The Days"), which describe in simple language the life of a blind Egyptian village boy. Taha Hussein's generation became more and more absorbed by the problems of the middle classes (to which most of them belonged), and this led them to realism in fiction. Some turned to fierce social criticism, depicting in their writings the dark side of everyday life in Egypt and elsewhere. The leading writer of this group is Maḥmūd Taymūr, who wrote short stories, a genre developed in Arabic by a Lebanese Christian who settled in the United States, the noted and versatile poet Khalil Gibran (Jibrān Khalīl Jibrān; died 1931). Muḥammad Ḥusayn Haykal (died 1956), a leading figure of Egyptian cultural and political life and the author of numerous historical studies, touched for the first time, in his novel *Zaynab* (1913), on the difficulties of Egyptian villagers. This subject quickly afterward became fashionable, although not all the writers had firsthand knowledge of the feelings and problems of the fellahin. The most fertile author of this group was al-ʿAqqād, who tirelessly produced biographies, literary criticism, and romantic poetry. To what extent the Islāmic reform movement led by Muḥammad ʿAbduh (died 1905) and his disciples, which centred on the journal *al-Manār* ("The Lighthouse"), has influenced present-day Arabic prose style cannot yet be ascertained. It has, however, been important in shaping the religious outlook of many authors writing in the 1920s and 1930s.

The diaspora. A considerable amount of Arabic literature has been produced by numerous writers who settled in non-Islāmic countries, especially in the United States and Brazil. Most of these writers came from Christian Lebanese families. A feeling of nostalgia often led them to form literary circles or launch magazines or newspapers. (The Arabic-language newspaper *al-Hudā* [or *Al-Hoda,* "The Guidance"], established in 1898, was published in New York City as *al-Hudā al-jadīdah* [*Al-Hoda Aljadidah,* or "The New Al-Hoda," or "The New Guidance"].) It was largely because of their work that the techniques of modern fiction and modern free verse entered Arabic literature and became a decisive factor in it.

The recurring theme in Lebanese fiction

One of the best known authors in this group was Amīn ar-Riḥānī (died 1941), whose descriptions of his journeys through the Arab world are informative and make agreeable reading. The fact that so many Lebanese emigrated to foreign countries led to the creation of a standard theme in Lebanese fiction: the emigrant who returns to his village.

Iraqi modern literature is best represented by "the poet of freedom" Maʿrūf ar-Ruṣāfī (died 1945), and Jamīl Sidqī az-Zahāwī (died 1936), whose satire "Rebellion in Hell" has incurred the wrath of the traditionalists.

Turkish literatures. The same changing attitude toward the function of literature and the same shift toward realism can be observed in Turkey. After 1839, Western ideas and forms were taken up by a group of modernists: Ziya Paşa (died 1880), the translator of Rousseau's *Émile* (which became a popular textbook for 19th-century Muslim intellectuals), was among the first to write in a less traditional idiom and to complain in his poetry—just as Ḥālī was to do in India a few years later—about the pitiable conditions of Muslims under the victorious Christians. Ziya Paşa, together with Şinasi (died 1871) and Namık Kemal (died 1888), founded an influential Turkish journal, *Tasvir-i Efkâr* ("Picture of Ideas"). The essential theme of the articles, novels, poems, and dramas composed by these authors is their fatherland (*vatan*), and they dared to advocate freedom of thought, democracy, and constitutionalism. Abdülhak Hâmid (died 1935), though considerably their junior, shared in their activities. In 1879 he published his epoch-making *Sahra* ("The Country"), a collection of ten Turkish poems that were the first to be composed in Western verse forms and style. Later, he turned to weird and often morbid subject matter in his poetic dramas. He, like his colleagues, had to endure political restrictions on writing, imposed as part of the harsh measures taken by Sultan Abdülhamid II against the least sign of liberal thought. Influenced by his work, later writers aimed to simplify literary language: Ziya Gökalp (died 1924) laid the philosophical foundations of Turkish nationalism; and Mehmed Emin, a fisherman's son, sang artless Turkish verses of his pride in being a Turk, throwing out the heavy rhetorical ballast of Arabo-Persian prosody and instead turning to the language of the people, unadulterated by any foreign vocabulary. The stirrings of social criticism could be discerned after 1907. Mehmed Akif (died 1936), in his masterly narrative poems, gave a vivid critical picture of conditions in Turkey before World War I. His powerful and dramatic style, though still expressed in traditional metres, is a testimony to his deep concern for the people's sorrows. It was he who composed the Turkish National Anthem after Mustafa Kemal Atatürk's victory; but soon afterward he left the country, disappointed with the religious policies of the Kemalists.

Atatürk's struggle for freedom also marks the real beginning of modern Turkish literature. The mainstream of novels, stories, and poems written during the 19th century had been replete with tears, world-weariness, and pessimism. But a postwar novel, *Ateşten gömlek* ("The Fire Shirt"), written by a woman, Halide Edib, reflected the brave new self-awareness of the Turkish nation. Some successful short stories about village life came from the pen of Ömer Seyfeddin (died 1920). The most gifted interpreter and harshest critic of Turkey's social structure was Sabaheddin Ali, who was murdered on his flight to Bulgaria in 1948. His major theme was the tragedy of the lower classes, and his writing is characterized by the same merciless realism that was later to be a feature of stories by many left-wing writers throughout the Islāmic world. The "great old man of Turkish prose," Yakup Kadri Karaosmanoğlu, displayed profound psychological insight, whether ironically describing the lascivious life in a Bektāshī centre or a stranger's tragedy in an Anatolian village. Most of the Turkish novelists of the 1920s and 1930s concentrated on the problems of becoming a modern nation, and in particular they reinterpreted the role of women in a liberated society.

The "great old man of Turkish prose"

Literary energies were set completely free when Atatürk introduced the Latin alphabet in 1928, hoping that his people would forget their Islāmic past along with the Arabic letters. From this time onward, especially after the language reform that was meant to rediscover the pre-Islāmic roots of the Turkish language, Turkish literature followed the pattern of Western literature in all major respects, though with local overtones. Poets experimented with new forms and new topics. They discovered the significance of the Anatolian village, neglected—even forgotten—during the Ottoman period. Freeing themselves from the traditional rules of Persian poetry, they adopted simpler forms from Europe. In some cases the skillful blending of inherited Ottoman grace and borrowed French lyricism produced outstandingly beautiful poems, such as those of Ahmed Haşim (died 1933) and of Yahya Kemal Beyatlı (died 1958), in which the twilight world of old Istanbul is mirrored in soft, evocative hues and melodious words. At the same time, the figure of Nazım Hikmet (died 1963) looms large in Turkish poetry. Expressing his progressive social attitude in truly poetical form, he used free rhythmical patterns quite brilliantly to enrapture his readers; his style, as well as his powerful, unforgettable images, has deeply influenced not only Turkish but also progressive Urdu and Persian poetry from the 1930s onward.

Persian literatures. In Iran, the situation to a certain extent resembled that in Turkey. While the last "classical" poet, Qāʾānī (died 1854), had been displaying the traditional glamorous artistry, his contemporary, the satirist Yaghmā (died 1859), had been using popular and comprehensible language to make coarse criticisms of contemporary society. As in the other Islāmic countries, a move toward simplicity is discernible during the last decades of the 19th century. The members of the polytechnic college Dār ol-Fonūn (founded 1851), led by its erudite principal Reẕā Qolī Khān Hedāyat, helped to shape the "new" style by making translations from European languages. Shāh Naṣer od-Dīn himself described his journeys to Europe in the late 1870s in a simple, unassuming style and in so doing set an example to future prose writers.

At the turn of the century, literature became for many younger writers an instrument of modernization and of revolution in the largest sense of the word. No longer did they want to complain, in inherited fixed forms, of some boy whose face was like the moon. Instead, the feelings and situation of women were stated and interpreted. Their oppression, their problems, and their grievances are a major theme of literature in this transition period of the first decades of the 20th century. The "King of Poets," Bahār (died 1951), who had been actively working before World War I for democracy, now devoted himself to a variety of cultural activities. But his poems, though highly classical in form, were of great influence; they dealt with contemporary events and appealed to a wide public.

One branch of modern Persian literature is closely connected with a group of Persian authors who lived in Berlin after World War I. There they established the Kaviani Press (named after a mythical blacksmith called Kaveh, who had saved the Iranian kingdom), and among the poems they printed were several by ʿAref Qazvīnī (died 1934), one of the first really modern writers. They also published the first short stories of Moḥammad ʿAlī Jamālzādeh, whose outspoken social criticism and complete break with the traditional inflated and pompous prose style inaugurated a new era of modern Persian prose. Many young writers adopted this new form, among them Ṣādeq Hedāyat (died 1951), whose stories—written entirely in a direct, everyday language with a purity of expression that was an artistic achievement—have been translated into many languages. They reflect the sufferings of living individuals; instead of dealing in literary clichês, they describe the distress and anxiety of a hopeless youth. The influence of Franz Kafka (some of whose work Hedāyat translated) is perceptible in his writing, and he has a tendency toward psychological probing shared by many Persian writers.

Authors of the Kaviani Press

As in neighbouring countries, women played a considerable role in the development of modern Persian literature. The lyrics of Parvīn Eʿteṣāmī (died 1940) are regarded as near classics, despite a trace of sentimentality in their sympathetic treatment of the poor. Some Persian writers whose left-wing political ideas brought them into conflict with the government left for the Tadzhik S.S.R. Of these, the gifted poet Lāhūtī (died 1957) is their most important representative.

India: Urdu and Persian. Persian literature in the Indian subcontinent did not have such importance as in earlier centuries, for English replaced Persian as the official language in 1835. Nevertheless, there were some outstanding poets who excelled in Urdu. One of them was Mīrzā Asadullāh Khān Ghālib (died 1869), the undisputed master of Urdu lyrics. He regarded himself, however, as the leading authority on high Persian style and was an accomplished writer of Persian prose and poetry. But much more important was a later poet, Sir Muḥammad Iqbāl (died 1938), who chose Persian to convey his message not only to the peoples of Muslim India but also to Afghans and Persians. Reinterpreting many of the old mystical ideas in the light of modern teachings, he taught the quiescent Muslim peoples self-awareness, urging them to develop their personalities to achieve true individualism. His first *maṡnavī*, called *Asrār-e khudī* (1915; "Secrets of the Self"), deeply shocked all those who enjoyed the dreamlike sweetness of most traditional Persian poetry. One of his later Persian works, *Payām-e Mashriq* (1923; "Message of the East"), is an effective answer to Goethe's *West-östlicher Divan* (1819). In the *Jāvīd-nāmeh* (1932) he poetically elaborated the old topic of the "heavenly journey," discussing with the inhabitants of the spheres a variety of political, social, and religious problems. Iqbāl's approach is unique. Although he used the conventional literary forms and leaned heavily on the inspiration of Jalāl ad-Dīn ar-Rūmī, he must be considered one of the select few poets of modern Islām who, because of their honesty and their capacity for expressing their message in memorable poetic form, appeal to many readers outside the Muslim world.

THE MODERN PERIOD

The modern period of Islāmic literatures can be said to begin after World War II. The topics discussed before then still appeared, but outspoken social criticism became an even more important feature. Literature was no longer a leisurely pastime for members of the upper classes. Writers born in the villages and from non-privileged classes began to win literary fame through their firsthand knowledge of social problems. Many writers started their careers as journalists, developing a literary style that retained the immediacy of journalistic observation.

Prose. In Egypt, a great change in literary preoccupations came about after 1952. The name of Najīb Maḥfūẓ is of particular importance. He was at first a novelist mainly concerned with the lower middle classes (his outstanding work is a trilogy dealing with the life of a Cairo family); but afterward he turned to socially committed literature, using all the techniques of modern fiction—of which he is the undisputed master in Arabic. Yūsuf Idrīs deals first and foremost with the problems facing poor and destitute villagers, a subject also treated in Sharqāwī's novel *al-Arḍ* (*The Earth;* 1954). In Turkey, Yaşar Kemal's village story *İnce Memed* has won acclaim for its stark realism. The young left-wing writers in Iraq and Syria share the critical and aggressive attitudes of their contemporaries in Turkey and Egypt and are involved in every political issue. Most of them have responded to the works of Bertolt Brecht and Karl Marx. They are also quite familiar with at least the externals of modern psychology. Freudian influence—often in its crudest form—can be detected in many modern short stories or novels in the Islāmic countries; and it is often the prelude to coarse descriptions of sexuality, appealing to the lowest instincts of the reading public. In the Near and Middle East, the existentialist philosophy gained many followers who tried to reflect its interpretation of life in their literary works. In fact, almost every current of modern Western philosophy and psychology, every artistic trend and attitude, has been eagerly adopted—though often only half-understood—by young Arab, Turkish, or Persian writers. Some of them, nevertheless, have achieved interesting results from time to time: an example is Laylā Baʿlabakkī, whose semiautobiographical novel, *I Live,* is regarded as an outstanding literary achievement in Arabic.

Poetry. *Arabic.* The new attitudes that have informed literature are even more conspicuous in poetry than in prose. Arabic poetry has at last freed itself completely from the fetters of classical tradition. Both French and English influences helped to shape the new art. The danger is that Western fashions are imitated uncritically, just as Arabic, Persian, or Turkish models were slavishly followed in the past. T.S. Eliot's poetry and criticism were influential in dethroning the Romanticism that many poets had adopted earlier, in the 1920s and '30s. One of the first and most important attempts at creating a modern Arabic poetic diction was made in the late 1940s by the Iraqi poet and critic Nāzik al-Malāʾikah, whose poems, in free but rhyming verse, give substance to the shadow of her melancholia. Free rhythm and a colourful imagination distinguish the best poems of the younger Arabs: even when their poems do not succeed, their experimentation, their striving for sincerity, their burning quest for identity, their rebellion against social injustice can be readily perceived. Indeed, one of the most noticeable aspects of contemporary Arabic poetry is its political engagement, evident in the poems of Palestinian writers such as Maḥmūd Darwīsh, whose verses once more prove the strength, expressiveness, and vitality of the Arabic language. An Iraqi, ʿAbdul Wahhāb al-Bayātī, combines political engagement with lyrical mysticism. Others, without withdrawing into a world of uncommitted dreams, manage to create an atmosphere that breaks up the harsh light of reality into its colourful components. Poets like the Lebanese Adonis (ʿAlī Aḥmad Saʿīd) and Tawfīq aṣ-Ṣāʾigh, or the Egyptian dramatist Ṣalāḥ ʿAbd aṣ-Ṣabur, make use of traditional imagery in a new, sometimes esoteric, often fascinating and daring way.

New use of traditional imagery

Persian. Almost the same situation developed in Iran. One notable poet was Forugh Farrokhzād, who wrote powerful yet very feminine poetry. Her free verses, interpreting the insecurities of the age, are full of longing; though often bitter, they are yet truly poetic. Poems by such critically minded writers as Seyāvūsh Kasrāʾī also borrow the classical heritage of poetic imagery, transforming it into expressions that win a response from modern readers.

Turkish. In Turkey, the adoption of Western forms began in the 1920s. Of major importance in modern Turkish literature was Orhan Veli Kanık, who combined perfect technique with "Istanbulian" charm. His work is sometimes melancholy, sometimes frivolous, but always convincing. He strongly influenced a group of poets whose names are connected with the avant-garde literary magazine *Varlik* ("Existence"). The powerful poetry of the leftist writer Nazım Hikmet has influenced progressive poets all over the Muslim world. It is still too early, however, to determine what will be most representative of modern Turkish poetry: a return to Anatolian subjects, sometimes in picturesque diction, influenced by earlier folk-poetry, or the continuation of lovely poems in praise of Istanbul; surrealism or a somewhat detached and ironic approach to a subject. The same question, indeed, could be raised of almost any contemporary literature in Islāmic lands.

The continuing problem of language

Contemporary features. In the Arab-speaking world, the problem of language has loomed large for many years. Classical high Arabic is still the common literary language of Morocco and Iraq, Tunisia and Kuwait. Spoken Arabic in dialectal variations is beginning to be used—but tentatively—in higher literature. It is more frequently employed in the popular spheres of theatre and cinema. But the local differences that exist in Arabic spoken from country to country have today become perceptible in literature; popular grammatical forms and syntactical constructions are occasionally used in modern poetry. A special problem arises in the North African countries, where French continues to be the chief literary language for most writers, especially in Morocco and Algeria. Yet there is no hard and fast rule: a leading member of the Senegal community, Amadou Bamba, who founded the politically important group of the Murīdīs, wrote (quite apart from practical words of wisdom in his mother tongue) some 20,000 mystically tinged verses in classical Arabic.

Throughout the Islāmic countries, the press and radio have helped to disseminate literary works; prizes for literary achievements have stimulated interest in writing; low-priced books have made the more or less valuable output of a growing number of writers available to the majority—the more so since literacy among the population steadily increases. But to what degree this means a continuation of the cultural role that Islāmic literatures have played in the formation and education of society over the centuries is not yet clear. Literature was never restricted to a privileged high society; in olden times even the illiterate villager and the "uneducated" womenfolk had a fund of poems, proverbs, songs, and quotations from classical sources that they knew by heart and to which they turned for both pleasure and spiritual strength.

One final symptom should be noted. The introduction of modern methods of criticism, of psychology and philosophy, has kindled a new interest in significant figures of the Islāmic past. Thus, to quote one instance, the figure of al-Ḥallāj (executed 922), who often served as a symbol figure of "the martyr of love" in both classical and folk poetry after the 11th century, has in recent years been made the subject of a Turkish drama, a Persian passion play, and an Arabic tragedy and plays an important role in Arabic, Turkish, Persian, and Indian Muslim lyrical poetry. He is interpreted as a symbol of suffering for one's ideals, and he is therefore acceptable both to conservative Muslims and to progressive social critics.

STUDY AND EVALUATION

The controversy over word and meaning

Early Islāmic criticism. The development of literature during the early Middle Ages soon produced among the Arabs much lively literary criticism. Even the choice of quotations made by the ancient grammarians from the classical stock of poetry implies a degree of critical (though subjective) activity. Attempts toward making a more objective study of poetic technique were first made in the late 9th century, when for the first time "beauties" and "faults" of verses were discussed and the ideals of the "new style" were defined by Ibn al-Mu'tazz in his *Kitāb al-badī'*. The relation between *lafẓ* (word) and *ma'nā* (meaning) has been a matter of some controversy—many earlier critics stress the importance of outward form rather than of content. There was some question, too, as to whether the most "poetical" verse was that which was the most "untrue"—that is to say, hyperbolic—or that which was closer to the heart of things. The matter was debated along with the problem of inspiration and imagination and their function in poetry. The most thorough analysis of the art of poetry was made by 'Abd al-Qāhir al-Jurjānī, who allowed equal weight to the idea and to the way it was expressed. An illuminating work about poetics was composed by the Tunisian critic al-Qarṭājannī (13th century), and this has been carefully studied by the German scholar Wolfhart Heinrichs in *Arabische Dichtung und griechische Poetik* (1969). This study analyzes al-Qarṭājannī's theories in relation to Aristotle's theories of poetics. (Heinrichs, one of the few Islāmic scholars specializing in the study of literary problems, has shown that classical Arabic criticism rarely interested itself in the poem as a whole but concentrated upon individual verses.) In later centuries, manuals of poetics and rhetoric written in every Islāmic country reveal the prevailing interest in purely formal problems.

Modern criticism. A similar interest long dominated the work of Western Orientalists. The first scholars who attempted to introduce Persian poetry to Western readers (such as Sir William Jones in the 18th century) thought it necessary to compare it with the compositions of Greek and Latin poetry. The verbal ingenuity of Ḥarīrī's *Maqāmāt* attracted the European scholars, who took great pleasure in disentangling the grammatically difficult forms. Pre-Islāmic poetry at first interested only the grammarian-antiquarian until its importance as a source of knowledge of early Bedouin life was recognized. The art of versification and problems of classical Arabic metrics are forever matters of discussion among Orientalists.

Although a large amount of translation, mainly from Persian poetry, was produced in the 18th and 19th centuries, most of it suffered for lack of proper understanding: the translators took the poetical statements about wine and love or the outbursts against established religious forms at their face value and failed to recognize them for the stereotyped forms and images they are. A deep study of the imagery of Persian, Turkish, and Arabic is required before their poetry and belles lettres can be properly understood and enjoyed. This was realized as early as 1818 by the Austrian Orientalist Joseph von Hammer-Purgstall (whose own translations from the three great Islāmic languages are, nevertheless, failures).

In the 20th century the critical study of imagery in Oriental poetry was taken up by Hellmut Ritter in his booklet *Über die Bildersprache Niẓāmīs* (1927; "On the Imagery of Neẓāmī"), which gives a most sensitive philosophical interpretation of Neẓāmī's metaphorical language and of the role of imagery in the structure of Neẓāmī's thought. Ritter's criticism is basic to the study of many other Persian poets. Slightly later, the Polish scholar Tadeusz Kowalski tried to interpret the "molecular" structure of Arabic literature—the absence of large units of thought or architectural structure—typical of the greater part of Islāmic literatures, which might be described as "carpet-like." This "molecular" structure can be related to the atomist theories and occasionalist world view embodied in Islāmic theology, which, unlike Christianity, does not admit of secondary causes and requires only short spans of hope from the faithful. In a number of articles, and in many books, E.G. von Grunebaum has pioneered this interpretation of literary structure. Other important critical works include S.A. Bonebakker's book on the rhetorical importance of *tawrīyah* (ambiguous wording); Manfred Ullmann's excellent study of *rajaz*-poetry and its place in Arabic literature; and C.H. de Fouchécour's detailed analysis of the descriptions of nature in early Persian poetry.

Among the Arabs themselves, modern literary criticism began during the early 1920s. Most famous was Ṭaha Ḥussein's attempt to prove the whole corpus of pre-Islāmic poetry as counterfeit. All the Islāmic countries, from Turkey to Pakistan, and especially Iran have sponsored reviews in which Western-trained scholars critically survey the literary achievements of the Islāmic world.

A full evaluation of literature as the most faithful mirror of past (and to some extent present-day) Islāmic life is still lacking. Notwithstanding the conventionalized style of most Islāmic poetry, a deeper study of individual poets' expressions, use of verbal and nominal forms, rhythmical preferences, and the like would certainly reveal more about the personalities of outstanding writers. The impact of poetry on the Islāmic mind was, and to some extent still is, much deeper than a modern Western reader might suppose. The poets must be viewed, therefore, in relation to their society, for their work corresponded to the measure of receptiveness, their new modes of expression developed according to the widening awareness of their audiences. They had to use a language and imagery to which those whom they addressed were accustomed. A new idea, embodied in traditional imagery and a beguiling metre, could capture the attention of thousands of people. The role of the poet as religious and political herald (even though his political thought was all too often subservient to courtly flattery) was widely acknowledged, and the impact of a poet like Muḥammad Iqbāl bears witness to the real power of poetical expression. Thus, even the most conventional Persian or Turkish poem can reflect certain attitudes of the Muslim mind more accurately than many a learned lecture. A modern short story, even if not particularly well wrought, often tells the reader more about the feelings and reactions of the people than scholarly sociological research papers can. The magic of language is still a living force in the East. (An.Sc.)

Music

The period of Islāmic music begins with the advent of Islām in about 610. A new art emerged, elaborated both from pre-Islāmic Arabian music and from important contributions by Persians, Byzantines, Turks, Berbers, and Moors. In this development the Arabian element acted as a catalyst, and, within a century, the new art was firmly established from Central Asia to the Atlantic. Such a fusion of musical styles succeeded because there were strong affinities between Arabian music and the music of the nations occupied by the expanding Arabic peoples. Not all Arab-dominated areas adopted the new art; Indonesia and parts of Africa, for example, retained native musical styles. The folk music of the Berbers in North Africa, the Moors in Mauretania, and other ethnic groups (*e.g.,* in Turkey) also remained alien to classical Islāmic music. The farther one looks from the axis reaching from the Nile Valley to Persia, the less one finds undiluted Islāmic music.

(It should be remembered that the word *music* and its concept were reserved for secular art music; separate names and concepts belonged to folk songs and to religious chants.)

NATURE AND ELEMENTS OF ISLĀMIC MUSIC

Islāmic music is characterized by a highly subtle organization of melody and rhythm, in which the vocal component predominates over the instrumental. It is based on the skill of the individual artist, who is both composer and performer and who benefits from a relatively high degree of artistic freedom. The artist is permitted, and indeed encouraged, to improvise. He generally concentrates on the details forming a work, being less concerned with following a preconceived plan than with allowing the music's structure to emerge empirically from its details. Melodies are organized in terms of *maqāmāt* (singular *maqām*), or "modes," characteristic melodic patterns with prescribed scales, preferential notes, typical melodic and rhythmic formulas, variety of intonations, and other conventional devices. The performer improvises within the framework of the *maqām,* which is also imbued with ethos (Arabic *ta'thīr*), a specific emotional or philosophical meaning attached to a musical mode. Rhythms are organized into rhythmic modes, or *īqā'āt* (singular *īqā'*), cyclical patterns of strong and weak beats.

Classical Islāmic music is the aristocratic music of the court and the upper class, which underwent development and modification in the hands of gifted musicians throughout several centuries. Rhythmic and melodic modes grew in number and complexity, and new vocal and instrumental genres arose. In addition, a body of theoretical works grew up, influencing both Islāmic and—in some cases—European music. Its later popularization did not alter its intimate and entertaining character.

The relation of music to poetry and dance. In pre-Islāmic times music was closely connected with poetry and dance. Being essentially vocal, pre-Islāmic music was an emotional extension of the solemn declamation of poems in Bedouin society. Later, the art of vocal composition itself was largely based upon prosody: only by respecting the poetic metre in the music could the text, when sung, be clear in meaning and correct in pronunciation and grammatical inflection. In turn, prosody itself was used to explain the musical rhythm.

The poet-musician

Words and rhetorical speech were the principal means through which the Bedouin expressed feelings. The *shā'ir,* or poet-musician, said to be possessed by supernatural powers, was feared and respected. His satirical song poems were a formidable arm against enemies, and his poems of praise enhanced the prestige of his tribe. Musician-poets, especially women, accompanied the warriors, inciting them by their songs, and those who fell in battle benefited from the elegies of the singer-poets. Musically, these elegies resembled the *ḥudā'* ("caravan song"), possibly used by camel drivers as a charm against the desert spirits, or *jinn.*

Music and dance were closely associated from early times. Bedouin music had a pronounced collective character, with well-defined functions and usages, and dance occupied an important place in Bedouin life. Most common was a simple communal dance that emphasized common, or social, rather than individual movement. Places of entertainment in the towns and oases employed professional dancers, mainly women. Art dancing embellished events in the courts of the Sāsānians, the pre-Islāmic rulers of Persia. In the Islāmic period, solo and ensemble forms of dance were an integral part of the intense musical activity in the palaces of the caliphs and in wealthy houses. Dance also was prominent in the *dhikr* ceremony of certain mystical fraternities; forms ranged from obsessional physical movements to refined styles similar to those of secular art dancing.

After the advent of Islām a deep change occurred in the social function of music. Emphasis was laid on music as entertainment and sensual pleasure rather than as a source of high spiritual emotion, a change mainly resulting from Persian influence. Knowledge of music was obligatory for the cultured person. Skilled professional musicians were highly paid and were admitted to the caliphs' palaces as courtesans and trusted companions. The term *ṭarab,* which designates a whole scale of emotions, characterizes the musical conception of the time and even came to mean music itself.

Music and religion. Fashionable secular music—and its clear association with erotic dance and drinking—stimulated hostile reactions from religious authorities. As Muslim doctrine does not sanction permitting or prohibiting a given practice by personal decision, the antagonists relied on forced interpretations of a few unclear passages in the Qur'ān (the sacred scripture of Islām) or on the *Ḥadīth* (traditions of the Prophet, sayings and practices that had acquired force of law). Thus both supporters and adversaries of music found arguments for their theses.

In the controversy, four main groups emerged: (1) uncompromising purists opposed to any musical expression; (2) religious authorities admitting only the cantillation of the Qur'ān and the call to prayer, or *adhān;* (3) scholars and musicians favouring music, believing there to be no musical difference between secular and religious music; and (4) important mystical fraternities, for whom music and dance were a means toward unity with God.

Religious music

Except in the Ṣūfī brotherhoods, Muslim religious music is relatively curtailed because of the opposition of religious leaders. It falls into two categories: the call to prayer, or *adhān* (in some places, *azān*), by the *mu'adhdhin,* or muezzin, and the cantillation of the Qur'ān. Both developed from relatively solemn cantillation to a variety of forms, both simple and highly florid. The cantillation of the Qur'ān reflected the ancient Arabic practice of declamation of poetry, with careful regard to word accents and inflections and to the clarity of the text. Yet it was possibly also influenced by early secular art song. Opponents of music considered the cantillation of the Qur'ān to be technically distinct from singing, and it acquired a separate terminology. Synagogues and the Eastern Christian churches, unhampered by such opposition, developed extensive musical repertories based on melodic modes: the Eastern churches used the eight modes of Byzantine music, while synagogue music followed the *maqām* system of Muslim art music.

Aesthetic traditions. Even in its most complicated aspects, Islāmic music is traditional and is transmitted orally. A rudimentary notational system did exist but it was used only for pedagogical purposes. A large body of medieval writing about music survives in which musical theory is related to various areas of intellectual activity, hence the extreme importance of understanding music as an element of the culture involved. The medieval writings fall mainly into two categories: (1) literary, encyclopaedic, and anecdotal sources, and (2) theoretical, speculative sources. The first group includes precious information on musical life, musicians, aesthetic controversies, education, and the theory of musical practice. The second deals with acoustics, intervals (distances between notes), musical genres, scales, measures of instruments, the theory of composition, rhythm, and the mathematical aspects of music. These documents show that, as in the modern era, medieval Islāmic music was principally an individual, soloistic art. Small ensembles were actually groups of soloists with the principal member, usually the singer, predominating. Being an essentially vocal music, it displayed many singing and vocal techniques, such as special vocal colour, guttural nasality, vibrato, and other stylistic ornaments. Although the music was based upon strict rules, preexisting melodies, and stylistic requirements, the performer enjoyed great creative freedom. The artist was expected to bring his contribution to a given traditional piece through improvisation, original ornamentation, and his own approach to tempo, rhythmic pattern, and the distribution of the text over the melody. Thus the artist functioned as both performer and composer.

The soloist's role

Melodic organization. Islāmic music is monophonic; *i.e.,* it consists of a single line of melody. In performance everything is related to the refinement of the melodic line and the complexity of rhythm. The notion of harmony is completely absent, although occasionally a simple combination of notes, octaves, fifths, and fourths, usually below the melody notes, may be used as an ornamentation. Among the elements contributing to the enrichment of the melody are microtonality (the use of intervals smaller than a Western half step or lying between a half step and a Western whole step) and the variety of intervals used. Thus the three-quarter tone, introduced into Islāmic music in the 9th or 10th century, exists alongside larger and smaller intervals. Musicians show a keen sensibility to nuances of pitch, often slightly varying even the perfect consonances, the fourth and fifth.

As the fourth is the basic melodic frame, theorists organized the intervals and their nuances into genres, or small units, often tetrachords (units the highest and lowest notes of which are a fourth apart), combining genres into larger units, or systems. More than 130 systems resulted; on these are based the musical scales of the *maqāmāt,* or modes. The scale of a *maqām* can thus be broken down into small units that are of importance in the formation of melodies. A *maqām* is a complex musical entity given distinct musical character by its given scale, small units, range and compass, predominant notes, and preexisting typical melodic and rhythmic formulas. It serves the musician as rough material for his own composition. Each *maqām* has a proper name that may refer to a place (as Hejaz, Iraq), to a famous man, or to an object, feeling, quality, or special event. Emotional or philosophical meaning (ethos, or *taʾthīr*) and cosmological background are attached to a *maqām* and also to the rhythmic modes. The Arabic term *maqām* is the equivalent of *dastgāh* in Persia, *naghmah* in Egypt, and *cbāṭ* in North Africa.

Rhythmic organization. Rhythms and their organization into cycles of beats and pauses of varying lengths (rhythmic modes, or *īqāʿāt*) are much discussed in theoretical writings and are of supreme importance in performance. Each cycle consists of a fixed number of time units with a characteristic distribution of strong and weak beats and pauses. In performance some of the pauses may be filled in, but the underlying pattern must be maintained. Parallel to the growth of the number of melodic modes—from 12 in the 8th century to more than 100 in the 20th—is the increase in the number of rhythmic modes from eight in the 9th century to more than 100 in the 20th.

Musical forms. The repertoire in common use comprises a wide variety of forms. One category includes unmeasured improvised pieces, such as the *layālī,* in which the singer puts forth the characteristics of the *maqām,* using long vocalises and meaningless syllables. An equivalent instrumental improvisation is called *taqsīm,* and this in some cases may be accompanied by a uniform pulsation, called *taqsīm ʿala al-wuḥdah.* The category of metrical songs embraces various poetic forms and metric structures, such as *qaṣīdah, dor,* and *muwashshaḥ.* Both categories, metrical and unmeasured, are almost always accompanied by either one or more instruments to enrich the performance. Important traditional forms combined both categories to create large compositions similar to a suite, using vocal and instrumental features. The whole was linked by the unity of the mode and a defined rhythmical development. Examples are the Andalusian *nūbah,* which survives in North Africa, the Persian *dastgāh,* the Turkish *fāṣil,* the Egyptian *waṣla,* and the Iraqi *macam.* Under the pressure of modernization and westernization have emerged new forms showing the influence of light dance music, operetta, and musical comedy.

Instruments of music. Instrumental music is not considered an independent art from vocal music. Yet many instruments were fully described by early writers, and their use in folk, art, religious, and military music pointed out. The most favoured instrument of ancient Near Eastern civilization, the harp, was gradually overshadowed by both long- and short-necked lutes.

Percussion instruments. Among idiophones (instruments the hard bodies of which vibrate to produce sound) commonly used are the *qaḍīb* ("percussion stick"), the *zil* and *sunūj* ("cymbals"), and the *kāṣāt,* or small finger cymbals. Membranophones, or vibrating membrane instruments, include a variety of tambourines, or frame drums, which all fall under the generic name *duff.* These include the North African *ghirbāl* and *bendīr,* instruments that have a number of "snares" across the skin and are used for folk dances; and the *dāʾirah,* or *ṭar,* with jingling plates or rings set in the frame. The *dāʾirah* and the vase-shaped drum *darabukka* (in Iran, *ẓarb*) are used in folk and art music, and the small kettledrums *naqqārah* and *nuqayrat* are used in art music and in military music (such as janissary music, the Turkish ensemble adopted by European military musicians). The large two-headed cylindrical drum, the *ṭabl* (Turkish *davul*), is generally played with the oboe-like *zornā* or *gayta* in processions and open-air ceremonies.

Wind instruments. Classed with the *zornā* and *gayta* as aerophones, or wind instruments, are the *būq,* or horn, the *nafīr,* or long trumpet, and a variety of flutes called *nāy* or *shabbābah.* Clarinetlike (single-reed) double-piped instruments such as the *dunay, zammārah,* and *urghūl* are used in folk events and open-air ceremonies.

Stringed instruments. Chordophones, or stringed instruments, constitute the most important family. The favourite instrument of Islāmic classical music is the *ʿūd,* a short-necked lute having four or five strings and resembling the Western lute, which derived from the *ʿūd.* In addition to holding musical supremacy, it was important in medieval theoretical and cosmological speculations. It has two derivatives in North Africa, the *kuwītra* and the *gunbrī.* The long-necked lutes favoured in Turkey, Iran, and the countries eastward include the *ṭunbūr, tār,* and *setār.* Another plucked instrument is the *qānūn,* or trapezoid-shaped psaltery, played at least from early medieval times. The trapezoidal dulcimer, or *sanṭūr,* the strings of which are struck with two thin sticks, is widespread and is especially prominent in Persian art music. Bowed lutes, or fiddles, include the *rabāb,* used by epic singers and beggars, and the *kamān,* or *kamanjā,* a hemispherically-shaped fiddle the body of which, like that of the *rabāb,* is pierced by the length of wood forming the neck (such instruments are known as spike fiddles). The violin, played either on the knee like the *kamanjā,* or beneath the collarbone, is also common.

Medieval contacts with Europe

The relation of Islāmic music to music of other cultures. The relation of Islāmic music to the West reveals itself in both musical theory and practice. By the 9th century many Greek treatises had been translated into Arabic. Arabic culture preserved Greek musical writings, and most of those that reached the West did so in their Arabic versions. Arab theorists followed Greek models, often developing them further. The Muslim occupation of Spain and Portugal and the Crusades to the Near East brought Europeans in contact with Arabic theoretical writings and the flourishing Islāmic art music. Musical instruments such as the lute, the rebec (a small bowed instrument derived from the *rabāb*), and the kettledrum (in the form of a pair of small kettledrums called nakers, from the Arabic *naqqārah*) became firmly established in European music. Arabic writings were translated, among them the *De scientiis,* a work on the arts and sciences by the great 10th-century philosopher and musician al-Fārābī (Latinized as Alpharabius). Such translations give further indication of the influence exerted by Muslim writers. Arabian influence on European medieval music is difficult to prove. Borrowed elements were possibly completely transformed. The influence of Islāmic music on European music is, at present, a subject of controversy.

As early as 711, Arab conquerors reached India, and Mongol and Turkmen armies later invaded the Near East, with resulting contact between Islāmic and Far Eastern music. There are similarities between the modal systems of India (the *rāga*s) and of the Near East (the *maqām* system) and between some cosmological and ethical conceptions of music. The migration of musical instruments from the Islāmic area to the Far East can also be traced. The Chinese oboe, the *sona,* apparently derived its name from its Near Eastern counterpart, the *zornā,* or *sornā.* The Indian long-necked lute sitar, having a different number of strings from the Persian *setār,* received its name, and perhaps part of its form, from the *setār.* The Chinese dulcimer, *yang ch'in* ("foreign zither"), originated in the Middle Eastern *sanṭūr.* On the other hand, the musical instruments appearing in the pre-Islāmic Ṭāq-e Bostān reliefs in Persia show a mouth organ similar to the Chinese *sheng,* indigenous to the Far East.

THE HISTORY OF ISLĀMIC MUSIC

The earliest extant writings on Islāmic music are from the end of the 9th century, more than 250 years after the advent of Islām. In the absence of historical documents, musicians, writers, and philosophers began to speculate on the origins of their music. They filled the gaps by legendary sources or vague traditions. Thus Lamak is said to have made the first lute from the leg of his dead son, whose loss he lamented with it. His lamentation is considered to be the first song.

Courtly music

The pre-Islāmic period. In nomadic encampments music emphasized every event in man's life, embellished social meetings, incited the warriors, encouraged the desert traveler, and exhorted the pilgrims to the black stone of the Kaʿbah (in Mecca), a holy shrine even in pre-Islāmic times. Among the earliest songs were the *ḥudāʾ* from which the *ghināʾ* derived, the *naṣb, sanad, rukbānī,* and the *hazāj,* a dancing song. In the markets of the Arabs, particularly the fair at the western Arabian town of ʿUkāẓ, competitions of poetry and musical performances were held periodically, attracting the most distinguished poet-musicians. Their music, more sophisticated than that practiced in the nomadic encampments, was related to that of the *qaynāt* ("singing girls"), who performed at court, in noble households, and in scattered taverns. Cultural contact with Byzantium was strong in the kingdom of Ghassān, where, in the 7th century, five Byzantine *qaynāt* were known to have performed songs of their homeland at court. The culture of the other Arabic kingdom of al-Ḥīrah under the Lakhmid dynasty was closely connected with that of Persia under the pre-Islāmic Sāsānian empire. The Sāsānians esteemed both secular and religious music. In the belief of the Mazdak sect (a dualistic Persian religion related to Manichaeanism, a Gnostic religion), music was considered as one of the four spiritual powers. In the king's entourage musicians occupied high rank. Some became famous, such as Bārbad, to whom is attributed the invention of the complicated pre-Islāmic system of modes. The compositions of Bārbad, who became a model of artistic achievement in Arabic literature, survived at least until the 10th century.

Sophisticated secular music

The beginning of Islām and the first four caliphs. Muḥammad was said to have been hostile to music and musicians; yet there are indications that he tolerated functional music such as war songs, pilgrimage chants, and public or private festival songs. In addition, he himself instituted in 622 or 623 the *adhān* ("call to prayer"), chanted by the *muʾadhdhin* (muezzin). For this task he chose the Abyssinian singer Bilāl, who became the patron of the *muʾadhdhin* and their guilds throughout the Islāmic world. Within 12 years after Muḥammad's death, the armies of Islām took possession of Syria, Iraq, Persia, Armenia, Egypt, and Cyrenaica (in modern Libya). The contact with the refined cultures of the conquered and the appearance of a new class of warriors who benefited from the spoils of the conquered nations deeply affected Arabian society. In spite of the austere regime of the four orthodox caliphs (632–660), joy of life and eagerness for pleasure dominated the two holy cities of Mecca and Medina. Wealthy men acquired slave musicians, who were often liberated and became the pillars of musical life. The wealthy competed with one another in the brilliance of the concerts held in their houses, and in sophisticated literary and musical salons, contests revealed and rewarded the best talents. In this milieu the great Islāmic musical tradition began to take shape, to be firmly established and codified in subsequent periods. A new generation of musicians was educated in the traditional manner and refined through constant hearing of the best music performed by the best masters. Through the contributions of the conquered "foreigners," and through intense emulation of their music, new techniques, improved instruments, and elaborated musical forms developed. Persian lute tuning was adopted for the lute (*ʿūd*), which became the classical instrument of the Arabs. Melodies and rhythms were regulated by a modal system that was later codified. Among the most famous female musicians was ʿAzza al-Maylāʾ, who excelled in *al-ghināʾ ar-raqīq,* or "gentle song." Her house was the most brilliant literary salon of Medina, and most of the famous musicians of the town came under her tutelage. Also famed were the female musician Jamīla, around whom clustered musicians, poets, and dignitaries; the male musician Ṭuways, who, attracted by the melodies sung by Persian slaves, imitated their style; and Ṣāʾib Khāthir, the son of a Persian slave. Songs were generally accompanied by the lute (*ʿūd*), the frame drum (*duff*), or the percussion stick (*qaḍīb*).

Further influence of conquered peoples

The Umayyad and ʿAbbāsid dynasties: classical Islāmic music. Under the Umayyad caliphate (661–750) the classical style of Islāmic music developed further. The capital was moved to Damascus (in modern Syria) and the courts were thronged with male and female musicians, who formed a class apart. Many prominent musicians were Arab by birth or acculturation, but the alien element continued to play a predominant role in Islāmic music. The first and the greatest musician of the Umayyad era was Ibn Misjaḥ, often honoured as the father of Islāmic music. Born in Mecca of a Persian family, he was a musical theorist and a skilled singer and lute player. Ibn Misjaḥ traveled to Syria and Persia, learning the theory and practice of Byzantine and Persian music and incorporating much of his acquired knowledge into the Arabian art song. Although he adopted new elements such as foreign musical modes, he rejected other musical traits as unsuitable to Arabian music. Knowledge of his contributions is contained in the most important source of information about music and musical life in the first three centuries of Islām. This is the 10th-century *Kitāb al-Aghānī,* or "Book of Songs," by Abū al-Faraj al-Iṣbahānī. In the 8th century Yūnus al-Kātib, author of the first Arabic book of musical theory, compiled the first collection of songs. Other notable musicians of the period were Ibn Muḥriz, of Persian ancestry; Ibn Surayj, son of a Persian slave and noted for his elegies and improvisations (*murtajal*); his pupil al-Gharīḍ, born of a Berber family; and the Negro Maʿbad. Like Ibn Surayj, Maʿbad cultivated a special personal style adopted by following generations of singers.

The Golden Age

By the end of the Umayyad period, the disparate elements of conqueror and conquered were fused into the style of classical Islāmic music. With the establishment of the ʿAbbāsid caliphate in 750, Baghdad (in modern Iraq) became the leading musical centre. The ʿAbbāsid caliphate is the period of the Golden Age in Islāmic music. Music, obligatory for every learned man, was dealt with in varied aspects—among them virtuosity, aesthetic theory, ethical and therapeutic goals, mystical experience, and mathematical speculation. The artist was required to possess technical proficiency, creative power, and almost encyclopaedic knowledge. Among the finest artists of the period were Ibrāhīm al-Mawṣilī and his son Isḥāq. Members of a noble Persian family, they were chief court musicians and close companions of the caliphs Hārūn ar-Rashīd and al-Maʾmūn.

Isḥāq, a singer, composer, and virtuoso lutenist, was the outstanding musician of his time. A man of wide culture, he is credited with authorship of nearly 40 works on music, which were subsequently lost. According to the "Book of Songs," he is the originator of the earliest Islāmic theory of melodic modes. Called *aṣbiʿ* ("fingers"), it structured the modes according to the frets of the lute and the fingers corresponding to them. Indications above each song in the "Book of Songs" show the mode, the type of third (major, minor, or neutral), and often the rhythmic mode. (The third is the interval encompassing three notes of the scale. It can vary considerably in exact size without losing its character. Western music uses the major and the minor third; much non-Western and folk music also uses a neutral third, between the major and minor in size.) The neutral third, introduced into Islāmic music about this time, increased the number of melodic modes from eight to 12 by making more intervals available from which to build melodies. At this time the number of rhythmic modes varied from six to eight, their actual structure and content differing from author to author.

Isḥāq and Ibrāhīm al-Mawṣilī actively participated in the contemporary controversy between modernism, a Persian romantic style tending toward exuberance of embellishments, and Arabian classicism, characterized by simplicity and artistic severity. The Mawṣilīs represented the older classical tradition; the proponents of modernism were Ibn Jāmiʿ and the celebrated singer Prince Ibrāhīm ibn al-Mahdī.

Theoretical writings

In the second half of the 8th century, the extensive Islāmic literature of music theory began to flourish. Greek treatises were translated into Arabic, and scholars, who were acquainted with the Greek writings, began to devote books or sections of books to the theory of music. In their works they expanded, changed, improved, or shed new light on Greek musical theory. The well-known philosopher al-Kindī, who was deeply immersed in Greek learning, wrote more than 13 musical treatises, including the earliest Arabic musical treatise that is known to have survived. He also dealt with the theory of ethos (*ta'thīr*) and with cosmological aspects of music. Members of the Ikhwān aṣ-Ṣafā, an important 10th-century brotherhood, dealt also with these two themes and advanced a theory of sound that went well beyond ancient Greek theories. Philosophers such as al-Fārābī, author of the monumental *Kitāb al-musīqī al-Kabīr* ("Grand Book on Music"), and Ibn Sīnā (known in Europe as Avicenna) dealt with such topics as the theory of sound, intervals, genres and systems, composition, rhythm, and instruments, as did others such as as-Sarakhsī, his contemporary Thābit ibn Qurrah, and Avicenna's pupil Ibn Zaylā. The last important theorist to emerge during the ʿAbbāsid period was Ṣafī ad-Dīn, who codified the elements of the modal practice as it was then known into a highly sophisticated system. His achievement became the chief model for subsequent generations. In the numerous treatises written between the 13th and 19th centuries, the system devised by Ṣafī ad-Dīn was split into multiple local traditions.

Islāmic music in Spain. Parallel to the flourishing of music at the eastern centres of Damascus and Baghdad, another important musical centre developed in Spain, first under the survivors of the Umayyad rulers and later under the Berber Almoravids (rulers of North Africa and Spain in the 11th and 12th centuries) and Almohads, who expanded into Spain after the fall of the Almoravids. In Spain, encounter with different cultures stimulated the development of the Andalusian, or Moorish, branch of Islāmic music. The most imposing figure in this development is Ziryāb (fl. 9th century), a pupil of Isḥāq al-Mawṣilī, who, because of the jealousy of his teacher, emigrated from Baghdad to Spain. A virtuoso singer and the leading musician at the court of Córdoba, Ziryāb introduced a fifth string to the lute, devised a number of new forms of composition, and developed a variety of new methods of teaching singing in his well-known school of music. Musical activity spread to large towns, and Seville became a leading centre of musical-instrument manufacture.

Poetic and musical forms

New poetic forms were developed, such as the *muwashshaḥ* and the *zajal,* that were freer in rhyme and metre than the classical *qaṣīdah* or formal ode. These innovations in prosody opened the way to further musical developments. Especially important was the *nawbah* ("suite"), a form that included songs and instrumental music, free or metrical, that were linked together by melodic mode and rhythmic patterns. The 24 traditional *nawbah*s were invested with symbolic and cosmological significance. After the expulsion of the Muslims from Spain in 1492 this musical tradition was transported to North African centres, where it partially survived.

After the Mongol invasion of Baghdad in 1258 and the Spanish reconquest of Granada in 1492, the magnificence of Islāmic culture gradually waned. Music continued to be cultivated, receiving new influences from Mongol and Turkmen conquerors. Persia enjoyed artistic independence for about 450 years, until 1918; but during this period a huge area, from the Balkans to Tunisia, was submitted to a strong Turkish influence, which itself was heavily influenced by Arab and Persian music.

The modern period. From the beginning of the 19th century, Islāmic music was affected by the intensification of contacts and relationships with Western music. For the first time Islāmic music existed in juxtaposition with Western music. For example, European composers and musicians were summoned to create military bands and conservatories in Turkey (1826) and in Persia (1856), and Giuseppe Verdi's opera *Aida* inaugurated the opera house in Cairo in 1871. Expanding contact with Western music caused certain alterations in traditional musical styles. There was a widespread musical renaissance, with two main centres: the leading school in Egypt was open to modernism and Western influences, while in Syria and Iraq traditional music was supported. Music in Syria and Iraq, together with North African, Iranian, and Turkish music, remained restricted to its own periphery. The Egyptian school developed Middle Eastern music in what can be called the mainstream style; and this music was widely diffused through the media of radio, television, recordings, and the cinema. Mainstream music borrowed instruments such as the cello, saxophone, and accordion; melodies and rhythms from European serious and light music; the concept of large ensembles; and the use of electronic amplification. Emphasis shifted from the display of individual virtuosity and personal creativity to performance as an ensemble, and the use of short songs underscored the separation, rather than the traditional union, of composer and performer. Classical and local genres coexist, however, with the innovative mainstream style.

Persian art music continues to be organized into 12 traditional modes, or *dastgāh,* each of which contains a repertory of from 20 to 50 small pieces called *gūsheh*s ("corners"). In performance of instrumental and vocal music, the artist improvises on the chosen *gūsheh*s of a *dastgāh* in a specific order.

Vocal music still predominates even in countries such as Iran, in which instrumental music is cultivated independently. Thus almost all of the Near Eastern musicians who are well known are singers; those particularly influential in the modern renaissance, in chronological order, include ʿAbduh al-Ḥamūlī, Dāhūd Ḥussnī, Sayyid Darwīsh, ʿAbd al-Wahhāb, Umm Kulthūm, Farid al-Aṭrash, Fayrouz, Rashid al-Hundarashi, Ṣadīqa al-Mulāya, and Muḥammad al-Gubanshi.

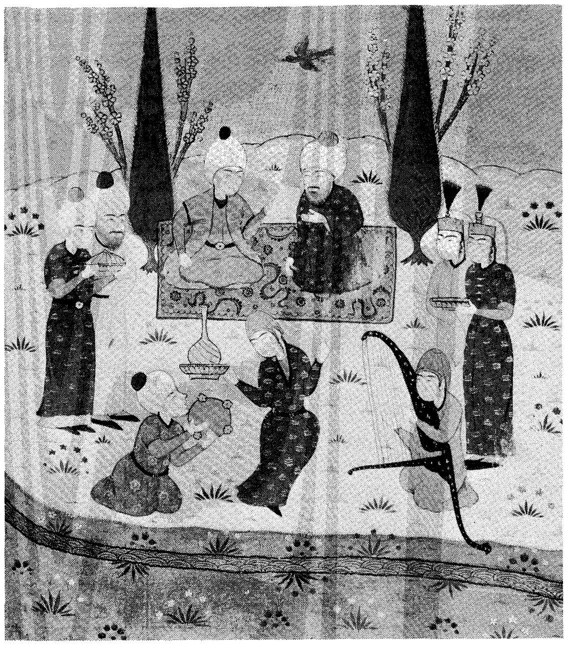
Theoretical developments

Modern Arab theorists also have produced valuable treatises. For example, the 19th-century theorists Michel Muchaqa of Damascus and Mohammed Chehab ad-Dīn of Cairo introduced the theoretical division of the scale into 24 quarter tones. In 1932 the international Congress of Arabian Music was held in Cairo, providing a forum for current analysis of subjects such as musical scales, modes, rhythms, and musical forms. (A.Sh.)

Dance and theatre

The performing arts have received comparatively little attention in the otherwise rich literature of the Islāmic peoples. This is most probably a result of the suspicions entertained by some orthodox Muslim scholars concerning the propriety of the dance and the theatre. Because this applies particularly in relation to the vexing theological question of human portrayal and its connection with idolatry, the performing arts have traditionally been regarded by the faithful with more than usual caution. Even as late as the 19th and early 20th centuries, most research on the subject, in what may loosely be called the Islāmic world, was carried out by Western scholars, chiefly from European nations; and only in the 20th century have indigenous scholars started publishing significant research on the subject.

There are no known references to the dance or theatre in pre-Islāmic Arabia, although nomad tribes were probably acquainted with the dance. The Islāmic peoples themselves seem to have developed this particular art form less than they did music or architecture; and, in addition to medieval Islām's cool attitude toward dance and theatre as art forms, it must be added that most women, leading a life of seclusion, could hardly be expected to play an active part in them. Nevertheless, there has been an active tradition of folk dance in most Islāmic countries, in addition to dancing as an entertainment spectacle and, particularly in Persia, as an art form. A ritual dance was instituted in



Dance as entertainment for the aristocracy, shown in "A Festive Party," manuscript illumination from the *Masnavī* of Jalāl ad-Dīn ar-Rūmī, AD 1295–96. In the British Museum (MS. OR. 7693, fol. 225 b.).

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

the Ṣūfī mystical order of the Mawlawīyah (Mevleviyah) in Turkey. The dance, performed by dervishes (members of the mystical order), is considered to be a manifestation of mystical ecstasy rather than an entertainment or an expression of aesthetic urges.

The theatre has not flourished as a major art under Islām, although as a form of popular entertainment, particularly in mime and shadow-puppet shows, it has persisted vigorously. Nevertheless, the theatre with live actors received support from the Ottomans in Turkey, and a live popular drama has been strong in Persia, where a passion play also took root. Otherwise, the theatrical record of Islām is meagre. Moreover, few neighbouring peoples had a well-developed theatre of their own; hence, outside stimulus was lacking, and the Islāmic disapproval of idolatry was so intense that, when the shadow theatre evolved in the East in the late Middle Ages, the puppets were regularly punched with holes to show that they were lifeless. Nonetheless, drama has had some ties with religion, as in Iran and other areas where the Shīʿite branch of Islām is concentrated. Here a passion play developed, rooted in traumatic memories of the bloody warfare of Islām's early years. This was a local phenomenon, uninfluenced by Christian Europe, and, though stereotyped, it movingly reenacted Shīʿite martyrdom.

Popular theatre

A popular theatre, frequently including dance, evolved independently from about the 17th century in some Muslim countries. West European and, later, U.S. influences were largely the main factors in the development of an artistic theatre in the 19th and 20th centuries. But conservative Muslims have consistently disapproved of theatre, and in Saudi Arabia, for example, no native theatrical establishment exists. In such an atmosphere, women's parts were at first taken by men; later, Christian and Jewish women took the roles, and only in the 20th century have Muslim women participated.

TYPES AND SOCIAL FUNCTIONS OF DANCE AND THEATRE

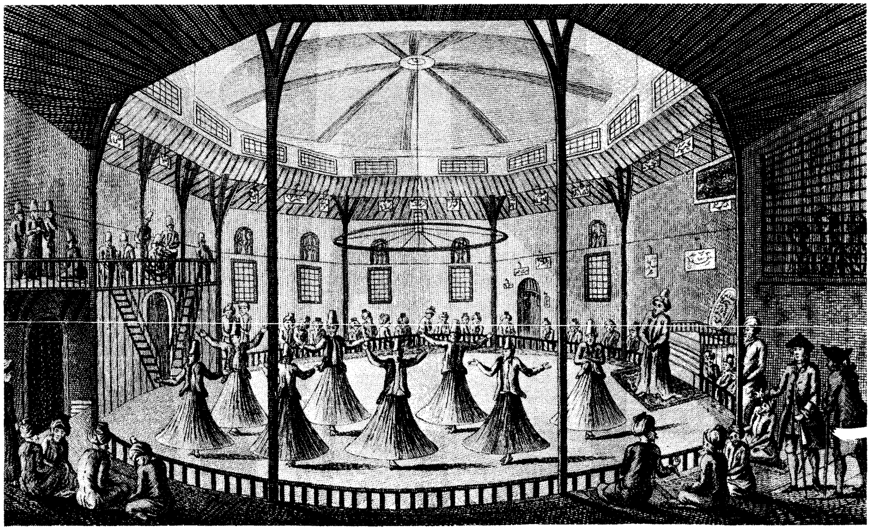
The dance. Folk dancing existed among medieval Islāmic peoples; but such sources as exist are mainly concerned with artistic dance, which was performed chiefly at the caliph's palace by skilled women. The aristocracy was quick to imitate this patronage by providing similar performances, its members vying with one another on festive occasions. One of these dances, the *kurrağ* (some times called *kurra*), developed into a song and dance festival held at the caliph's court. Since the latter part of the 19th century, the dancing profession has lost ground to the performance of U.S., Latin-American, and western European dances in cabarets. In a reaction that set in after World War II, fervent nationalists have tried to create native dance troupes, revive traditional motifs in costume and interpretation, and adapt tribal figures to modern settings. Few traditional dances have survived unchanged; among those that have are the dervish dances, performed mainly in Turkey.

Folk dance. Though now performed and fostered chiefly as an expression of national culture, folk dances were long regarded as pure entertainment and were either combined with theatrical shows or presented alone. Dance performances, accompanied by music, took place in a special hall or outdoors; many dancers, particularly the males, were also mimes. Sometimes the dance enacted a pantomime, as in Turkey, of physical love or of a stag hunt, representing the pursuit of a suspicious husband deceived by his wife.

Folk dance, except in Iran, has almost always been mimetic or narrative, a tradition still fostered by many tribes.

Dance as entertainment. The Turks considered dancing as a profession for the low-born; as a result, most dancers

By courtesy of the University of Chicago Library



Dervishes dancing in a *tekke,* engraving by J. Fougeron from *A Tour to the East in the Years 1763 and 1764, with Remarks on the City of Constantinople and the Turks,* by Frederick Calvert, 6th Baron Baltimore, 1767.

were members of minority groups: mostly Greeks, Jews, and Armenians. This judgment has usually applied to the status of professional dancers and indeed to most professional entertainers at most periods and in most societies until modern times. In 19th-century Egypt, both male and female dancers were regarded as public entertainers. Many of the women entertainers (*ghawāzī*) belonged to a single tribe and were usually considered little better than prostitutes. The erotic element in dancing has become focused in the belly dance, which has become the leading form of exhibition dance in modern Turkey and the Arab countries.

The mimetic tradition of folk dance has blended well in countries of the Sunnite persuasion with comedy and with the passion-play tragedy in Shīʿite countries. In recent years, however, the theatre has been increasingly divorced from the dance, with most plays being consciously modeled on European patterns; only in the operetta does the old combination remain.

Dance as an art form. In pre-Islāmic times in Iran, dance was both an art form and a popular entertainment. There are pictures of dancers in miniatures, on pottery, and on walls, friezes, and coins. Some of these ancient dances lived on partially in tribal dances, but again, under Islām's restrictions on women, the art became a male monopoly. Women were permitted to dance in private, however, as in the harem.

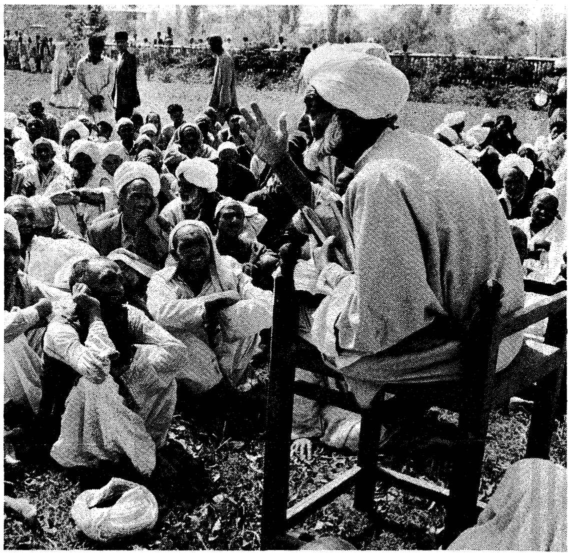
Iran is perhaps the only Muslim country with a tradition of dance regarded as an art form. When revived after World War II, folk dancing was encouraged and adapted for the foundation of a national ballet.

Muslim orthodoxy's very uncertainty over the exact status of the artistic dance ensured that it was always considered as an adjunct to music. Accordingly, although there are many detailed treatises on Islāmic music, none is available on dance.

Dervish dancing. There is one outstanding example of pure dance: that of the whirling dervishes, an art that has been practiced for more than seven centuries. The procedure is part of a Muslim ceremony called the *dhikr,* the purpose of which is to glorify God and seek spiritual perfection. Not all dervish orders dance; some simply stand on one foot and move the other foot to music. Those who dance, or rather, whirl, are the Mawlawī dervishes, an order that was founded by the Persian poet and mystic Jalāl ad-Dīn ar-Rūmī at Konya, in Anatolia, in the 13th century.

The performance, for which all of the participants don tall, brown, conical hats and black mantles, takes place in a large hall in the *tekke,* the building in which the dervishes live. The dervishes sit in a circle listening to music. Then, rising slowly, they move to greet the *shaykh,* or master, and cast off the black coat to emerge in white shirts and waistcoats. They keep their individual places with respect to one another and begin to revolve rhythmically. They throw back their heads and raise the palms of their right hands, keeping their left hands down, a symbol of giving and taking. The rhythm accelerates, and they whirl faster and faster. In this way they enter a trance in an attempt to lose their personal identities and to attain union with the Almighty. Later they may sit, pray, and begin all over again. The *dhikr* ceremony always ends with a prayer and a procession.

Brian Brake—Magnum



A storyteller in Kashmir.

The theatre. In lands where the Sunnite sect was strong, mime shows were frequent and popular attractions during the later Middle Ages. The Ottoman sultans were accompanied on military campaigns by their own troupe of actors; and, as the Ottoman Empire grew larger and richer, the court became ever more partial to entertainment, whether at the accession of a sultan, a royal wedding, a circumcision, an official visit, or a victory. On such occasions, dances and theatrical performances played their part along with parades, fireworks, music, mock fights, and circus performances in one huge, sumptuous pageant. This lavishing of entertainment reached a height of splendour that the admiring Ottoman aristocracy strove to imitate throughout the empire. In Arabia and North Africa, popular shows on a lesser scale were performed in the open air. Another aspect of the Islāmic theatre was represented in the shadow plays, which were given chiefly to while away the time during the month of fasting, Ramaḍān (the sacred ninth month of the Muslim year).

Actors at the Ottoman court

Among Shīʿites the passion play was regularly performed, by both professional and amateur actors. The performance always took place during the first 10 days of the month of Muḥarram (the first in the Muslim year), the period when the suffering and death of the descendants and relatives of the fourth caliph ʿAlī were commemorated. For generations this largely theatrical event served as a focal point of the year, gripping audiences in total involvement with its blend of symbolism and realism.

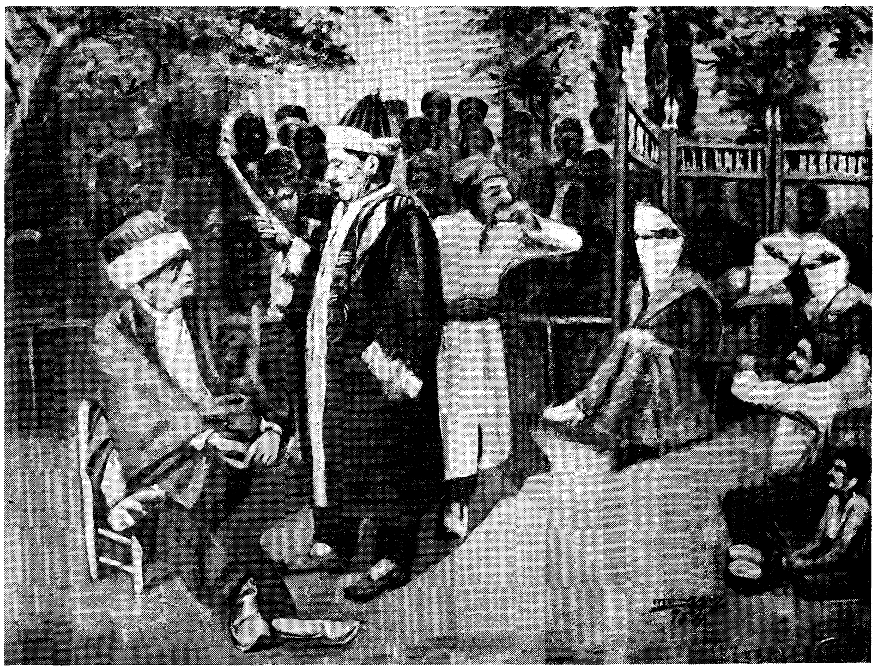
Mime shows. In the medieval Muslim theatre, mime shows aimed to entertain rather than to uplift their audiences. Regrettably, few mime shows were recorded in writing, and those that were recorded were set down primarily to serve as guidelines for directors, who might tamper with the wording, as in the improvisation of the Italian commedia dell'arte. Some plays were on historical themes, but preference was for comedies or farces with an erotic flavour. The audience was largely composed of the poor and uneducated.

A rudimentary theatrical form, the mime show long enjoyed widespread popularity in Anatolia and other parts of the Ottoman Empire. Called *meddah* (eulogist) or *mukallit* (imitator) in Turkish, the mimic had many similarities to his classical Greek forerunners. Basically he was a storyteller who used mimicry as a comic element, designed to appeal to his largely uneducated audience. By gesture and word he would imitate animals, birds, or local dialects; he was very popular in Arabic- and Turkish-speaking areas. Even today, he has not been wholly supplanted in the Islāmic world by literacy or by such modern entertainments as radio, television, and the cinema. Sometimes several *meddah*s performed together, and this may have been the source of a rural theatrical performance.

Ortaoyunu. The *ortaoyunu* (middle show) was the first type of genuine theatre the Turks, and possibly other Muslim peoples, ever had. The Ottoman sultans provided subsidies for *ortaoyunu* companies of actors, who consequently became generally accepted; also some were retained by the princes of the Romanian principalities under Ottoman rule. The fact that they continued to enjoy popularity to World War I may be explained by their simple dramatic appeal, which was coupled with sharp satire of the well-to-do and the ruling classes (but hardly ever of Islām). This irreverence frequently resulted in fines and imprisonment for the actors, but it never produced a basic change of style.

The first real theatre

During the 19th and 20th centuries, the *ortaoyunu* was generally performed in an open square or a large coffeehouse. There was no stage, and props were simple: they generally comprised a table or movable screen, while other



Ortaoyunu theatre, painting by Muazzez. In the collection of Dr. Metin And.
Metin And

objects were represented by paintings glued on paper. An orchestra of about four musicians enlivened the show and gave the performers, who were all male, their cues. Roles were generally stereotyped, with stock characters, such as a dandy, the foreign physician, and regional types (Kurds, Albanians, Armenians, Arabs, and Jews) quarreling and fighting in slapstick style. Mimicry was important, and some actors changed roles and costumes. The plot was flimsy, a mere frame for the dialogue, which was itself frequently improvised.

The marionette theatre. In comparison with *ortaoyunu,* the marionette theatre, although popular in Turkistan (under the name of *çadir hayâl*) and other parts of Muslim Central Asia, never really caught on in the Ottoman Empire.

Shadow plays (Karagöz). On the other hand, the shadow play had been widely popular for many centuries in Turkish- or Arabic-speaking countries. Its essence, like that of the mime shows, was entertainment without moral import; and few plays were recorded in writing beyond a sketch of the action. Most were comedies and farces that were performed for the enjoyment of an audience that was, for the most part, very poor and uneducated.

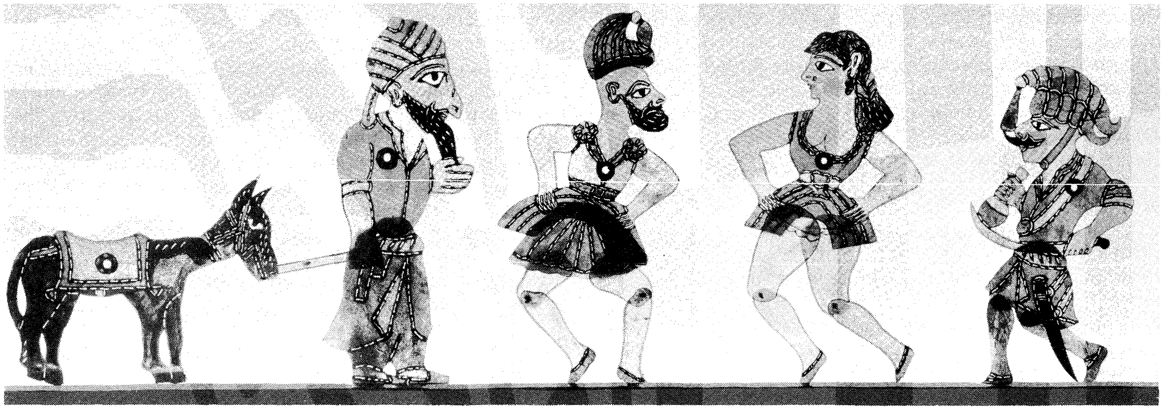
In Turkey, the Karagöz (a character, "Black-eye") theatre was the prevalent form of shadow play. This art apparently came from China or perhaps from Southeast Asia, as the French term *ombres chinoises* indeed hints, though the prevailing element of the grotesque was probably inherited from ancient Greece by way of Byzantium. The Karagöz was well known in Turkey during the 16th century but was so fully developed that it must have been introduced much earlier, and it quickly spread from Syria to North Africa and the Greek islands. Its performers were in great demand at the sultan's court as well as elsewhere, and they soon organized their own guild. Since only the framework of the play was sketched in writing, there was scope for a great deal of impromptu wit, and Karagöz shows, like the *ortaoyunu,* were inevitably satirical. But with the coming of motion pictures the Karagöz declined, and performances are now mostly confined to the month of Ramaḍān.

In the traditional performance of the Karagöz, the stage is separated from the audience by a frame holding a sheet; the latter has shrunk over the years from about six by 7½ feet (1.8 by 2.3 metres) to about three by two feet (0.9 by 0.6 metres). The puppets, which are flat and made of leather, are controlled by the puppeteers with rods and are placed behind the screen. An oil lamp is then placed still farther back so that it will throw the puppets' shadows onto the screen.

Traditional characters

A standard shadow play has three main elements: introduction, dialogue, and plot. The introduction is fairly stereotyped and consists of an argument and usually a quarrel between Karagöz and Hacivat, the two most com-

Marc Riboud—Magnum



Karagöz shadow puppets.
From left: Yahudi (the Jew) with donkey, Karagöz, Zenne (the woman), and Tasuz Deli Bekir.

mon characters. The former is a simple, commonsense fellow, while the latter is more formal and polished, if shallow and pedantic. The dialogue between the two varies with the occasion but always contains impromptu repartee, though most puppet masters have at least 28 different plots in stock—a different one for each night of Ramaḍān. Some are historical, many ribald, but all are popular entertainment. Additional characters or animals may be introduced, calling for great skill on the part of the puppet master and his assistant in manipulating several simultaneously, as well as in reciting the text in changing tones and playing music. Some have one or two musicians to help.

Mimicry and caricature, while essential to both the *meddah* and the *ortaoyunu,* are technically more developed in the shadow play. Here entire productions are based on a comedy of manners or of character. In addition to the stock characters from various ethnic groups, there is, for example, the drug addict who wraps his narcotic in dissolving gum before the fast begins so as not to sin, the light-headed Turk ("he who eats his inheritance") who is a prodigal and a debauchee, the highway robber, the stutterer, and the policeman.

Karagöz is the most frequently performed but not the sole type of shadow play in Muslim countries. In Egypt a shadow theatre is known to have existed as early as the 13th century, long before records of Karagöz shows were kept in Turkey. A physician, Muḥammad ibn Dāniyāl, wrote three shadow plays that have survived. They were performed in the 13th century and display humour and satire and the lampooning of match-making and marriage. These plays also introduce a parade of popular contemporary characters, many of whom earn their living in shady or amusing trades. A positively phallic element is as evident here as it is in the Karagöz.

The comic theatre in Iran

Iranian popular theatre. Popular theatre existed among the Iranians, who were proud of a long-lived cultural tradition and preserved their national language under Arab domination: indeed, even their branch of Islām, Shīʿism, set them apart from the Sunnism of the majority of Islām. The Ottomans' failure to conquer Iran increased competition between the respective intellectual elites. Iran had inherited a considerable theatrical tradition from pre-Islāmic times; it is not surprising that a popular comic theatre flourished there. The central figure of this theatre was the *Katchal Pahlavān* (or "bald actor"), and mimicry was important, both in comedy and in pantomime. The *Baqqal-Bāzī* ("Play of the Grocer"), in which a grocer repeatedly quarrels with his good-for-nothing servant, is a typical example of the popular comic tradition. The marionette theatre, or *Lobet-Bāzī,* while using Iranian puppets, was similar to its Turkish counterpart. At least five puppets appeared, and singing was an integral part of a production that sometimes resembled Italian and French puppet shows. The *ortaoyunu,* particularly in the region of Azerbaijan, is almost identical with the Turkish of the same name. The shadow play in Iran, however, has always been less popular and obscene than the Ottoman or Arab Karagöz.

Passion plays (taʿziyah). Quite different was the passion play, derived mainly from early Islāmic lore and assembled as a sequence of tragedies representing Shīʿite martyrdom. Both shadow and passion play were interlarded with musical prologues, accompaniment, and interludes; but these were not necessarily an integral part, serving rather to create a mood.

A preoccupation with religion is characteristic of Persian theatrical performances, and, during the first 10 days of the month of Muḥarram, the martyrdom of ʿAlī's descendants at the hands of the Umayyads is reenacted. Although these shows are also performed among Shīʿite Turks in Central Asia and Shīʿite Arab communities in Iraq and elsewhere, Iran is their centre. Some plays are satirical, directed against wrongdoers, but most form a set of tragedies, performed as passion plays on these 10 successive days. Named *taʿziyah* ("consolation"), this type of drama is an expression of Persian patriotism and, above all, of piety, both elements combining in an expression of the national religion, Shīʿism.

Background of the *taʿziyah*

In order to understand the mood of the *taʿziyah* it is necessary to remember that storytellers in Iran recite the gruesome details of the martyrdom of Ḥasan, Ḥusayn, and other descendants of ʿAlī all year long. Thus prepared, people swell the street processions during the days of Muḥarram, chain themselves, flagellate their bodies, and pierce their limbs with needles, shouting in unison and carrying images of the martyrs, made of straw and covered with blood—contrary to the injunctions of Islām. Sometimes men walk in the processions with heads hidden and collars bloodied, all part of a pageant dating from the 9th or 10th century. Its peak is reached daily in the play describing the martyrdom of ʿAlī's family and entourage, which used to be presented in the large mosques, but which, when the mosques proved too small, was given a special place. The roles of reciter of the martyrdom and of participant in a procession have blended over the years to produce the *taʿziyah* play, in which the reciters march in procession to the appointed place and there recite their pieces, which can be considered as a prologue before the play itself begins.

The chief incidents narrated in the *taʿziyah* are not necessarily presented in chronological order, but in any case the *taʿziyah* texts (manuscripts from the 17th and 18th centuries, thenceforth, printed texts) give an inadequate impression of their forceful effect. Indeed, the audience identifies itself so closely with the play that foreigners have, on occasion, been manhandled. Since half of the actors play the supporters of the ʿAlids and half play their opponents, the latter are sometimes attacked and beaten at the end of the play. The decor, too, is half-realistic and half-symbolic: blood is real, yet sand is represented by straw. The stage effects are frequently overdone, and this clearly further excites the audience. For instance, Ḥusayn's gory head is made to recite holy verses; or an armless warrior is seen to kill his opponent with a sword he holds in his teeth. The horses are real, although most of the other animals are played by humans. In general, the actors, though chiefly nonprofessional, infect the audience with their enthusiasm and absorption.

DANCE AND THEATRE IN MODERN TIMES

Developments in dance. Insofar as dance is related to the modern theatre, there is little difference between Muslim production and its European or American counterpart. Dance and drama are combined according to the artistic needs of the production or the personal tastes of the producer and director. Perhaps more important is the dance itself, independently performed as artistic self-expression. The geographical centre of folk dance is in the area east of the Mediterranean, though remnants of other cultures have survived. There are Balkan traces in western and northern Turkey, for example, and Berber and even black African traces in Morocco and elsewhere in North Africa.

Arab countries. In some Arab countries, dancing is popular, varying in town, village, or with nomad tribe. In the town, dancing is generally reserved for special occasions, chiefly Western social dances. On the other hand, villages have such favourites as the *dabkah.* The *dabkah* is danced mainly by men and is quite common in festivities in the area between northern Syria and southern Israel; for instance, the Druzes (sectarian Arab communities located in Lebanon, Syria, and Israel) are very fond of it. The performers dance in a straight line, holding handkerchiefs high in the air, while the first dancer in the row gives the sign for stepping or jumping. Among the Bedouin almost any pretext suffices for dancing, although since the mid-20th century dancing has been practiced most often at weddings and similar festivities. Usually two male dancers, or two rows of male dancers, repeatedly advance toward each other or the audience and retire. To this basic figure, there are numerous variations that give the different dances their names.

Turkey. The Turks are also lovers of music and dance and when they meet frequently sing and dance. There is no single national dance popular throughout the country; dances vary in the numbers required, some being for solo performance, others designed for pairs or groups, though nearly all have instrumental accompaniment. As illustration of the possibilities of a basic step, there are at least 40 variations of the group dance known as *bar,* a chain dance. Again, several folk dances have characteristics akin to pantomime, breaking up into five main types of imitation: village life, nature, combat, courtship, and animals or birds.

Opera is popular in Turkey, reflected in a long tradition of invitations to foreign companies, and the musical theatre, which frequently includes dancing, is also widespread. On the other hand, classical ballet was unknown until a school of ballet was opened by foreign teachers with government encouragement. Although most of the ballet performances are in Istanbul, they are well received on tour.

Iranian ballet

Iran. In Iran a national dance company was formed with government support after World War II, and ancient customs were revived. Until it was closed in 1979, the Iranian ballet company was outstanding in the Muslim world, drawing on ancient war dances, fire-priest dances, dervish dances, and tribal folklore, as well as on scenes and decor from painting, sculpture, and the rich imagery of classical Persian poetry. Various folk dances are likewise performed all over Iran; they are accompanied by music and reflect local traditions and customs. Some are mimetic, others erotic, others, again, war dances (chiefly in the mountain areas) and comic dances (usually with masks). Many of these are dying out as new tastes and customs evolve, and Iranian dance companies have tried to preserve some of these dying forms.

The contemporary theatre. The modern Muslim theatre is almost wholly a western European importation, unconnected with the traditional medieval theatre, which has almost completely disappeared, although there are vestiges of it.

Arab countries. Contemporary Arabic theatre owes much to the imaginative daring of the Naqqāsh family in 19th-century Beirut, which was then under Turkish rule. Significantly, they were Christians, then better educated and more cosmopolitan than Muslims, and they had the advantages of Beirut's contacts with Europe and position as the headquarters of missionary activity. A Beirut Maronite (a Roman Catholic following the Syrio-Antiochene rite, widespread in the area), Mārūn an-Naqqāsh (died 1855), who knew French and Italian as well as Arabic and Turkish, adapted Molière's *L'Avare* ("The Miser") and presented it on a makeshift stage in Beirut in 1848. He did so before a select audience of foreign dignitaries and local notables, and he wrote his play in colloquial Arabic and revised the plot to suit the taste and views of his audience. Further, he changed the locale to an Arab town and Arabicized the names of the participants. Other touches included instrumental and vocal music and the playing of women's roles by men, in the traditional manner. The above features characterized the Arabic theatre for about half a century. An-Naqqāsh, together with his family, composed and presented two other musical plays, one based on Molière's *Tartuffe,* the other on the story, in *The Thousand and One Nights,* of Abū al-Ḥasan, who became caliph for a day.

Egyptian theatre

Soon the main centre of Arabic theatre moved to Egypt, whose comparatively tolerant autonomy offered an atmosphere for literary and artistic creativity more congenial than other parts of the Ottoman Empire. Syrian and Lebanese intellectuals and actors emigrated there, particularly after the anti-Christian riots of 1860 in Syria. Though a somewhat crippled Arabic theatre continued in Syria, its influence was carried into Egypt by émigrés and later spread to other Arabic-speaking regions. The number of theatres, a potentially large public, the munificence of Egypt's rulers, increasing prosperity under British rule after 1882, and increasing education soon made Egypt the centre of Arabic theatre, a position it has successfully maintained since.

The colloquial Arabic of Egypt was increasingly employed in the theatre, and several companies toured the country and neighbouring parts. The composition of these companies was fluid, for the actors were prone to be fickle in their loyalties; nevertheless, certain types of Egyptian theatre can be discerned in the late 19th century and during the early 20th. Some, like the company of Salāmah Ḥijāzī, used music to such an extent that their productions approached being labelled opera or operetta. Others, like that of ʿAlī al-Kassār, specialized in downright farce, expressed in revue form, with a Nubian hero, the "Barbarin," who made a specialty of ridicule and mimicry. Yet others, like the company of Najīb ar-Rīḥānī, oscillating between outright farce and comedy, skillfully depicted contemporary Egyptian manners; in particular, Najīb ar-Rīḥānī created a character called Kish-Kish Bey, whose misadventures and unsolicited advice on every subject have made him a classic creation. A conventional theatre sprang up in Egypt, too, catering to a growing number of intellectuals and presenting dramas and tragedies in polished, literary Arabic. Its chief exponent was Jūrj Abyaḍ, who had spent time studying acting in Paris. In contrast, Yūsuf Wahbī's National Troupe performed realistic plays, usually dramas or melodramas, using either colloquial or literary Arabic and sometimes a combination of both.

The plays performed by the Egyptian troupes and others in Arabic-speaking lands developed through three overlapping but distinguishable stages: adaptations, translations, and original plays. Adaptations came first in the 19th century (see above). Translations of established works appealed to a discriminating public, but original plays, part of the evolution of modern Arabic literature, reflected a growing interest in political and social problems. The decline of foreign influence and the arrival of political independence encouraged creativity, which, however much under European influence, has some original works to its credit. Two 20th-century Arabic playwrights, both Egyptian, are Tawfīq al-Ḥakīm, a sensitive shaper of both social and symbolic dramas, and Maḥmūd Taymūr, a novelist and comedy writer who strikes deep into Egypt's social problems.

Turkey. The development of the modern Turkish theatre strongly resembles its Arabic counterpart. In Istanbul, theatrical performances were not unusual among the diplomatic and international set, and some local Turks were acquainted with them. Nonetheless, Turkish plays for live actors—barring *ortaoyunu*—date only from 1839. The first Turkish playhouse was built in Pera (now Beyoğlu), significantly in the middle of the foreign and embassy quarter of Istanbul. Many of the actors were members of non-Muslim minorities, such as the Armenian; and the first plays presented in Turkish were adaptations from the French, chiefly Molière. They were done during the 1840s, when music was an important item.

The Gedik Paşa Theatre

The Gedik Paşa Theatre, which is named for the area in Istanbul where it was located, was the first theatre in which Turkish plays were produced by native actors speaking in Turkish. The actors received a salary, and local writers presented their own plays. Originally built for foreign companies, the theatre was reconstructed in 1867 and reopened in 1868 for a Turkish company headed by an Armenian, Agop, who was later converted to Islām and changed his name to Yakup. For almost 20 years the Gedik Paşa Theatre was the dramatic centre of the city; and plays in translation were soon followed by original plays, several with a nationalist appeal, such as Namık Kemal's *Vatan yahut Silistre* (*Fatherland*), which was first produced in 1873. The actors had to struggle against prejudice and the playwrights against censorship (some of them were imprisoned or exiled), but the Turkish theatre spread beyond Istanbul in the 1870s and 1880s to such places as Adana (in southern Anatolia) and Bursa (just south of Istanbul, across the Sea of Marmara).

After the Young Turk Revolution of 1908, censorship was not relaxed, but interest in the theatre grew, particularly over political matters; and plays about the new constitution were written and performed. After the foundation of the Turkish Republic in 1923, the state subsidized several theatre companies and a school for dramatic arts, and an opera house was built in Ankara. Official support not only gave financial encouragement but also implied a change of attitude over such matters as the participation of Muslim women in productions.

By the middle of the 20th century, theatrical life was mostly centred on Istanbul and Ankara, although theatres and companies continued in the small towns too.

A growing number of original plays, some of which were influenced by American literature, have been written and produced; the standard has been higher than it was before World War I, when Turkish poetry and fiction were rather more impressive than the drama. Subjects, too, have been more diverse since that time. To topics such as the position of women, marriage and divorce, and the character of Islāmic institutions—all popular under the Ottomans—have been added the Greco-Turkish War, education, village conditions, secularization, class struggle, and psychological problems. The Dormen Theatre was founded in Istanbul in 1955 by Haldun Dormen; in the 1971 World Theatre season in London the company performed *A Tale of Istanbul,* a comedy that included elements of folklore, a puppet show, singing, and a belly dance. The Dormen Theatre also produces 20th-century Western plays.

Western European and Russian influences in Iran

Iran. In Iran the birth of the modern theatre dates from the second half of the 19th century. Adaptations and translations from European plays appeared in Persian, often with the location and names suited to Iran. Molière, again, was a favourite and western European influence considerable, though Russian literature also left its mark, particularly in Azerbaijan, whose northern population had a chance to watch Russian actors during World War I.

Playwrights began to write original plays almost at once; one of the earliest playwrights was an Azerbaijani, named Akhundof, living in the Caucasus. He wrote seven comedies ridiculing Persian and Causasian Muslim society; all were translated into Persian and printed in 1874. Other plays likewise showed pronounced yearnings for social reform presented in a satirical style; some of these were published in a magazine called *Tyatr* ("Theatre"), which first appeared in 1908. Another type was the patriotic play, extolling Iran's history.

Some pre-World War I pieces were designed for reading rather than production. They were performed usually in schools, but there were hardly any professional actors, and the stage and props were very simple. After World War I, suitable halls were built in Tehrān and other cities, but the iron hand of Reza Shah (1925–41) curtailed development through continuous censorship and surveillance. After 1942 many new companies were formed, and there was speedy development, with growing interest in social and political subjects, though competition from foreign films was considerable. The revolutionary Islāmic regime established in 1979 severely curtailed theatrical activity.

(J.M.L.)

Visual arts

In order to answer whether or not there is an aesthetic, iconographic, or stylistic unity to the visually perceptible arts of Islāmic peoples, it is first essential to realize that no ethnic or geographical entity was Muslim from the beginning. There is no Islāmic art, therefore, in the way there is a Chinese art or a French art. Nor is it simply a period art, like Gothic art or Baroque art, for once a land or an ethnic entity became Muslim it remained Muslim, a small number of exceptions like Spain or Sicily notwithstanding. Political and social events transformed a number of lands with a variety of earlier histories into Muslim lands. But, since early Islām as such did not possess or propagate an art of its own, each area could continue, in fact often did continue, whatever modes of creativity it had acquired. It may then not be appropriate at all to talk about the visual arts of Islāmic peoples, and one should instead consider separately each of the areas that became Muslim: Spain, North Africa, Egypt, Syria, Mesopotamia, Iran, Anatolia, India. Such, in fact, has been the direction taken by some recent scholarship. Even though tainted at times with parochial nationalism, the approach has been useful in that it has focused attention on a number of permanent features in different regions of Islāmic lands that are older than and independent from the faith itself and from the political entity created by it. Iranian art, in particular, exhibits a number of features (certain themes such as the representation of birds or an epic tradition in painting) that owe little to its Islāmic character since the 7th century. Ottoman art shares a Mediterranean tradition of architectural conception with Italy rather than with the rest of the Muslim world.

Such examples can easily be multiplied, but it is probably wrong to overdo their importance. For if one looks at the art of Islāmic lands from a different perspective, a totally different picture emerges. The perspective is that of the lands that surround the Muslim world or of the times that preceded its formation. For even if there are ambiguous examples, most observers can recognize a flavour, a mood in Islāmic visual arts that is distinguishable from what is known in East Asia (China, Korea, and Japan) or in the Christian West. This mood or flavour has been called decorative, for it seems at first glance to emphasize an immense complexity of surface effects without apparent meanings attached to the visible motifs. But it has other characteristics as well; it is often colourful, both in architecture and in objects; it avoids representations of living things; it gives much prominence to the work of artisans and counts among its masterpieces not merely works of architecture or of painting but also the creations of weavers, potters, and metalworkers. The problem is whether these uniquenesses of Islāmic art, when compared to other artistic traditions, are the result of the nature of Islām or of some other factor or series of factors.

These preliminary remarks suggest at the very outset the main epistemological peculiarity of Islāmic art: it consists of a large number of quite disparate traditions that, when seen all together, appear distinguishable from what surrounded them and from what preceded them through a series of stylistic and thematic characteristics. The key question is how this was possible, but no answer can be given before the tradition itself has been properly defined.

Such a definition can only be provided in history, through an examination of the formation and development of the arts through the centuries. For a static sudden phenomenon is not being dealt with, but rather a slow building up of a visual language of forms with many dialects and with many changes. Whether or not these complexities of growth and development subsumed a common structure is the challenging question facing the historian of this artistic tradition. What makes the question particularly difficult to answer is that the study of Islāmic art is still so new. Many monuments are unpublished or at least insufficiently known, and only a handful of scientific excavations have investigated the physical setting of the culture and of its art. Much, therefore, remains tentative in the knowledge and appreciation of works of Islāmic art, and what follows is primarily an outline of what is known with a number of suggestions for further work into insufficiently investigated areas.

Media and techniques

Each artistic tradition has tended to develop its own favourite mediums and techniques. Some, of course, such as architecture, are automatic needs of every culture; and, for reasons to be developed later, it is in the medium of architecture that some of the most characteristically Islāmic works of art are found. Other techniques, on the other hand, acquire varying forms and emphases. Sculpture in the round hardly existed as a major art form, and, although such was also the case of all Mediterranean arts at the time of Islām's growth, one does not encounter the astounding rebirth of sculpture that occurred in the West. Wall painting existed but has generally been poorly preserved; the great Islāmic art of painting was limited to the illustration of books. The unique feature of Islāmic techniques is the astounding development taken by the so-called decorative arts—*e.g.,* woodwork, glass, ceramics, metalwork, textiles. New techniques were invented and spread throughout the Muslim world—at times even beyond its frontiers. In dealing with Islām, therefore, it is quite incorrect to think of these techniques as the "minor" arts. For the amount and intensity of creative energies spent on the decorative arts transformed them into major artistic forms, and their significance in defining a profile of the aesthetic and visual language of Islāmic peoples is far greater than in the instances of many other cultures. Furthermore, since, for a variety of reasons to be discussed later, the Muslim world did not develop until quite late the notion of "noble" arts, the decorative arts have reflected far better the needs and ambitions of the culture as a whole. The kind of conclusion that can be reached about Islāmic civilization through its visual arts thus extends far deeper than is usual in the study of an artistic tradition, and it requires a combination of archaeological, art-historical, and textual information.

Definition of a culture through an art

An example may suffice to demonstrate the point. Among all the techniques of Islāmic visual arts, the most important one was the art of textiles. Textiles, of course, were used for daily wear at all social levels and for all occasions. But clothes were also the main indicators of rank, and they were given as rewards or as souvenirs by princes, high and low. They were a major status symbol, and their manufacture and distribution were carefully controlled through a complicated institution known as the *ṭirāz.* Major events were at times celebrated by being depicted on silks. Many texts have been identified that describe the hundreds of different kinds of textiles that existed. Since textiles could easily be moved, they became a vehicle for the transmission of artistic themes within the Muslim world and beyond its frontiers. In the case of this one technique, therefore, one is not dealing simply with a medium of the decorative arts but with a key medium in the definition of a given time's taste, of its practical functions, and of the ways in which its ideas were distributed. The more unfortunate point is that the thousands of fragments that have remained have not yet been studied in a sufficiently systematic way, and in only a handful of instances has it been possible to relate individual fragments to known texts. When more work has been completed, however, a study of this one medium should contribute significantly to the commercial, social, and aesthetic history of Islām, as well as explain much of the impact that Islāmic art had beyond the frontiers of the Muslim world.

The following survey of Islāmic visual arts, therefore, will be primarily a historical one, for it is in development through time that the main achievements of Islāmic art can best be understood. At the same time, other features peculiar to this tradition will be kept in mind: the varying importance of different lands, each of which had identifiable artistic features of its own, and the uniqueness of some techniques of artistic creativity over others.

ORIGINS

Assimilation of earlier artistic traditions

Islāmic visual arts were created by the confluence of two entirely separate kinds of phenomena: a number of earlier artistic traditions and a new faith. The arts inherited by Islām were of extraordinary technical virtuosity and stylistic or iconographic variety. All the developments of arcuated and vaulted architecture that had taken place in Iran and in the Roman Empire were available in their countless local variants. Stone, baked brick, mud brick, and wood existed as mediums of construction, and all the complicated engineering systems developed particularly in the Roman Empire were still utilized from Spain to the Euphrates. All the major techniques of decoration were still used, except for monumental sculpture. In secular and in religious art, a more or less formally accepted equivalence between representation and represented subject had been established. Technically, therefore, as well as ideologically, the Muslim world took over an extremely sophisticated system of visual forms; and, since the Muslim conquest was accompanied by a minimum of destruction, all the monuments, and especially the attitudes attached to them, were passed on to the new culture.

The second point about the pre-Islāmic traditions is the almost total absence of anything from Arabia itself. While archaeological work in the peninsula may modify this conclusion in part, it does seem that Islāmic art formed itself entirely in some sort of relationship to non-Arab traditions. Even the rather sophisticated art created in earlier times by the Palmyrenes or by the Nabataeans had almost no impact on Islāmic art, and the primitively conceived ḥaram in Mecca, the only pre-Islāmic sanctuary maintained by the new faith, remained as a unique monument that was almost never copied or imitated despite its immense religious significance. The pre-Islāmic sources of Islāmic art are thus entirely extraneous to the milieu in which the new faith was created. In this respect the visual arts differ considerably from most other aspects of Islāmic culture.

This is not to say that there was no impact of the new faith on the arts, but to a large extent it was an incidental impact, the result of the existence of a new social and political entity rather than of a doctrine. Earliest Islām as seen in the Qur'ān or in the more verifiable accounts of the Prophet's life simply do not deal with the arts, either on the practical level of requiring or suggesting forms as expressions of the culture or on the ideological level of defining a Muslim attitude toward images. In all instances, concrete Qur'ānic passages later used for the arts had their visual significance extrapolated.

There is no prohibition against representations of living things, and not a single Qur'ānic passage refers clearly to the mosque, eventually to become the most characteristically Muslim religious building. In the simple, practical, and puritanical milieu of early Islām, aesthetic or visual questions simply did not arise.

Influence of Islām in establishing a new artistic tradition

The impact of the faith on the arts occurred rather as the fledgling culture encountered the earlier non-Islāmic world and sought to justify its own acceptance or rejection of new ways and attitudes. The discussion of two examples of particular significance illustrates the point. One is the case of the mosque. The word itself derives from the Arabic *masjid,* "a place where one prostrates one's self (in front of God)." It was a common term in pre-Islāmic Arabic and in the Qur'ān, where it is applied to sanctuaries in general without restriction. If a more concrete significance was meant, the word was used in construct with some other term, as in *masjid al-ḥaram* to refer to the Meccan sanctuary. There was no need in earliest times for a uniquely Muslim building, for any place could be used for private prayer as long as the correct direction (*qiblah,* originally Jerusalem, but very soon Mecca) was observed and the proper sequence of gestures and pious statements was followed. In addition to private prayer, which had no formal setting, Islām instituted a collective prayer on Fridays, where the same ritual was accompanied by a sermon from the *imām* (leader of prayer, originally the Prophet, then his successors, and later legally any able-bodied Muslim) and by the more complex ceremony of the *khuṭbah,* a collective swearing of allegiance to the community's leadership. This ceremony served to strengthen the common bond between all members of the *ummah,* the Muslim "collectivity," and its importance in creating and maintaining the unity of early Islām has often been emphasized. There were two traditional locales for this event in the Prophet's time. One was his private house, whose descriptions have been preserved; it was a large open space with private rooms on one side and rows of palm trunks making a colonnade on two other sides, the deeper colonnade being on the side of the *qiblah.* The Prophet's house was not a sanctuary but simply the most convenient place for the early community to gather. Far less is known about the second place of gathering for the Muslim community. It was used primarily on major feast days, such as the end of the fasting period or the feast of sacrifice. It was called a *muṣallā,* literally "a place for prayer," and *muṣallā*s were usually located outside city walls. Nothing is known about the shape taken by *muṣallā*s, but in all probability they were as simple as pre-Islāmic pagan sanctuaries: large enclosures surrounded by a wall and devoid of any architectural or ornamental feature.

Altogether then there was hardly anything that could be identified as a holy building or as an architectural form. To be complete, one should add two additional features. One is an action, the call to prayer (*adhān*). It became, fairly rapidly, a formal moment preceding the gathering of the faithful. One man would climb on the roof and proclaim that God is great and that men must congregate to pray. There was no formal monument attached to the ceremony, though it led eventually to the ubiquitous minaret. The other early feature was an actual structure. It was the *minbar,* a chair with several steps on which the Prophet would climb in order to preach. The monument itself had a pre-Islāmic origin, but Muḥammad transformed it into a characteristically Muslim form.

With the exception of the *minbar,* only a series of actions was formulated in early Islāmic times. There were no forms attached to them, nor were any needed. But, as the Muslim world grew in size, the contact with many other cultures brought about two developments. On the one hand there were thousands of examples of beautiful religious buildings that impressed the conquering Arabs. But, more importantly, the need arose to preserve the restricted uniqueness of the community of faithful and to express its separateness from other groups. Islāmic religious architecture began with this need and, in ways to be described later, created a formal setting for the activities, ceremonies, and ideas that had been formless at the outset.

Muslim iconoclasm

A second and closely parallel development of the impact of the Islāmic religion on the visual arts is the celebrated question of a Muslim iconoclasm. As has already been mentioned, the Qur'ān does not utter a word for or against the representation of living things. It is equally true that from about the middle of the 8th century a prohibition had been formally stated, and thenceforth it would be a standard feature of Islāmic thought, even though the form in which it is expressed has varied from absolute to partial and even though it has never been totally followed. The justification for the prohibition tended to be that any representation of a living thing was an act of competition with God, for he alone can create something that is alive. It is striking that this theological explanation reflects the state of the arts in the Christian world at the time of the Muslim conquest—a period of iconoclastic controversy. It may thus be suggested that Islām developed an attitude toward images as it came into contact with other cultures and that its attitude was negative because the arts of the time appeared to lead easily to dreaded idolatry. While it is only by the middle of the 8th century that there is actual proof of the existence of a Muslim doctrine, it is likely that, more or less intuitively, the Muslims felt a certain reluctance toward representations from the very beginning. For all monuments of religious art are devoid of any representations; even a number of attempts at representational symbolism in the official art of coinage were soon abandoned.

This rapid crystallization of Islāmic attitudes toward images has considerable significance. For practical purposes, representations are not found in religious art, although matters are quite different in secular art. Instead there occurred very soon a replacement of imagery with calligraphy and the concomitant transformation of calligraphy into a major artistic medium. Furthermore, the world of Islām tended to seek means of representing the holy other than by images of men, and one of the main problems of interpretation of Islāmic art is that of the degree of means it achieved in this search. But there is a deeper aspect to this rejection of holy images. Although the generally Semitic or specifically Jewish sources that have been given to Islāmic iconoclasms have probably been exaggerated, the reluctance imposed by the circumstances of the 7th century transformed into a major key of artistic creativity the magical fear of visual imagery that exists in all cultures but that is usually relegated to a secondary level. This uniqueness is certainly one of the main causes of the abstract tendencies that are among the great glories of the tradition. Even when a major art of painting did develop, it remained always somehow secondary to the mainstream of the culture's development.

Both in the case of the religious building and in that of the representations, therefore, it was the contact with pre-Islāmic cultures in Muslim-conquered areas that compelled Islām to transform its practical and unique needs into monuments and to seek within itself for intellectual and theological justifications for its own instincts. The great strength of early Islām was that it possessed within itself the ideological means to put together a visual expression of its own, even though it did not develop at the very beginning a need for such an expression.

One last point can be made about the origins of Islāmic art. It concerns the degree of importance taken by the various artistic and cultural entities conquered by the Arabs in the 7th and 8th centuries, for the early empire had gathered in regions that had not been politically or even ideologically related for centuries. During the first century or two of Islām, the main models and the main sources of inspiration were certainly the Christian centres around the Mediterranean. But the failure to capture Constantinople and to destroy the Byzantine Empire also made these Christian centres inimical competitors, whereas the whole world of Iran became an integral part of the empire, even though the conquering Arabs were far less familiar with the latter than with the former. A much more complex problem is posed by conversions, for it is through the success of the militant Muslim religious mission that the culture expanded so rapidly. Insofar as one can judge, it is the common folk, primarily in cities, who took over the new faith most rapidly; and thus there was added in early Islāmic culture a folk element whose impact may have been larger than has hitherto been imagined.

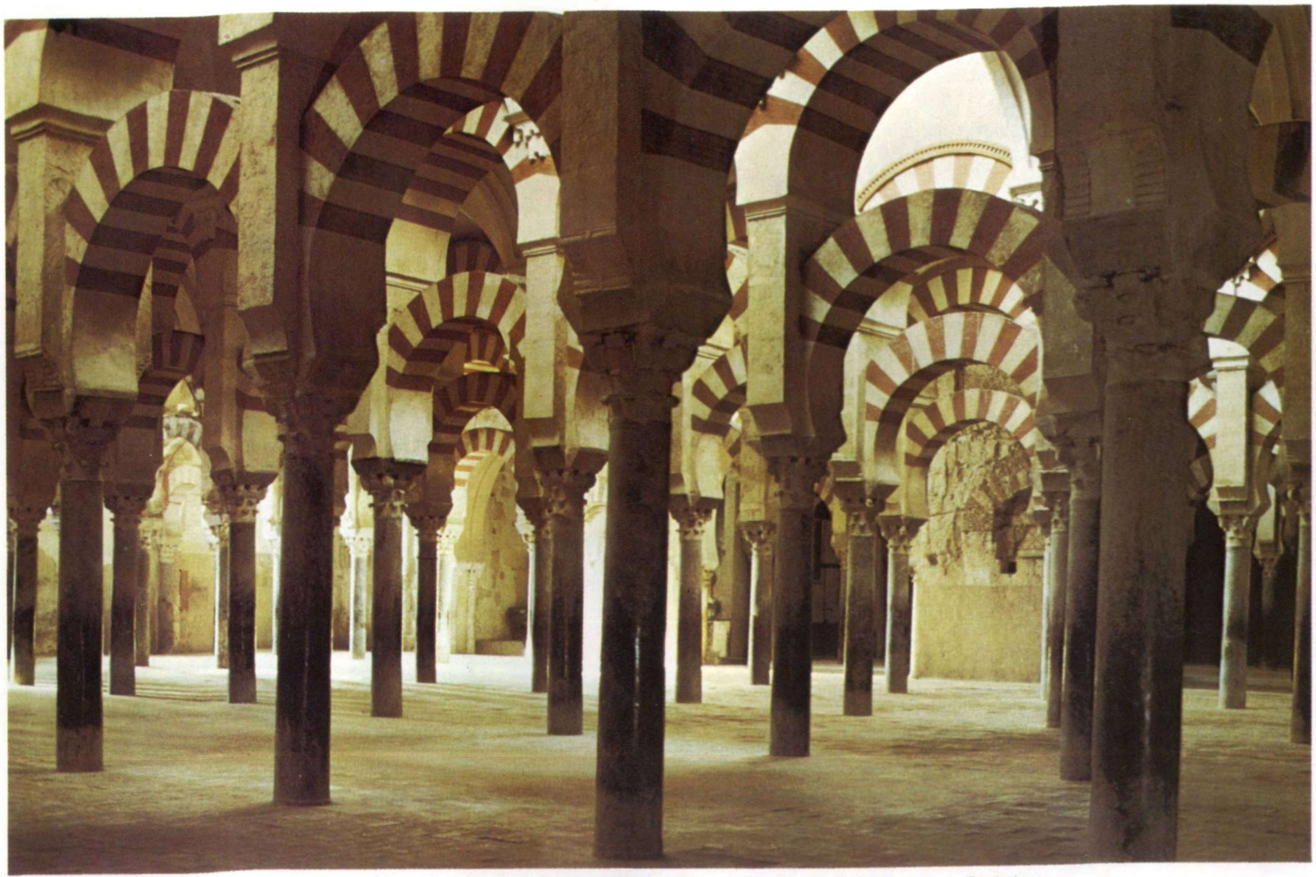
Folk element in Islāmic art

These preliminary considerations on the origins of Islāmic art have made it possible to outline several of the themes and problems that remained constant features of the tradition: a self-conscious sense of uniqueness when compared to others; a continuous reference to its own Qur'anic sources; a constant relationship to many different cultures; a folk element; and a variety of regional developments. None of these features remained constant, not even those aspects of the faith that affected the arts. But while they changed, the fact of their existence, their structural presence, remained a constant of Islāmic art.

EARLY PERIOD: THE UMAYYAD AND 'ABBĀSID DYNASTIES

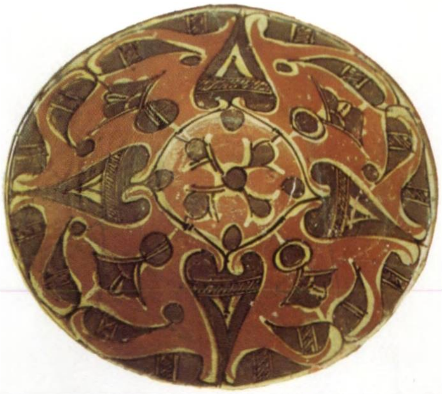
Of all the recognizable periods of Islāmic art, this is by far the most difficult one to explain properly, even though it is quite well documented. There are two reasons for this difficulty. On the one hand, it was a formative period, a time when new forms were created that identify the aesthetic and practical ideals of the new culture. Such periods are difficult to define when, as in the case of Islām, there was no artistic need inherent to the culture itself. The second complication derives from the fact that Muslim conquest hardly ever destroyed former civilizations with its own established creativity. Material culture, therefore, continued as before, and archaeologically it is almost impossible to distinguish between pre-Islāmic and early Islāmic artifacts. Paradoxical though it may sound, there is an early Islāmic Christian art of Syria and Egypt, and in many other regions the parallel existence of a Muslim and of a non-Muslim art continued for centuries. What did happen during early Islāmic times, however, was the establishment of a dominant new taste, and it is the nature and character of this taste that has to be explained. It occurred first in Syria and Iraq, the two areas with the largest influx of Muslims and with the two successive capitals of the empire, Damascus under the Umayyads and Baghdad under the early 'Abbāsids. From Syria and Iraq this new taste spread in all directions and adapted itself to local conditions and local materials, thus creating considerable regional and chronological variations in early Islāmic art.

From a historical point of view two major dynasties are involved. One is the Umayyad dynasty, which ruled from 661 to 750 and whose monuments are datable from 680 to 745. It was the only Muslim dynasty ever to control the whole of the Islāmic-conquered world. The second dynasty is the 'Abbāsid dynasty; technically its rule extended as late as 1258, but in reality its princes ceased to be a significant cultural factor after the second decade of the 10th century. The 'Abbāsids no longer controlled Spain, where an independent Umayyad caliphate had been established; and in Egypt as well as in northeastern Iran a number of more or less independent dynasties appeared, such as the Ṭūlūnids or the Sāmānids. Although recent research tends to make the conclusion less certain than it used to be for the Sāmānids and northeast Iran, the initial impulse for the artistic creativity of these dynasties came from the main 'Abbāsid centres in Iraq. While in detailed studies it is possible to distinguish between Umayyad and 'Abbāsid art or between the arts of various provinces, the key features of the first three centuries of Islāmic art (roughly through the middle of the 10th century) are the interplay between local or imperial impulses and the creation of new forms and functions.

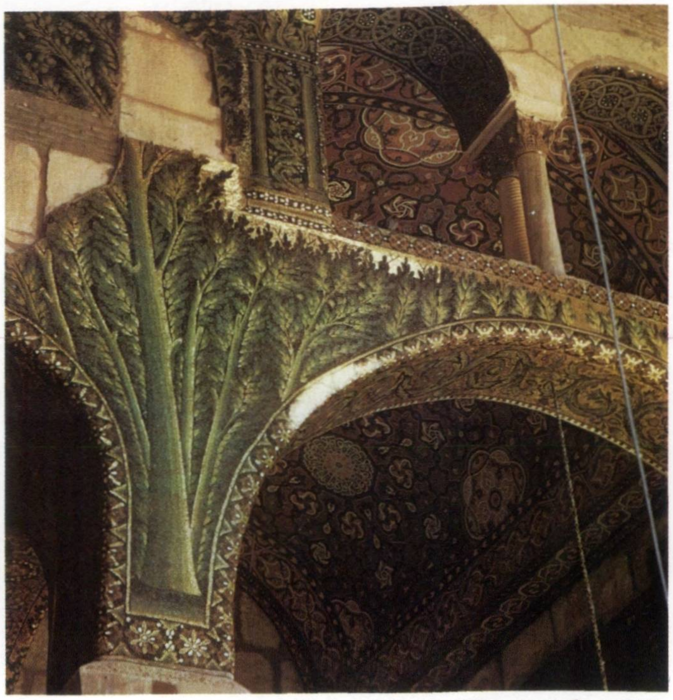


Interior of the Great Mosque of Córdoba, Spain, begun 785. The building is now a Christian cathedral.

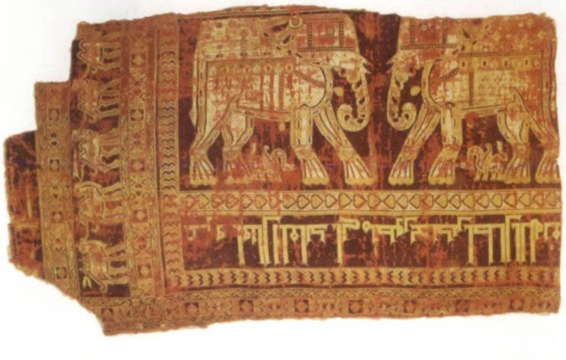
Umayyad and ʿAbbasid Art



Bowl from Nīshāpūr, lead-glazed earthenware with a slip decoration. In the Victoria and Albert Museum, London.



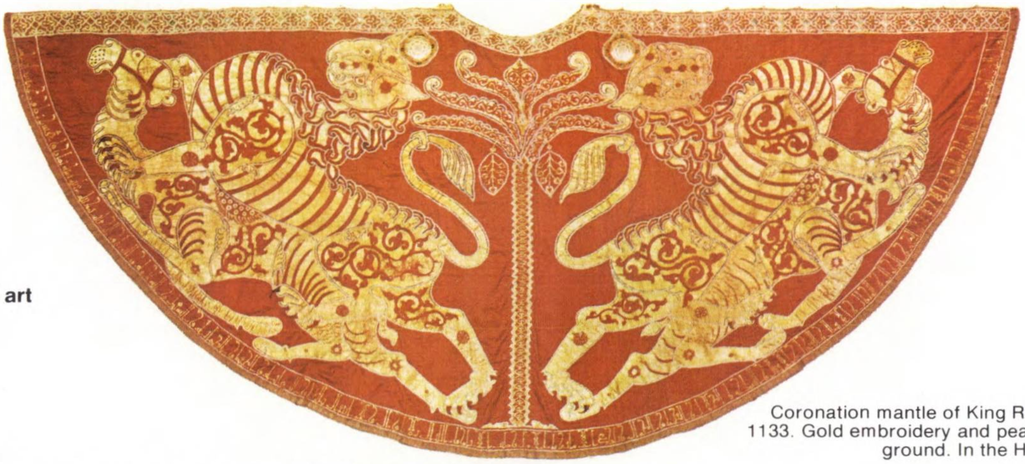
Mosaics decorating the portico of the Great Mosque of Damascus, Syria, 715.



Woven silk bearing the inscription "Glory and happiness to Qaid Abul Mansur Nudjkatin; may God continue his prosperity," 10th century. In the Louvre, Paris. 94 × 52 cm.

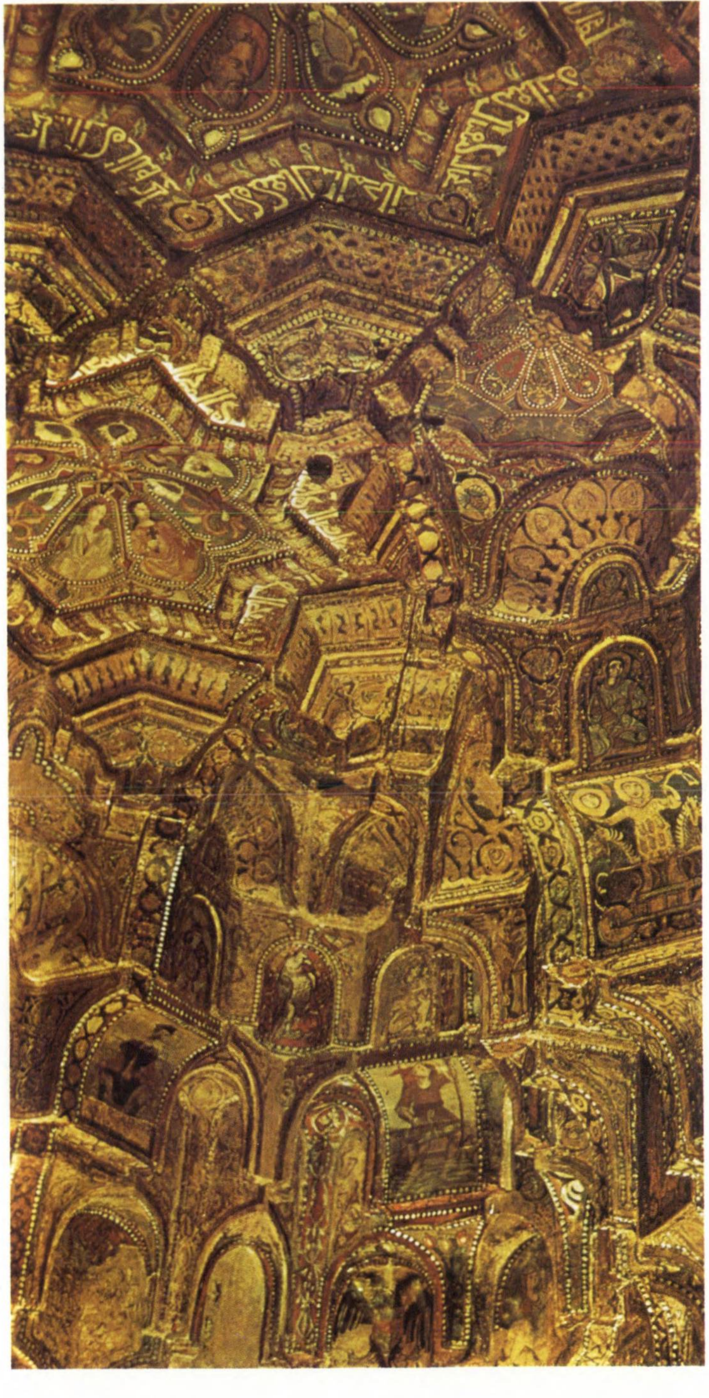
Plate 1: By courtesy of (centre left) the Victoria and Albert Museum, London; photographs, (top) Alfonso Gutierrez Escera—Ostman Agency, (bottom left) Cliche Musees Nationaux, (bottom right) J.E. Dayton, London

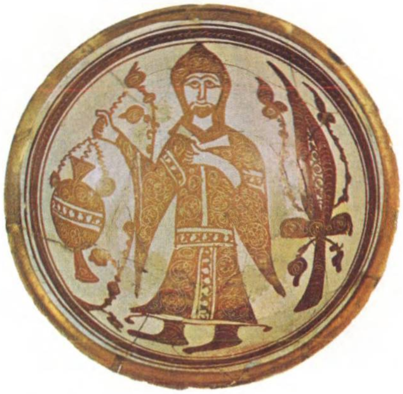
Fatimid art



Coronation mantle of King Roger II of Sicily, 1133. Gold embroidery and pearls on a red silk ground. In the Hofburg, Vienna.

Ceiling of the Capella Palatina, Palermo, Sicily. The chapel was built by the Norman kings of Sicily and decorated by Fātimid artists.





Bowl of lustre-ware by the potter Sa'ad, depicting a Christian priest swinging a censer, first half of the 12th century. In the Victoria and Albert Museum, London.

Bronze griffin. In the Camposanto, Pisa, Italy.

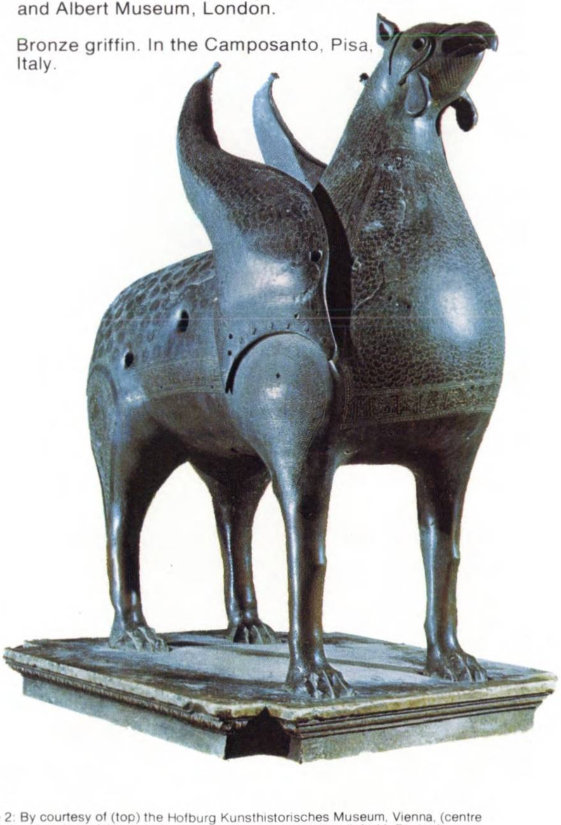
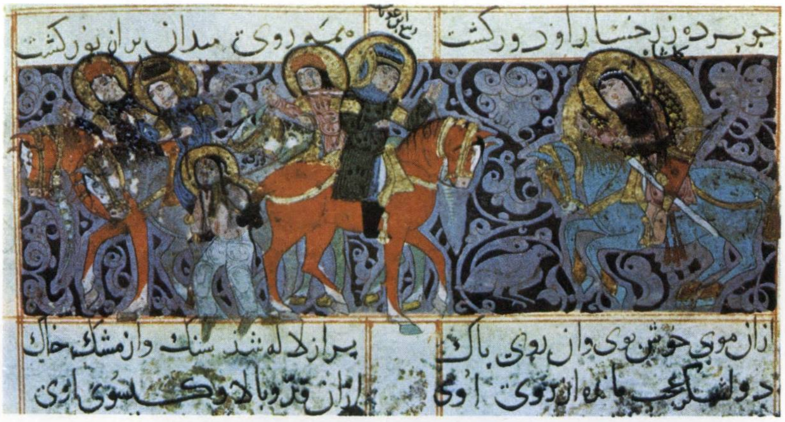


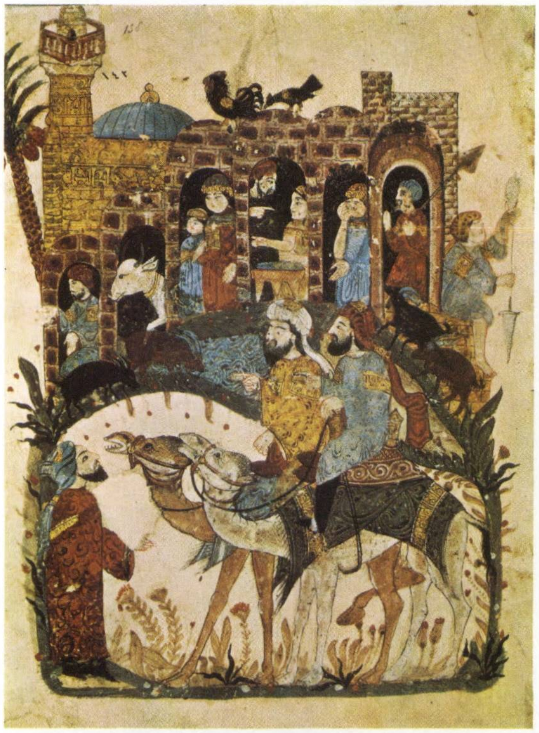
Plate 2: By courtesy of (top) the Hofburg Kunsthistorisches Museum, Vienna, (centre right) the Victoria and Albert Museum, London; photographs, (left) M. Desjardins—Realities, (bottom right) SCALA—Art Resource/EB Inc.

Plate 3: By courtesy of (top right) the Topkapı Saray Museum, Istanbul, (centre left) the Bibliotheque Nationale, Paris, (bottom left) the Smithsonian Institution, Freer Gallery of Art, Washington, D.C.; photograph, (right) Roland and Sabrina Michaud—Rapho/Photo Researchers from EB Inc.

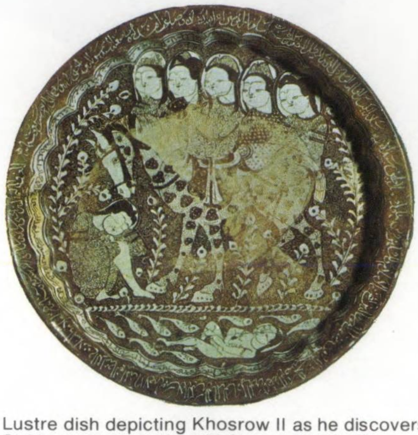
Seljuq art



"Golshāh has removed her veil during a battle," miniature from *Varqeh o-Golshāh,* 13th century. In the Topkapı Saray Museum, Istanbul. (Ms. Hazine 841, fol. 22.) 10.2 × 29.7 cm.

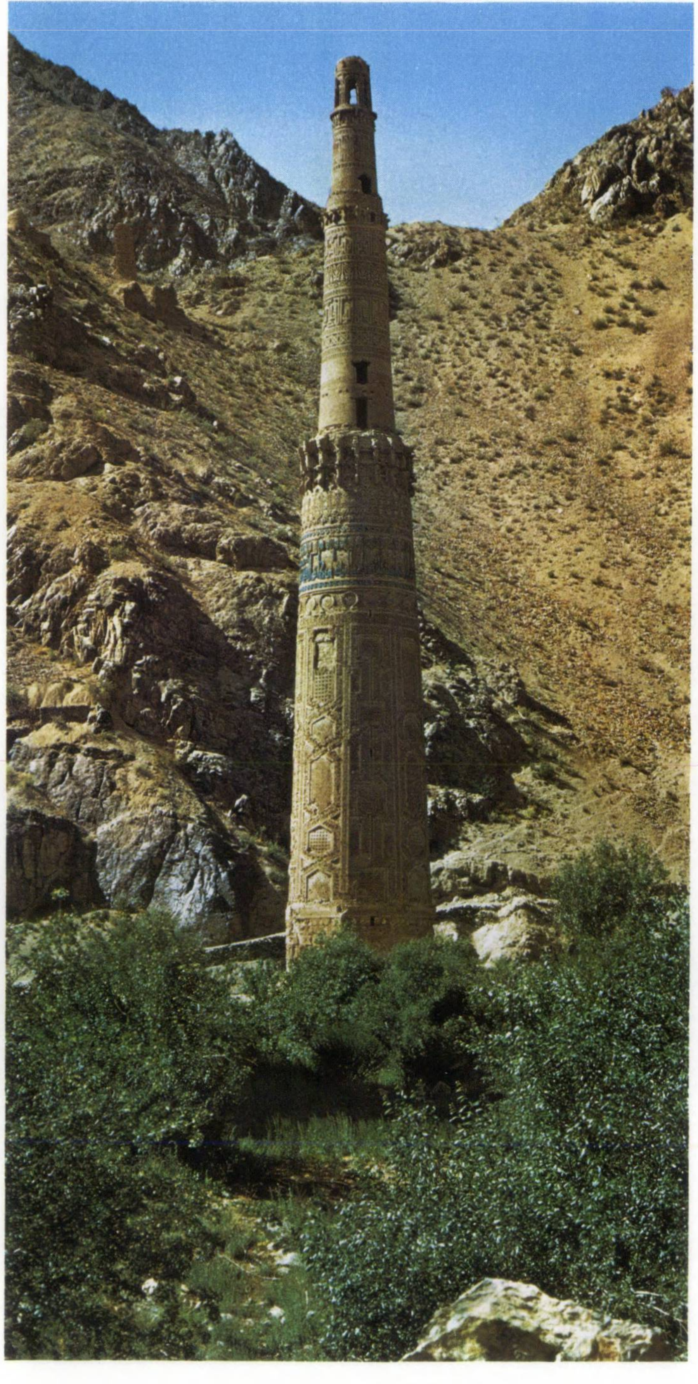


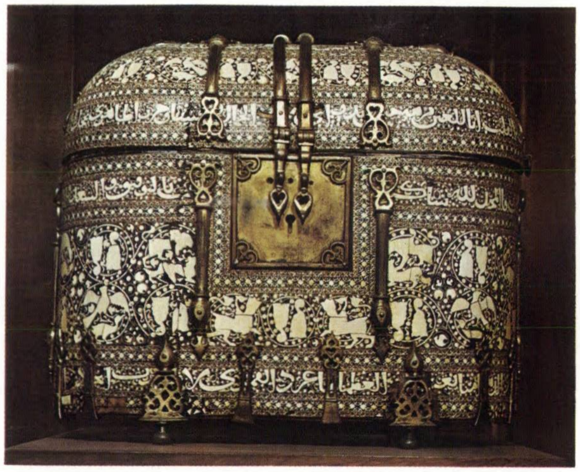
Discussion near a village, miniature painted by Yahyā ibn Mahmūd al-Wāsitī from the 43rd *maqāmah* of the *Maqāmāt* ("Assemblies") of al-Harīrī, 1237. In the Bibliothèque Nationale, Paris. (Ms. Arabe 5847, folio 138 r.). 34.8 × 26 cm.



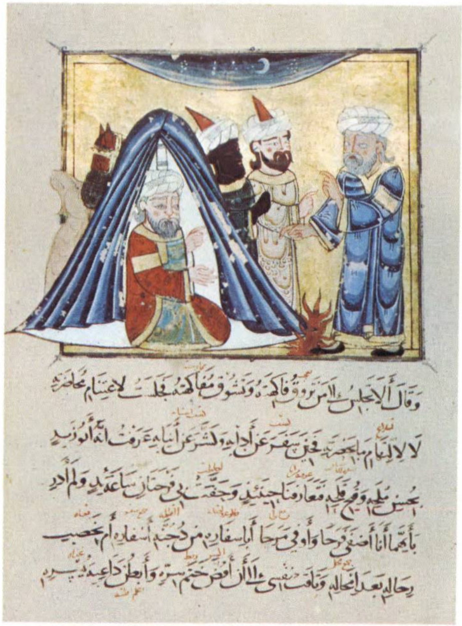
Lustre dish depicting Khosrow II as he discovers Shīrīn bathing, by Sayyid Shams ad-Dīn al Husani, from Kāshān, Iran, *c.* 1210. In the Freer Gallery of Art, Washington, D.C.

The minaret of Jām, Afghanistan, 1116–1202.



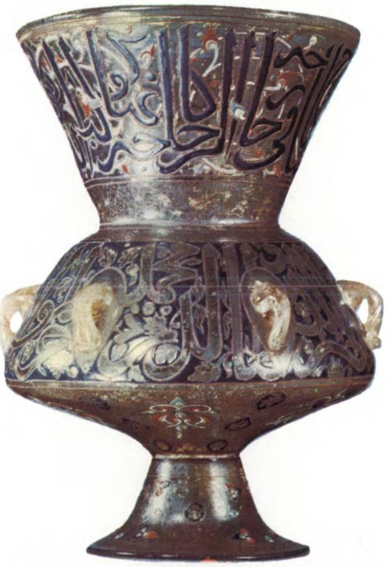


Ivory casket, 13th century. In the Palazzo Reale, Capella Palatina, Palermo, Sicily. 39 × 40 × 24 cm.

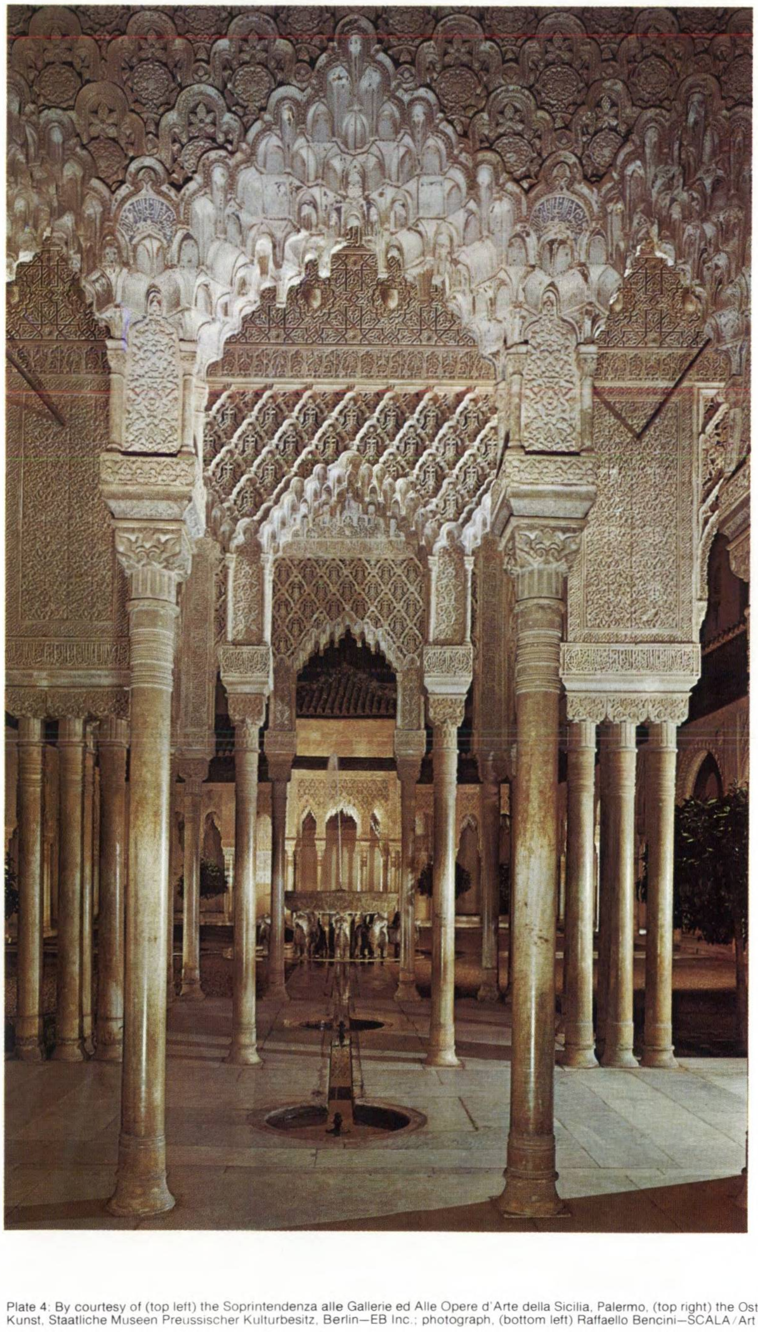


Al-Hārith talks to Abū Zayd in his tent, miniature from the 26th *maqāmah* of the *Maqāmāt* ("Assemblies") of al-Harīrī, probably Egyptian, 1334. In the Österreichische Nationalbibliothek, Vienna (MS. A. F. 9, folio 87 v). Miniature only, 13.7 × 15.8 cm.

Moorish and Mamluk art

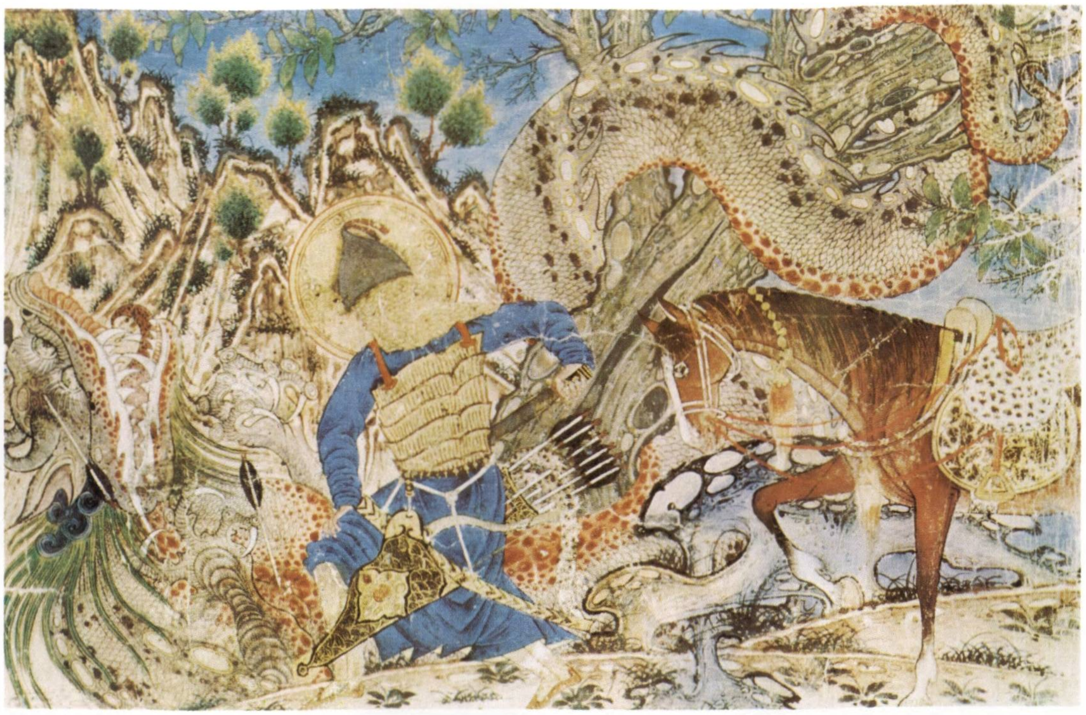


Hanging mosque lamp, enamelled and gilded glass, from Aleppo, Syria, *c.* 1300. In the Museum für Islamische Kunst, Staatliche Museen Preussischer Kulturbesitz, West Berlin.



Court of the Lions, the Alhambra, Granada, Spain, 14th century.

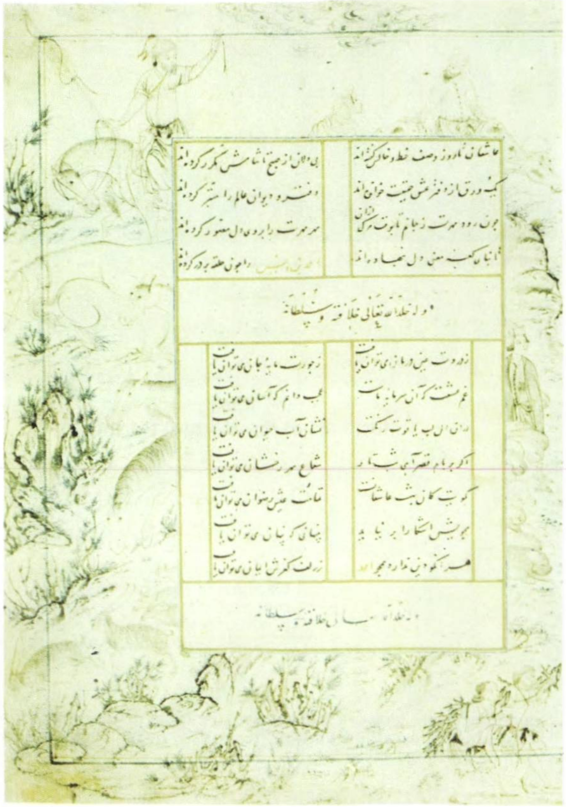
Plate 4: By courtesy of (top left) the Soprintendenza alle Gallerie ed Alle Opere d'Arte della Sicilia, Palermo, (top right) the Osterreichische Nationalbibliothek, Vienna, (bottom right) the Museum fur Islamische Kunst, Staatliche Museen Preussischer Kulturbesitz, Berlin—EB Inc.; photograph, (bottom left) Raffaello Bencini—SCALA/Art Resource from EB Inc.



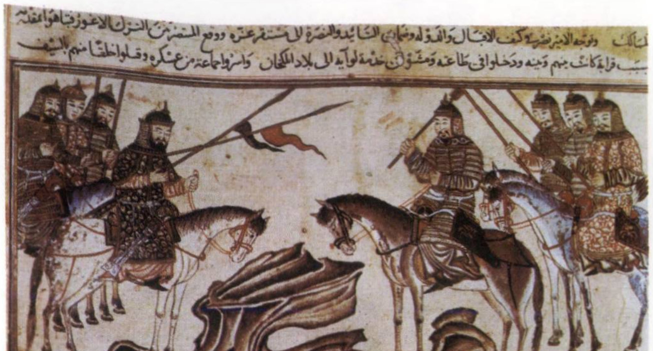
Bahrām Gūr killing a dragon, illustration from the *Shāh-nāmeh* ("Book of Kings") of Ferdowsī, known as the Demotte *Shāh-nāmeh*, 1320–60, from Tabriz, Iran. In the Cleveland Museum of Art. Height 40.6 cm.

Il-Khanid art of the Mongol Period

Pottery bowl from Kāshān, Iran, late 14th century. In the Victoria and Albert Museum, London.

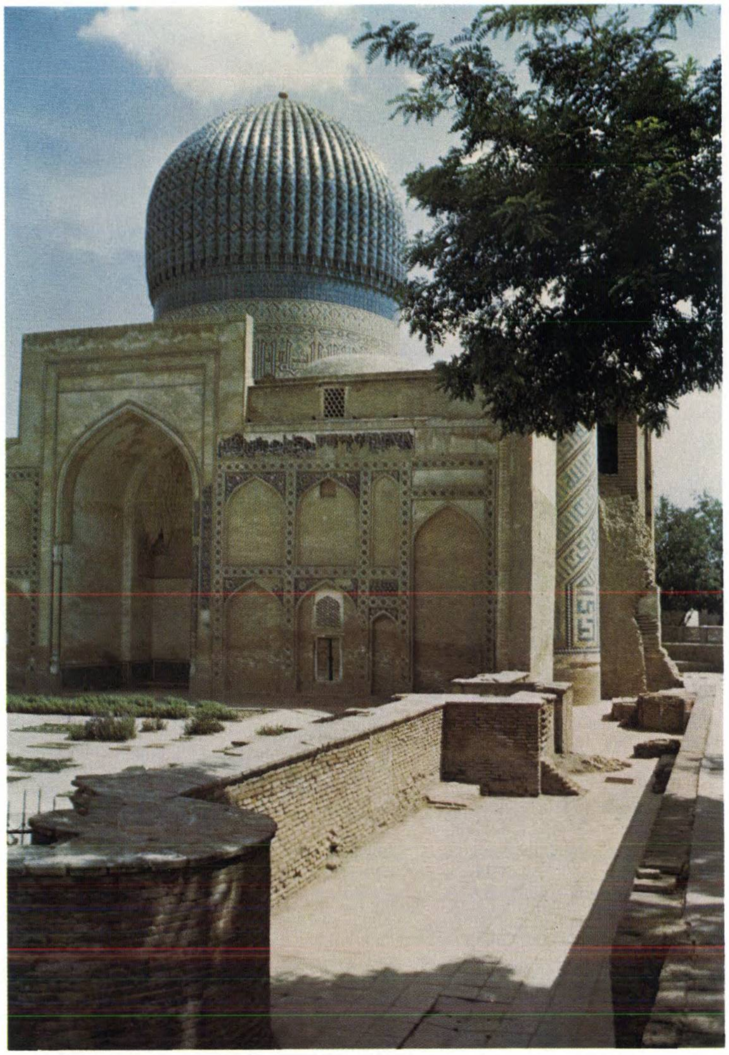


Diwān of Sultan Ahmad, pastoral border painted by Junayd, *c.* 1405, from Baghdad. In the Freer Gallery of Art, Washington, D.C. 29.2 × 20.3 cm.

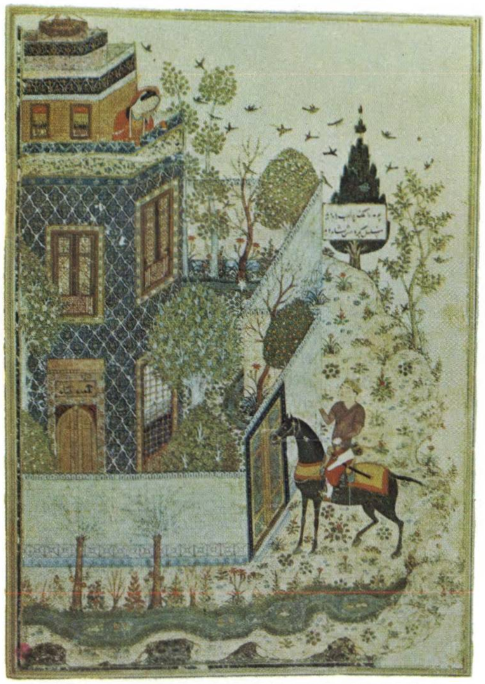


Mongol warriors, miniature from Rashīd ad-Dīn's *History of the World*, 1307. In the Edinburgh University Library, Scotland. Miniature only, 25 × 11.4 cm.

Plate 5: By courtesy of (top) the Cleveland Museum of Art, Ohio, Grace Rainey Rogers Fund, (centre left) the Victoria and Albert Museum, London, (bottom left) the Edinburgh University Library, Scotland, (bottom right) the Smithsonian Institution, Freer Gallery of Art, Washington, D.C.; photograph (centre left) John Webb

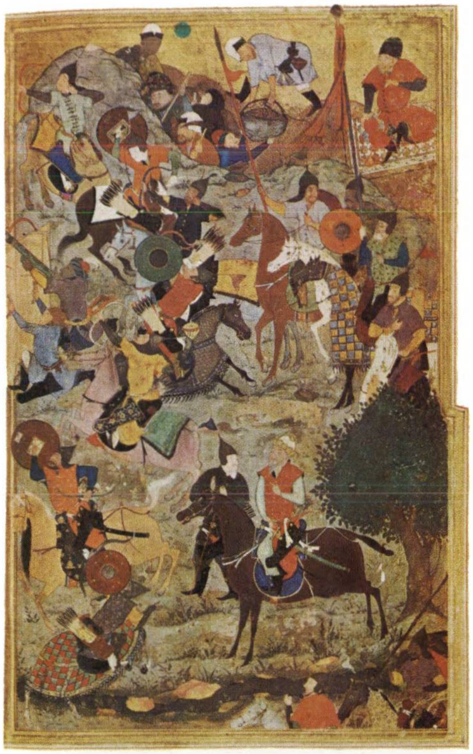


The mausoleum of Timur at Samarkand, 1434.



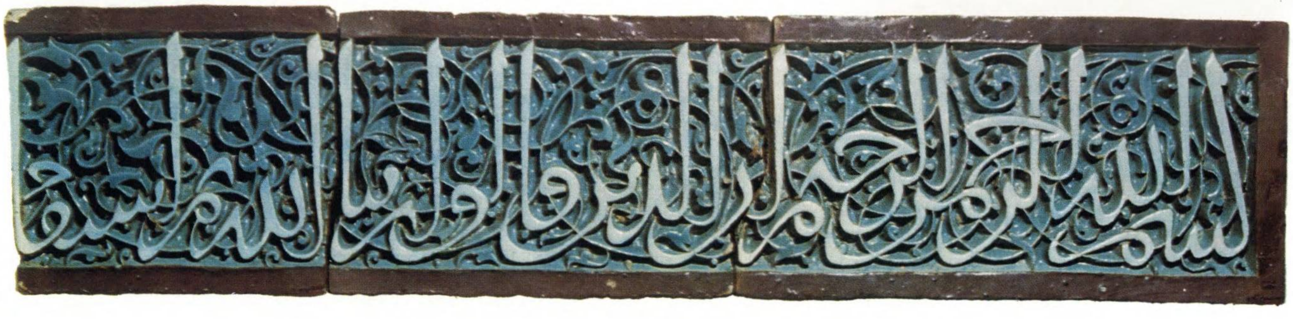
"Prince Humāy at the Gate of Humāyun's Castle," miniature painted by Junayd for the *Khamseh* of Khwāju Kermānī, 1396. In the British Library (MS. Add 18113, folio 18v). 29 × 20.2 cm.

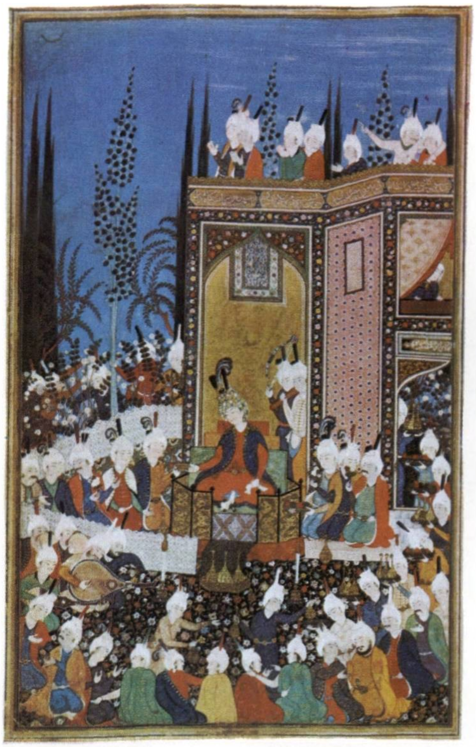
Timurid art of the Mongol Period



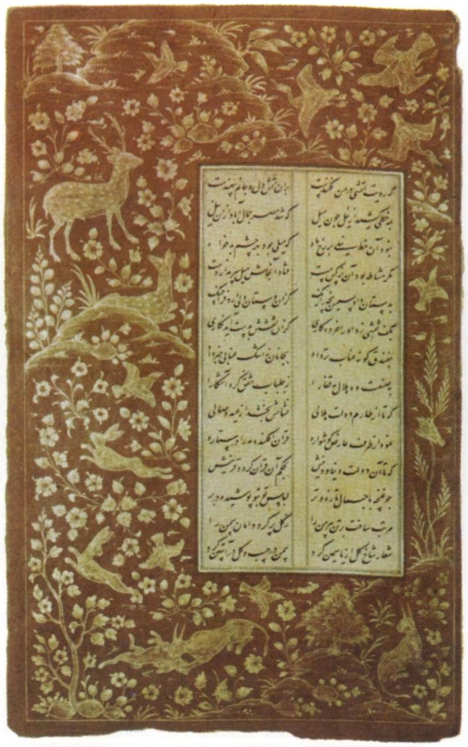
Capture of the fortress of the Knights Hospitallers at Smyrna, miniature from a *Zafar-nāmeh* (a life of Timur) by Behzād, *c.* 1490, from Herāt. In the John Work Garrett Library, Johns Hopkins University, Baltimore. 25.2 × 13 cm.

Section of relief tilework from the mausoleum of Bayram Khān at Fatḥābād, Uzbekistan, late 14th to early 15th century. In the Victoria and Albert Museum, London. Length 1.52 m.



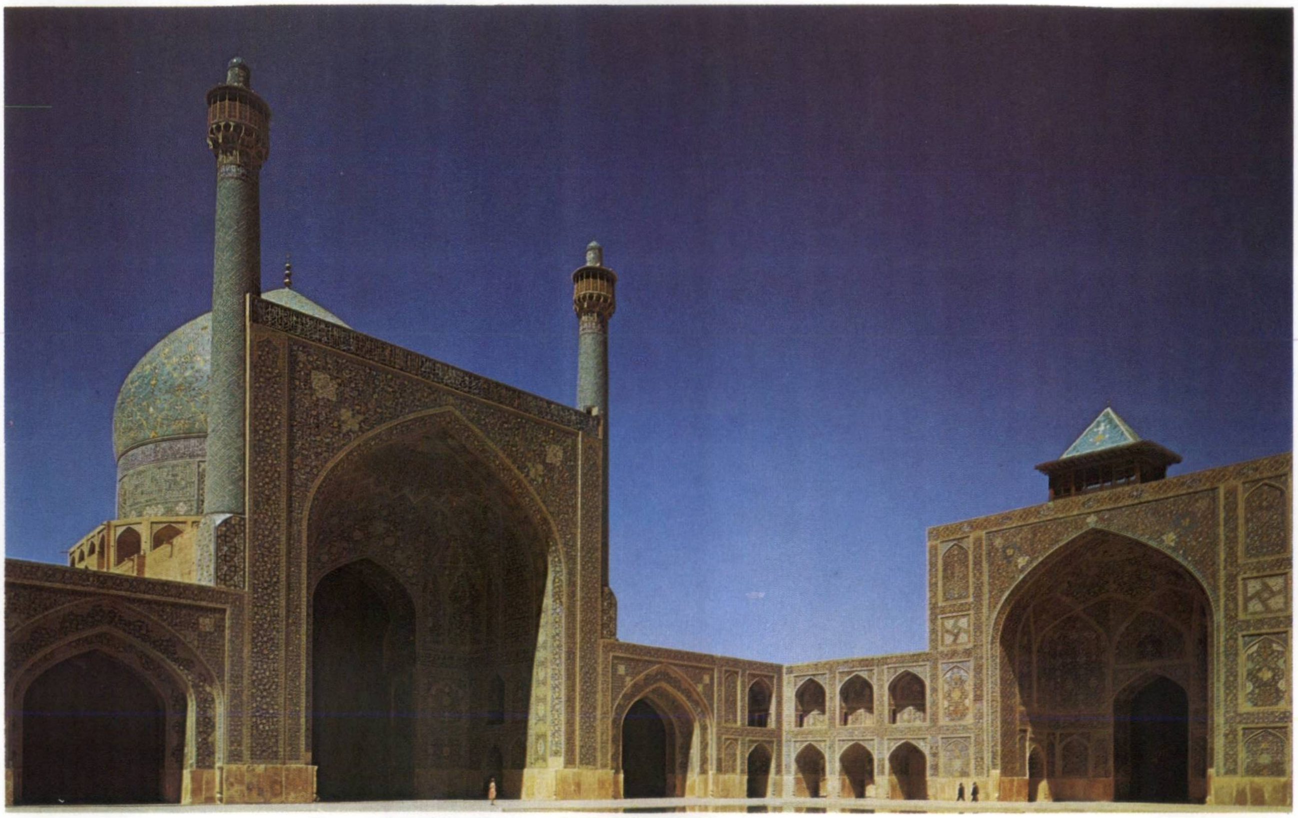


The Feast of 'Id, illustration from a *Divān* of Hāfez, signed Sultan Muhammad, *c.* 1520. In a private collection. 24 × 16 cm.



Miniature from *Yusof o-Zalikhā* by Jāmī, the text in small *nasta 'liq* calligraphy, 1557. In the Freer Gallery of Art, Washington, D.C. 25.2 × 15 cm.

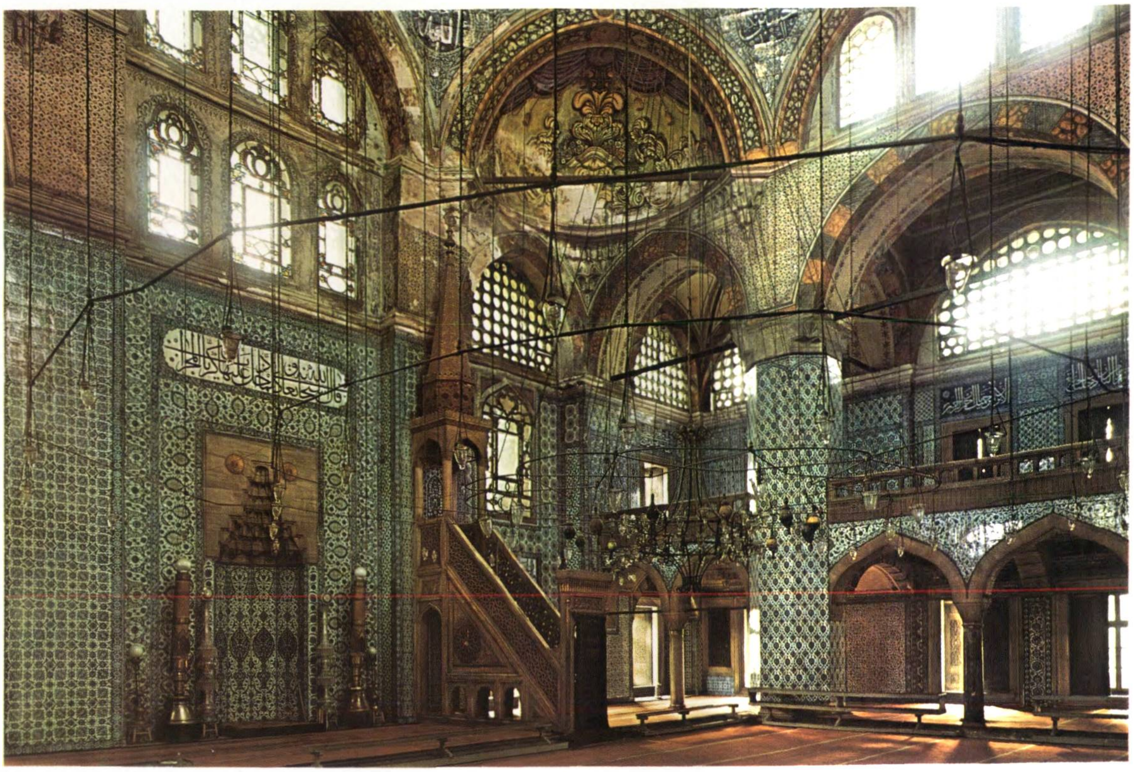
Safavid art



Two *eyvān*s of the Masjed-e Shāh of 'Abbās I the Great at Isfahan, Iran, 17th century.

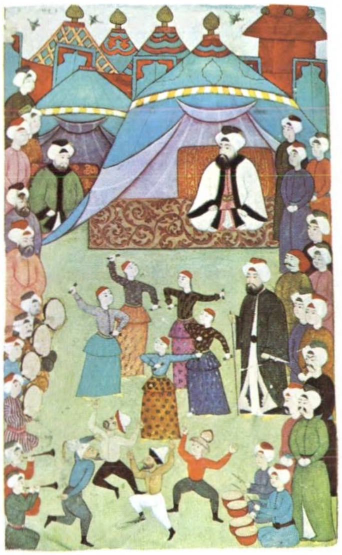
Plate 6: By courtesy of (centre right) the John Work Garrett Library, Johns Hopkins University, Baltimore, (bottom) Victoria and Albert Museum, London; photographs, (top right) reproduced by permission of the British Library, (top left) Alex Langley—Photo Researchers/EB Inc., (centre right) Walters Art Gallery, Baltimore, (bottom) John Webb

Plate 7: By courtesy of (top left) a private collector, Cambridge, Mass. (formerly in the collection of Louis J. Cartier), (top right) the Smithsonian Institution, Freer Gallery of Art, Washington, D.C.; photographs, (top left) Fogg Art Museum, Cambridge, Mass., (bottom) Ray Manley—Shostal/EB Inc.

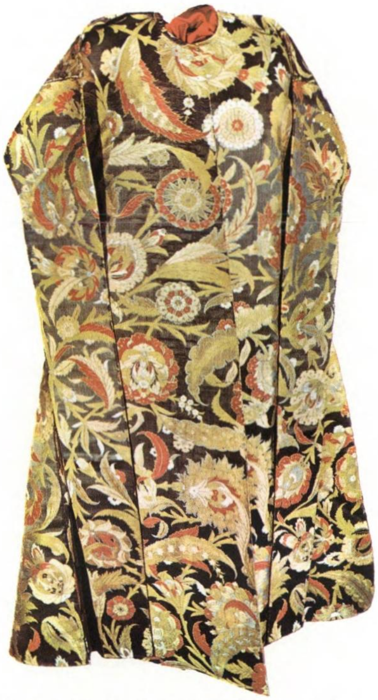


Interior of the Rüstem Paşa Mosque, Istanbul, showing the coloured tile decoration.

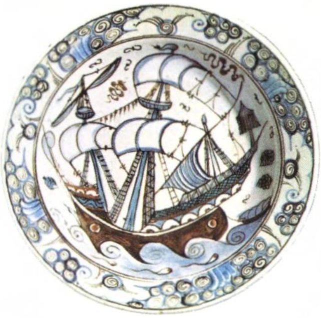
Ottoman art



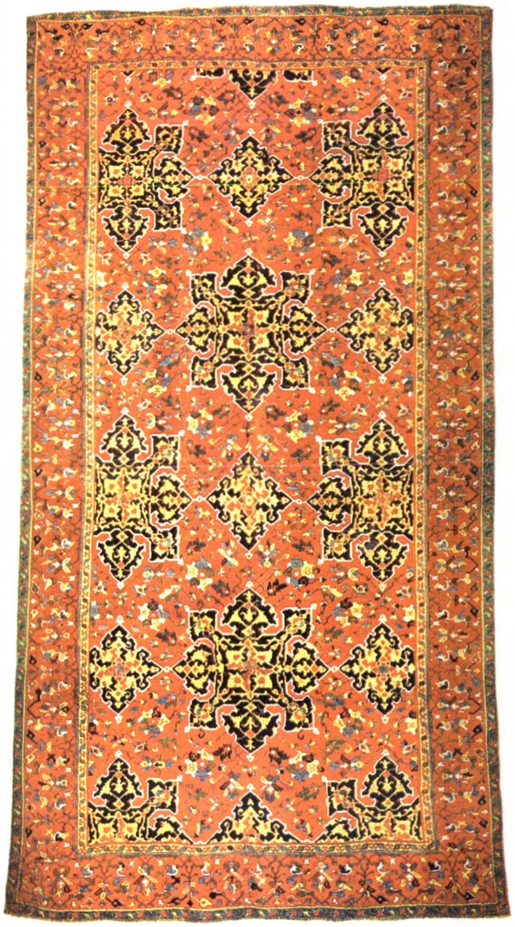
The Sultan watching dancers and comedians in the Hippodrome, illustration from the *Surname-i Vehbi* of Ahmed III (1703–30), painted by Levnî. In the Topkapı Saray Museum, Istanbul.



Silk caftan said to be that of Bayezid II (1481–1512). In the Topkapı Saray Museum, Istanbul.



Isnik ware dish, second half of the 16th century. In the Victoria and Albert Museum, London. Diameter 30.5 cm.



"Star Ushak" carpet from western Anatolia, late 16th to early 17th century. In the Metropolitan Museum of Art, New York City. 2.17 × 4.27 m.

Plate 8: By courtesy of (bottom left, centre) the Topkapı Saray Museum, Istanbul, (bottom centre) the Victoria and Albert Museum, London, (bottom right) the Metropolitan Museum of Art, New York, gift of Joseph V. McMullen; photographs, (top, bottom left, centre) Ara Guler—EB Inc., (bottom centre) John Webb, (bottom right) Otto E. Nelson

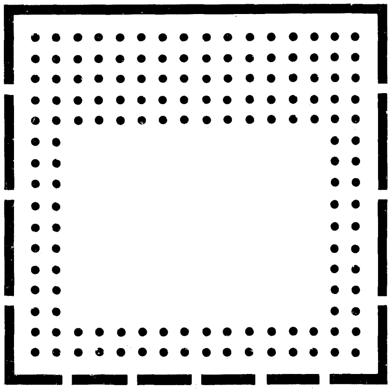
It is possible to study these centuries as a succession of clusters of monuments, but, since there are so many of them, a study can easily end up as an endless list. It is preferable, therefore, to centre the discussion of Umayyad and ʿAbbāsid monuments on the functional and morphological characteristics that identify the new Muslim world and only secondarily be concerned with stylistic progression or regional differences.

Origin of the mosque

Architecture. *Religious buildings.* The one obviously new function developed during this period is that of the mosque, or *masjid.* The earliest adherents of Islām used the private house of the Prophet in Medina as the main place for their religious and other activities and *muṣallā*s without established forms for certain holy ceremonies. The key phenomenon of the first decades that followed the conquest is the creation outside of Arabia of *masjid*s in every centre taken over by the new faith. These were not simply or even primarily religious centres. They were rather the community centres of the faithful, in which all social, political, educational, and individual affairs were transacted. Among these activities were common prayer and the ceremony of the *khuṭbah.* The first mosques were built primarily to serve as the restricted space in which the new community would take its own collective decisions. It is there that the treasury of the community was kept, and early accounts are full of anecdotes about the immense variety of events, from the dramatic to the scabrous, that took place in mosques. Since even in earliest times the Muslim community consisted of several superimposed and interconnected social systems, mosques reflected this complexity, and, next to large mosques for the whole community, tribal mosques and mosques for various quarters of a town or city are also known.

None of these early mosques has survived, and no descriptions of the smaller ones have been preserved. There do remain, however, accurate textual descriptions of the large congregational buildings erected at Kūfah and Basra in Iraq and at al-Fusṭāṭ in Egypt. At Kūfah a larger square was marked out by a ditch, and a covered colonnade known as a *ẓullah* (a shady place) was put up on the *qiblah* side. In 670 a wall pierced by many doors was built in place of the ditch, and colonnades were put up on all four sides, with a deeper one on the *qiblah.* In all probability the Basra mosque was very similar, and only minor differences distinguished the ʿAmr ibn al-ʿĀṣ mosque at al-Fusṭāṭ. Much has been written about the sources of this type of building, but the simplest explanation may be that this is the very rare instance of the actual creation of a new architectural type. The new faith's requirement for centralization, or a space for a large and constantly growing community, could not be met by any existing architectural form. Almost accidentally, therefore, the new Muslim cities of Iraq created the hypostyle mosque (a building with the roof resting on rows of columns). A flexible architectural unit, a hypostyle structure could be square or rectangular and could be increased or diminished in size by the addition or subtraction of columns. The single religious or symbolic feature of the hypostyle mosque was a *minbar* (a pulpit) for the preacher, and the direction of prayer was indicated by the greater depth of the colonnade on one side of the structure.

From E. Kuhnel, *Islamic Art and Architecture*



Plan of reconstructed mosque at Kūfah.

Versatility of the hypostyle plan

The examples of Kūfah, Basra, and al-Fusṭāṭ are particularly clear because they were all built in newly created cities. Matters are somewhat more complex when discussing the older urban centres taken over by Muslims. Although it is not possible to generalize with any degree of certainty, two patterns seem to emerge. In some cases, such as Jerusalem and Damascus and perhaps in most cities conquered through formal treaties, the Muslims took for themselves an available unused space and erected on it some shelter, usually a very primitive one. In Jerusalem this space happened to be a particularly holy one—the area of the Jewish Temple built by Herod I the Great, which had been left willfully abandoned and ruined by the triumphant Christian empire. In Damascus it was a section of a huge Roman temple area, on another part of which there was a church dedicated to St. John the Baptist. Unfortunately too little is known about other cities to be able to demonstrate that this pattern was a common one. The very same uncertainty surrounds the second pattern, which consisted in forcibly transforming sanctuaries of older faiths into Muslim ones. This was the case at Ḥamāh in Syria and at Yazd-e Khvāst in Iran, where archaeological proof exists of the change. There are also several literary references to the fact that Christian churches, Zoroastrian fire temples, and other older abandoned sanctuaries were transformed into mosques. Altogether, however, these instances probably were not too numerous, because in most places the Muslim conquerors were quite anxious to preserve local tradition and because few older sanctuaries could easily serve the primary Muslim need of a large centralizing space.

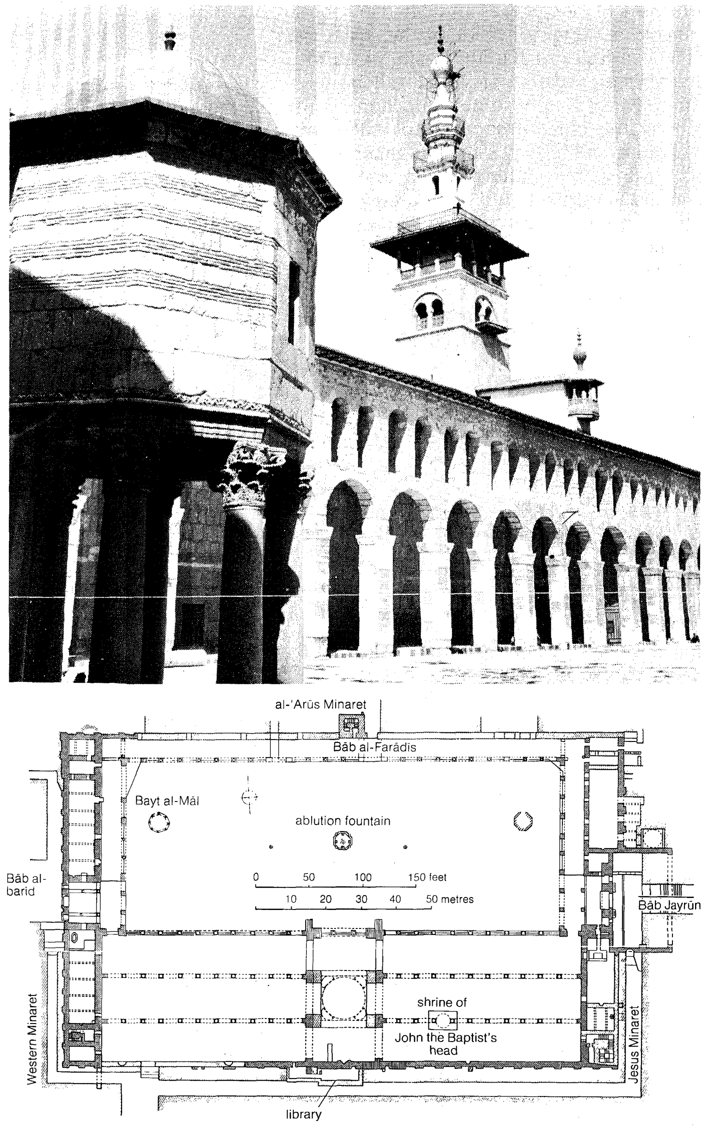
Transformation of existing pre-Islāmic sanctuaries into mosques

During the 50 years that followed the beginning of the Muslim conquest, the mosque, until then a very general concept in Islāmic thought, became a definite building reserved for a variety of needs required by the community of faithful in any one settlement. Only in one area, Iraq, did the mosque acquire a unique form of its own, the oriented hypostyle. Neither in Iraq nor elsewhere is there evidence of symbolic or functional components in mosque design. The only exception is that of the *maqṣūrah* (literally "closed-off space"), an enclosure, probably in wood, built near the centre of the *qiblah* wall. Its purpose was to protect the caliph or his replacement, for several attacks against major political figures had taken place. But the *maqṣūrah* was never destined to be a constant fixture of mosques, and its typological significance is limited.

During the rule of the Umayyad prince al-Walīd I (705–715), a number of complex developments within the Muslim community were crystallized in the construction of three major mosques, at Medina, Jerusalem, and Damascus. The very choice of these three cities is indicative: the city in which the Muslim state was formed and in which the Prophet was buried; the city held in common holiness by Jews, Christians, and Muslims, to which was rapidly accruing the mystical hagiography surrounding the Prophet's ascension into heaven; and the ancient city that became the capital of the new Islāmic empire. A first and essential component of al-Walīd's mosques was, thus, their imperial character; they were to symbolize the permanent establishment of the new faith and of the state that derived from it. They were no longer purely practical shelters but willful monuments.

Mosques of Medina, Jerusalem, and Damascus

Although the plans of al-Aqṣā Mosque in Jerusalem and of the mosque of Medina can be reconstructed with a fair degree of certainty, only the one at Damascus has been preserved with comparatively minor alterations and repairs. In plan the three buildings appear at first glance to be quite different from each other. The Medina mosque was essentially a large hypostyle with a courtyard. The colonnades on all four sides were of varying depth. Al-Aqṣā Mosque consisted of an undetermined number of naves (possibly as many as 15) parallel to each other in a north–south direction. There was no courtyard because the rest of the huge esplanade of the former Jewish temple served as the open space in front of the building. The Umayyad Mosque of Damascus is a rectangle 515 by 330 feet (157 by 100 metres) whose outer limits and three gates



Great Mosque of Damascus, Syria, built by al-Walīd I, 705–715. (Top) Courtyard with the Bayt al-Māl (treasury) on the left, beyond which can be seen one of the three towers that were the first minarets in Islām. (Below) Plan.

Paul Almasy

are parts of a Roman temple (a fourth Roman gate on the *qiblah* side was blocked). The interior consists of an open space surrounded on three sides by a portico and of a covered space of three equal long naves parallel to the *qiblah* wall that are cut in the middle by a perpendicular nave.

Appearance of the *mihrāb* and minaret

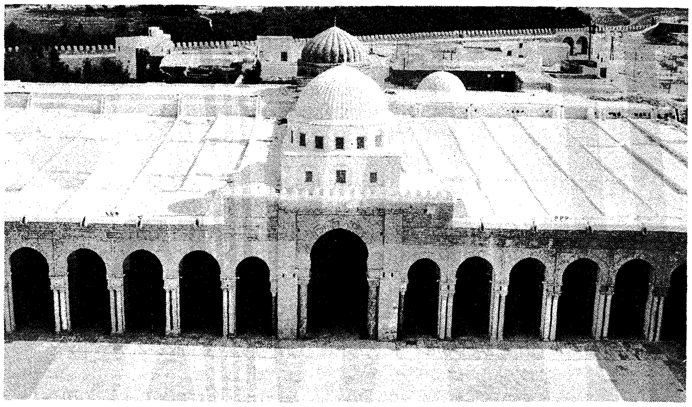
The three buildings share several important characteristics. They are all large spaces with a multiplicity of internal supports; and although only the Medina mosque is a pure hypostyle, the Jerusalem and Damascus mosques have the flexibility and easy internal communication characteristic of a hypostyle building. All three mosques exhibit a number of distinctive new practical elements and symbolic meanings. Many of these occur in all mosques; others are only known in some of them. The *miḥrāb,* for example, appears in all mosques. This is a niche of varying size that tends to be heavily decorated. It occurs in the *qiblah* wall, and, in all probability, its purpose was to commemorate the symbolic presence of the Prophet as the first *imām,* although there are other explanations. It is in Damascus only that the ancient towers of the Roman building were first used as minarets to call the faithful to prayer and to indicate from afar the presence of Islām (initially minarets tended to exist only in predominantly non-Muslim cities). All three mosques are also provided with an axial nave, a wider aisle unit on the axis of the building, which served both as a formal axis for compositional purposes and as a ceremonial one for the prince's retinue. Finally, all three buildings were heavily decorated with marble, mosaics, and woodwork. At least in the mosque of Damascus, it is further apparent that there was careful concern for the formal composition—a balance between parts that truly makes this mosque a work of art. This is particularly evident in the successful relationship established between the open space of the court and the facade of the covered *qiblah* side.

When compared to the first Muslim buildings of Iraq and Egypt, the monuments of al-Walīd are characterized by the growing complexity of their forms, by the appearance of uniquely Muslim symbolic and functional features, and by the quality of their construction. While the dimensions, external appearance, and proportions of any one of them were affected in each case by unique local circumstances, the internal balance between open and covered areas and the multiplicity of simple and flexible supports indicate the permanence of the early hypostyle tradition.

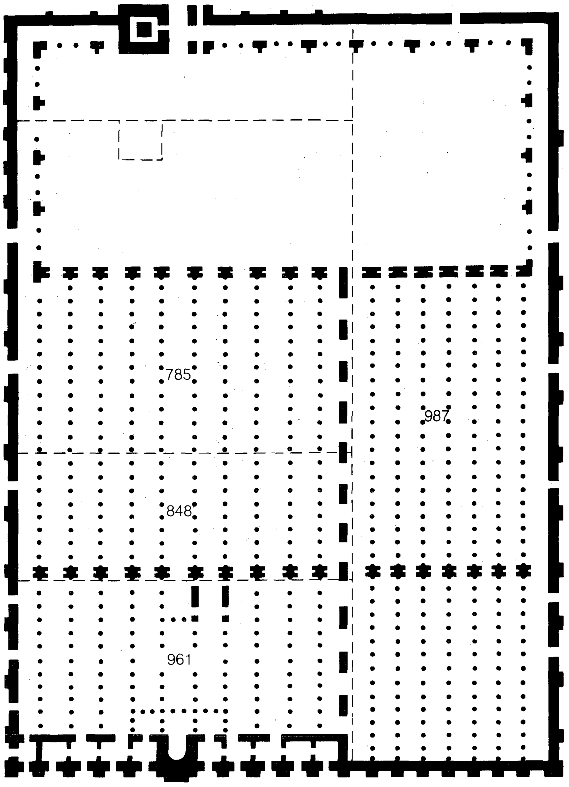
Either in its simplest form, as in Medina, or in its more formalized shape, as in Damascus, the hypostyle tradition dominated mosque architecture from 715 to the 10th century. As it occurs at Nīshāpūr in northeastern Iran, Sīrāf in southern Iran, al-Qayrawān (Kairouan) in Tunisia, and Córdoba in Spain, it can indeed be considered as the classic early Islāmic type. Its masterpieces occur in Iraq and in the West. The monumentalization of the early Iraqi hypostyle is illustrated by the two ruined structures in Sāmarrā', with their enormous sizes (790 by 510 feet [240 by 156 metres] for one and 700 by 440 feet [213 by 135 metres] for the other), their multiple entrances, their complex piers, and, in one instance, a striking separation of the *qiblah* area from the rest of the building. The best preserved example of this type is the mosque of Ibn Ṭūlūn at Cairo (876–879), where a semi-independent governor, Aḥmad ibn Ṭūlūn, introduced Iraqi techniques and succeeded in creating a masterpiece of composition.

Two classic examples of early mosques in the western Islāmic world of interest are preserved in Tunisia and Spain. In al-Qayrawān the Great Mosque was built in stages between 836 and 866. Its most striking feature is the formal emphasis on the building's T-like axis punctuated by two domes, one of which hovers over the earliest preserved ensemble of *miḥrāb, minbar,* and *maqṣūrah.* At Córdoba the earliest section of the Great Mosque was built in 785–786. It consisted simply of 11 naves with a wider central one and a court. It was enlarged twice in length, first between 833 and 855 and again from 961 to 965 (it was in the latter phase that the celebrated *maqṣūrah* and *miḥrāb,* comprising one of the great architectural ensembles of early Islāmic art, were constructed). Finally, in 987–988 an extension of the mosque was completed to the east that increased its size by almost one-third without destroying its stylistic unity. The constant increases in the size of this mosque are a further illustration of the flexibility of the hypostyle and its adaptability to any spatial requirement. The most memorable aspects of the Córdoba mosque, however, lie in its construction and decoration. The particularly extensive and heavily decorated

Carl Frank—Photo Researchers



The Great Mosque in al-Qayrawān, Tunisia, built between 836 and 866.



Plan of the Great Mosque at Córdoba, Spain, showing dates of different additions.
From G. Marcais, *L'Architecture Musulmane D'Occident*

miḥrāb area exemplifies a development that started with the Medina mosque and would continue: an emphasis on the *qiblah* wall.

Early aberrant plans

Although the hypostyle mosque was the dominant plan, it was not the only one. From very early Islāmic times, a fairly large number of aberrant plans also occur. Most of them were built in smaller urban locations or were secondary mosques in larger Muslim cities. It is rather difficult, therefore, to evaluate whether their significance was purely local or whether they were important for the tradition as a whole. Since a simple type of square subdivided by four piers into nine-domed units occurs at Balkh in Afghanistan, at Cairo, and at Toledo, it may be considered a pan-Islāmic type. Other types, a single square hall surrounded by an ambulatory, or a single long barrel-vault parallel or perpendicular to the *qiblah,* are rarer and should perhaps be considered as purely local. These are particularly numerous in Iran, where it does seem that the mainstream of early Islāmic architecture did not penetrate very deeply. Unfortunately, the archaeological exploration of Iran is still in its infancy, and many of the mud-brick buildings from the early Islāmic period have been destroyed or rebuilt beyond recognition. As a result, it is extremely difficult to determine the historical importance of monuments found at Neyrīz, Moḥammadīyeh (near Nāʾīn), Fahraj (near Yazd), or Hazareh (near Samarkand). For an understanding of the mosque's development and of the general dynamics of Islāmic architecture however, an awareness of these secondary types, which may have existed outside of Iran as well, is essential.

The function of the mosque, the central gathering place of the Muslim community, became the major and most original completely Muslim architectural effort. The mosque was not a purely religious building, at least not at the beginning; but, because it was restricted to Muslims, it is appropriate to consider it as such. This, however, was not the only type of early Islāmic building to be uniquely Muslim. Three other types can be defined architecturally, and a fourth one only functionally.

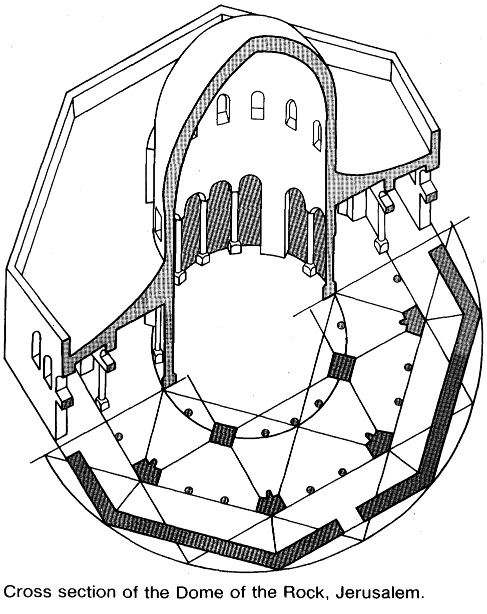
The first type, the Dome of the Rock in Jerusalem, is a unique building. Completed in 691, this masterwork of Islāmic architecture is the earliest major Islāmic monument. Its octagonal plan, use of a high dome, and building techniques are hardly original, although its decoration is unique. Its purpose, however, is what is most remarkable about the building. Since the middle of the 8th century, the Dome of the Rock has become the focal centre of the most mystical event in the life of the Prophet: his ascension into heaven from the rock around which the building was erected. According to an inscription preserved since the erection of the dome, however, it would seem that the building did not originally commemorate the Prophet's ascension but rather the Christology of Islām and its relationship to Judaism. It seems preferable, therefore, to interpret the Dome of the Rock as a victory monument of the new faith's ideological and religious claim on a holy city and on all the religious traditions attached to it.

The *ribāṭs* of the frontier

The second distinctly Islāmic type of religious building is the little-known *ribāṭ*. As early as in the 8th century, the Muslim empire entrusted the protection of its frontiers, especially the remote ones, to warriors for the faith (*murābiṭūn,* "bound ones") who lived, permanently or temporarily, in special institutions known as *ribāṭ*s. Evidence for these exist in Central Asia, Anatolia, and North Africa. It is only in Tunisia that *ribāṭ*s have been preserved. The best one is at Sūsah, Tunisia; it consists of a square fortified building with a single fairly elaborate entrance and a central courtyard. It has two stories of private or communal rooms. Except for the prominence taken by an oratory, this building could be classified as a type of Muslim secular architecture. Since no later example of a *ribāṭ* is known, there is some uncertainty as to whether the institution ever acquired a unique architectural form of its own.

The diverse functions of the Muslim mausoleum

The last type of religious building to develop before the end of the 10th century is the mausoleum. Originally Islām was strongly opposed to any formal commemoration of the dead. But three independent factors slowly modified an attitude that was eventually maintained only in the most strictly orthodox circles. One factor was the growth of the Shīʿite heterodoxy, which led to an actual cult of the descendants of the Prophet through his son-in-law ʿAlī. The second factor was that, as Islām strengthened its hold on conquered lands, a wide variety of local cultic practices and especially the worship of certain sacred places began to affect the Muslims, resulting in a whole movement of Islāmization of ancient holy places by associating them with deceased Muslim heroes and holy men or with prophets. The third factor is not, strictly speaking, religious, but it played a major part. As more or less independent local dynasties began to grow, they sought to commemorate themselves through mausoleums. Not many mausoleums have remained from these early centuries, but literary evidence is clear on the fact that the Shīʿite sanctuaries of Karbalāʾ and an-Najaf, both in Iraq,



Cross section of the Dome of the Rock, Jerusalem.

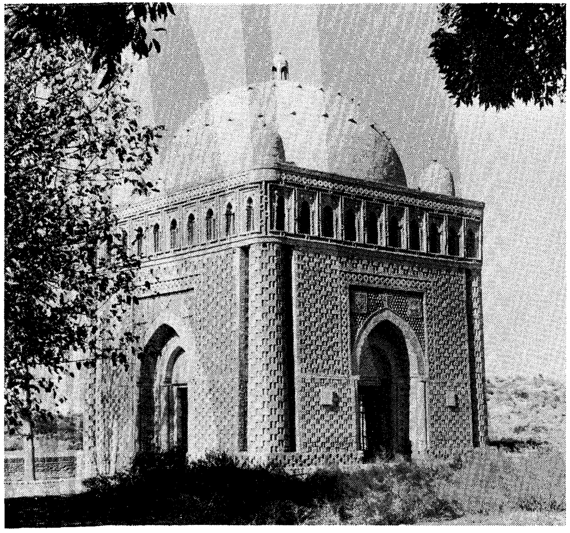
and Qom, Iran, already possessed monumental tombs. At Sāmarrā' an octagonal mausoleum had been built for three caliphs. The masterpieces of early funerary architecture occur in Central Asia, such as the royal mausoleum of the Sāmānids (known incorrectly as the mausoleum of Esmā'īl the Sāmānid) at Bukhara (before 942), which is a superb example of Islāmic brickwork. In some instances a quasi-religious character was attached to the mausoleums, such as the one at Tim (976), which already has the high facade typical of so many later monumental tombs. In all instances the Muslims took over or rediscovered the ancient tradition of the centrally planned building as the characteristic commemorative structure.

The *madrasah*

The fourth kind of Muslim building is the *madrasah,* an institution for religious training set up independently of mosques. It is known from texts that such privately endowed schools existed in the northeastern Iranian world as early as in the 9th century, but no description exists of how they were planned or looked.

Secular architecture. Whereas the functions of the religious buildings of early Islām could not have existed without the new faith, the functions of secular Muslim architecture have *a priori* no specifically Islāmic character. This is all the more so since one can hardly point to a significant new need or habit that would have been brought from Arabia by the conquering Muslims and since so little was destroyed in the conquered areas. It can be assumed, therefore, that all pre-Islāmic functions such as living, trading, and manufacturing continued in whatever architectural setting they may have had. Only one exception is certain. With the disappearance of Sāsānian kingship, the pre-Islāmic Iranian imperial tradition ceased, and elsewhere conquered minor kings and governors left their palaces and castles. A new imperial power was created, located first in Damascus, then briefly in the northern Syrian town of ar-Ruṣāfah, and eventually in Baghdad and Sāmarrā' in Iraq. New governors and, later, almost independent princes took over provincial capitals, which were sometimes old seats of government and, at other times, were new Muslim centres. In all instances, however, there is no reason to assume that for an architecture of power or of pleasure early Muslims would have felt the need to modify pre-Islāmic traditions. In fact there is much in early Islāmic secular architecture that can be used to illustrate secular arts elsewhere—in Byzantium, for example, or even in the West. If any new political or social entity is to succeed in preserving an identity of its own, however, it must give to its secular needs certain directions and emphases that will eventually establish a unique cultural image. This is what happened in the development of Umayyad and early ʿAbbāsid secular architecture.

Three factors contributed to the evolution of a new secular architecture. One was that the accumulation of an immense wealth of ideas, workers, and money in the hands of the Muslim princes settled in Syria and Iraq gave rise to a unique palace architecture. The second factor was the impetus given to urban life and to trade. New cities were founded from Sijilmāssah on the edge of the Moroccan Sahara to Nīshāpūr in northeastern Iran, and 9th-century Arab merchants traded as far away as China. Thus the second topic, to be treated below, will be the urban design and commercial architecture. The third factor is that, for the first time since Alexander the Great, a world extending from the Mediterranean to India became culturally unified. As a result, decorative motifs, design ideas, structural techniques, and artisans and architects—which until then had belonged to entirely different cultural traditions—were available in the same places. Early Islāmic princely architecture has become the best known and most original aspect of early Islāmic secular buildings.



Josephine Powell, Rome

Royal mausoleum of the Sāmānids, Bukhara, Uzbek S.S.R., (before 942).

The country princely palace complex

There are basically three kinds of these princely structures. The first type consists of 10 large rural princely complexes found in Syria, Palestine, and Transjordan dating from around 710 to 750: ar-Ruṣāfah, Qaṣr al-Ḥayr East, Qaṣr al-Ḥayr West, Jabal Says, Khirbat Minyah, Khirbat al-Mafjar, Mshattā, Qaṣr 'Amrah, Qaṣr al-Kharānah, and Qaṣr aṭ-Ṭūbah. Apparently these examples of princely architecture belong to a group of more than 60 ruined or only textually identifiable rural complexes erected by Umayyad princes. In the past a romantic theory had developed about their locations, suggesting that the remoteness of their sites expressed an atavistic hankering on the part of the Umayyad Arab rulers for the desert or at least the semiarid steppe that separates the permanently cultivated areas of Syria and Palestine from their original home in the north Arabian wilderness. This theory has been disproved, for every one of these has turned out to have been a major agricultural or trade centre, some of which were developed even before the Muslim conquest. Private palaces were built, notably at ar-Ruṣāfah, Qaṣr al-Ḥayr West, Khirbat al-Mafjar, Qaṣr ʿAmrah, and Mshattā. These must be considered as early medieval equivalents of the *villae rusticae* so characteristic in the ancient Roman period. Although each of these had a number of idiosyncrasies that were presumably inspired by the needs and desires of its owner, all of these structures tend to share a number of features that can best be illustrated by Khirbat al-Mafjar.

This palace, the richest of them all, contained a residential unit consisting of a square building with an elaborate entrance, a porticoed courtyard, and a number of rooms or halls arranged on two floors. Few of these rooms seem to have any identifiable function, although at Khirbat al-Mafjar a private oratory, a large meeting hall, and an anteroom leading to a cool underground pool have been identified. The main throne room was on the second floor above the entrance. Its plan is not known but probably resembled the preserved throne rooms or reception halls at Qaṣr 'Amrah and Mshattā, which consisted of a three-aisled hall ending in an apse (semicircular or polygonal domed projection) in the manner of a Roman basilica.

Next to an official residence, there usually was a small mosque, generally a miniaturized hypostyle in plan. The most original feature of these establishments was the bath. The bathing area itself is comparatively small, but every bath had its own elaborate entrance and contained a large hall that, at least in the instance of Khirbat al-Mafjar, was heavily decorated and of an unusual shape. It would appear that these halls were for pleasure—places for music, dancing, and probably occasional orgies. In some instances, as at Qaṣr 'Amrah, the same setting may have been used for both pleasure and formal receptions.

These palaces are important illustrations of the luxurious taste and way of life of the new Near Eastern aristocrats, who settled in the countryside and transformed some of it into places of pleasure. This aspect of these establishments is peculiar to the Umayyad dynasty in Syria and Palestine. Outside of this area and period only one comparable structure has been found—at Ukhayḍir in Iraq, which dates from the early 'Abbāsid period. A number of princely residences of the Central Asian or North African countryside are still too little known but appear not to have had the same development. The other important lesson to draw from them is that few of their features are original. All of them derive from the architectural vocabulary of pre-Islāmic times, and it is in the artistic traditions of the Mediterranean world that most of their sources are found, although the Mshattā throne room does have a number of Sāsānian elements. For this reason these palaces should be considered as major examples of pre-Islāmic secular architecture, for as interesting as these monuments are, they are not part of the Islāmic tradition.

The urban palace

A second type of princely architecture—the urban palace—has been preserved only in texts or literary sources, with the exception of the palace at Kūfah in Iraq. Datable from the very end of the 7th century, this example of princely architecture seems to have functioned both as a residence and as the *dār al-imārah,* or centre of government. This dual function is reflected in the use of separate building units and in the absence of much architectural decoration, which suggests that it reflected an austere official taste. Although suggestions concerning the plans used are occasionally encountered in literary sources, this information is not sufficient to define these early urban official buildings of the Muslims. Nothing is known, for instance, about the great Umayyad palace in Damascus aside from the fact that it had a green dome.

Also poorly documented is a development in urban aristocratic buildings that seems to have begun with the ʿAbbāsids during the last decades of the 8th century. This involved the construction of smaller palaces, probably pavilions in the midst of gardens in or around major cities.

The ʿAbbāsid palace-city

The third type of early Islāmic princely architecture is the palace-city. Several of these huge palaces are part of the enormous mass of ruins at Sāmarrā', the temporary ʿAbbāsid capital from 838 to 883. Jawsaq al-Khāqānī, for instance, is a walled architectural complex nearly one mile to a side that in reality is an entire city. It contains a formal succession of large gates and courts leading to a cross-shaped throne room, a group of smaller living units, basins and fountains, and even a racetrack. Too little is known about the architectural details of these huge walled complexes to lead to more than very uncertain hypotheses. Their existence, however, suggests that they were settings for the very elaborate ceremonies developed by the ʿAbbāsid princes, especially when receiving foreign ambassadors. An account, for instance, in Khaṭīb al-Baghdādī's (died 1071) *Ta'rīkh Baghdad* ("History of Baghdad") of the arrival in Baghdad of a Byzantine envoy in 914 illustrates this point. The meeting with the caliph was preceded by a sort of formal presentation intended to impress the ambassador with the Muslim ruler's wealth and power. Treasures were laid down, thousands of soldiers and slaves in rich clothes guarded them, lions roared in the gardens, and on gilded artificial trees mechanical devices made silver birds chirp. The ceremony was a fascinating mixture of a traditional attempt to recreate paradise on earth and a rather vulgar exhibition of wealth that required a huge space, as in the Sāmarrā' palaces. Another important aspect of these palace-cities is that they became part of a myth. The walled enclosure in which thousands lived a life unknown to others and into which simple mortals did not penetrate without bringing their own shroud was transformed into legend. It became the mysterious City of Brass of *The Thousand and One Nights,* and it is from its luxurious glory that occasionally a caliph such as Hārūn ar-Rashīd escaped into the "real" world. Even though there is inadequate information on the ʿAbbāsid palace-city, it was clearly a unique early Islāmic creation, and its impact can be detected from Byzantium to Hollywood.

Urban design

Islāmic secular architecture has left considerable information about cities, for systematic urbanization was one of the most characteristic features of early Muslim civilization. It is much too early to draw any sort of conclusion about the actual physical organization of towns, about their subdivisions and their houses, for only at al-Fusṭāṭ (Cairo) and Sīrāf in Iran is the evidence archaeologically clear, and much of it has not yet been properly published. A huge task remains to be done of relating immense amounts of textual material with scraps of archaeological information scattered from Central Asia to Spain, such as the outer walls and impressive gateway preserved at ar-Raqqah in Syria. In general it can be said that there does not seem to have been any idealized master plan for the internal arrangement of an urban site in contradistinction to Hellenistic or Roman towns. Even mosques or palaces were often located eccentrically and not in the middle of the town. Extraordinary attention was paid to water distribution and conservation, as demonstrated by the magnificent 9th-century cisterns in Tunisia, the 9th-century Nilometer (a device to measure the Nile's level) in Cairo, and the elaborate dams, canals, and sluices of Qaṣr al-Ḥayr in Syria. The construction of commercial buildings on a monumental scale occurred. The most spectacular example is the caravansary of Qaṣr al-Ḥayr East, with its magnificent gate.

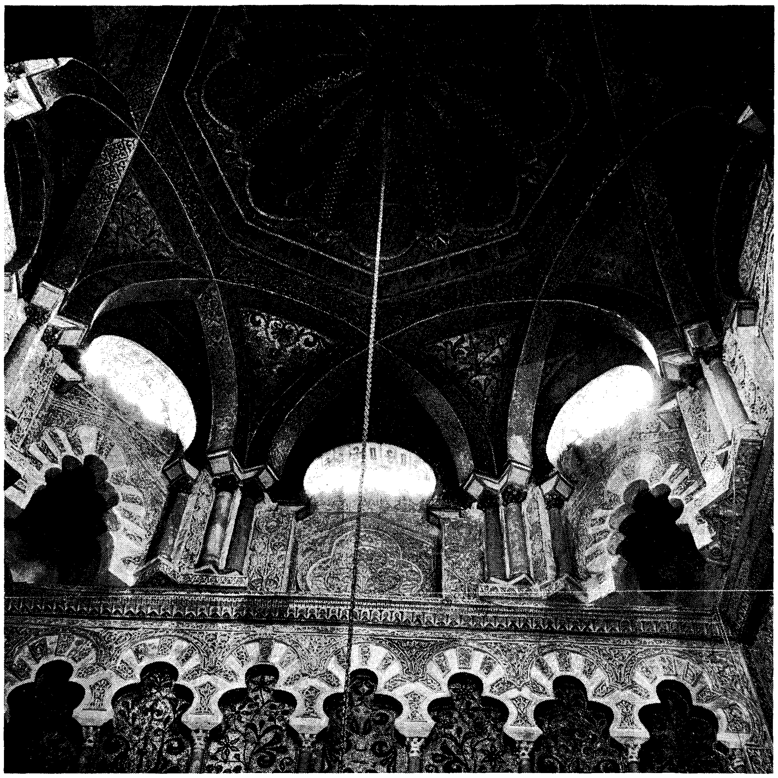
The concern for palaces and cities that characterized early Islāmic secular architecture shows itself most remarkably in the construction of Baghdad between 762 and 766–767 by the ʿAbbāsid caliph al-Manṣūr. It was a walled round city whose circular shape served to demonstrate Baghdad's symbolic identity as the navel of the universe. A thick ring of residential quarters was separated by four axial, commercial streets entered through spectacular gates. In the centre of the city there was a large open space with a palace, a mosque, and a few administrative buildings. By its size and number of inhabitants, Baghdad was unquestionably a city; however, its plan so strongly emphasized the presence of the caliph that it was also a palace.

Building materials and technology. The early Islāmic period, on the whole, did not innovate much in the realm of building materials and technology but utilized what it had inherited from older traditions. Stone and brick continued to be used around the Mediterranean, while mud brick usually covered with plaster predominated in Iraq and Iran, with a few notable exceptions like Sīrāf, where a masonry of roughly cut stones set in mortar was more common. The most important novelty was the rapid development in Iraq of a baked brick architecture in the late 8th and 9th centuries. Iraqi techniques were later used in Syria at ar-Raqqah and Qaṣr al-Ḥayr East and in Egypt. Iranian brickwork appears at Mshattā in Jordan. The mausoleum of the Sāmānids in Bukhara is the earliest remaining example of the new brick architecture in northeastern Iran. Wood was used consistently but has usually not been very well preserved, except in Palestine and Egypt where climatic (extreme dryness of Egypt), religious (holiness of Jerusalem sanctuaries), or historic (Egypt was never conquered) factors contributed to the continuous upkeep of wooden objects or architectural elements.

As supports for roofs and ceilings, early Islāmic architecture used walls and single supports. Walls were generally continuous, often buttressed with half towers, and rarely (with exceptions in Central Asia) were they articulated or broken by other architectural features. The most common single support was the base–column–capital combination of Mediterranean architecture. Most columns and capitals were either reused from pre-Islāmic buildings or were directly imitated from older models. In the 9th century in Iraq a brick pier was used, a form that spread to Iran and Egypt. Columns and piers were covered with arches. Most often these were semicircular arches; the pointed, or two-centred, arch was known, but it does not seem that its property of reducing the need for heavy supports had been realized. The most extraordinary technical development of arches occurs in the Great Mosque at Córdoba, where, in order to increase the height of the building in an area with only short columns, the architects created two rows of superimposed horseshoe arches. Almost immediately they realized that such a succession of superimposed arches constructed of alternating stone and brick could be modified to create a variety of patterns that would alleviate the inherent monotony of a hypostyle building. A certain ambiguity remains, however, as to whether ornamental effect or structural technology was the predominate concern in the creation of these unique arched columns.

Arches and vaulting

The majority of early Islāmic ceilings were flat. Gabled wooden roofs, however, were erected in the Muslim world west of the Euphrates and simple barrel vaults to the east. Vaulting, either in brick or in stone, was used, especially in secular architecture. Domes were employed frequently



Dome of the *miḥrāb* in the Great Mosque of Córdoba, Spain, *c.* 961.
Fritz Henle—Photo Researchers/EB Inc.

in mosques, consistently in mausoleums, and occasionally in secular buildings. Almost all domes are on squinches (supports carried across corners to act as structural transitions to a dome). Most squinches, as in the al-Qayrawān domes, are classical Greco-Roman niches, which transform the square room into an octagonal opening for the dome. In Córdoba's Great Mosque a complex system of intersecting ribs is encountered, while at Bukhara the squinch is broken into halves by a transverse half arch. The most extraordinary use of the squinch occurs in the mausoleum at Tim, where the surface of this structural device is broken into a series of smaller three-dimensional units rearranged into a sort of pyramidal pattern. This rearrangement is the earliest extant example of *muqarnas,* or stalactite-like decoration that would later be an important element of Islāmic architectural ornamentation. The motif is so awkwardly constructed at Tim that it must have derived from some other source, possibly the ornamental device of using curved stucco panels to cover the corners and upper parts of walls found in Iran at Nīshāpūr.

Architectural decoration. Early Islāmic architecture is most original in its decoration. Mosaics and wall paintings followed the practices of antiquity and were primarily employed in Syria, Palestine, and Spain. Stone sculpture existed, but stucco sculpture, first limited to Iran, spread rapidly throughout the early Islāmic world. Not only were stone or brick walls covered with large panels of stucco sculpture, but this technique was used for sculpture in the round in the Umayyad palaces of Qaṣr al-Ḥayr West and Khirbat al-Mafjar. The latter was a comparatively short-lived technique, although it produced some of the few instances of monumental sculpture anywhere in the early Middle Ages. A variety of techniques borrowed from the industrial arts were used for architectural ornamentation. The *miḥrāb* wall of al-Qayrawān's Great Mosque, for example, was covered with ceramics, while fragments of decorative woodwork have been preserved in Jerusalem and Egypt.

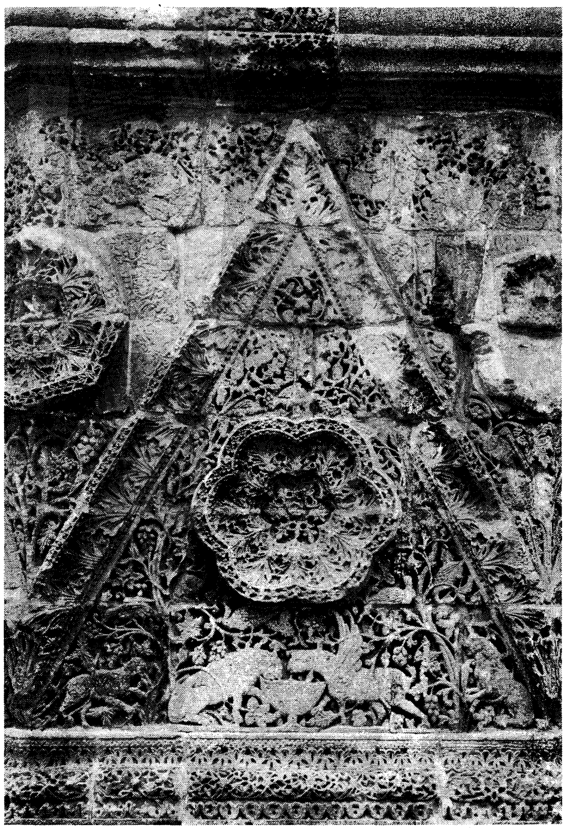
The themes and motifs of early Islāmic decoration can be divided into three major groups. The first kind of ornamentation simply emphasizes the shape or contour of an architectural unit. The themes used were vegetal bands for vertical or horizontal elements, marble imitations for the lower parts of long walls, chevrons or other types of borders on floors and domes, and even whole trees on the spandrels or soffits (undersides) of arches as in the Umayyad Mosque of Damascus or the Dome of the Rock; all these motifs tend to be quite traditional, being taken from the rich decorative vocabularies of pre-Islāmic Iran or of the ancient Mediterranean world.

Three kinds of early architectural decoration

The second group consists of decorative motifs for which a concrete iconographic meaning can be given. In the Dome of the Rock and the Umayyad Mosque of Damascus, as well as possibly the mosques of Córdoba and of Medina, there were probably iconographic programs. It has been shown, for example, that the huge architectural and vegetal decorative motifs at Damascus were meant to symbolize a sort of idealized paradise on earth, while the crowns of the Jerusalem sanctuary are thought to have been symbols of empires conquered by Islām. But it is equally certain that this use of visual forms in mosques for ideological and symbolic purposes was not easily accepted, and most later mosques are devoid of iconographically significant themes. The only exceptions fully visible are the Qur'ānic inscriptions in the mosque of Ibn Ṭūlūn at Cairo, which were used both as a reminder of the faith and as an ornamental device to emphasize the structural lines of the building. Thus the early Islāmic mosque eventually became austere in its use of symbolic ornamentation, with the exception of the *miḥrāb,* which was considered as a symbol of the unity of all believers.

Like religious architecture, secular buildings seem to have been less richly decorated at the end of the early Islāmic period than at the beginning. The paintings, sculptures, and mosaics of Qaṣr al-Ḥayr West, Khirbat al-Mafjar, Qaṣr ʿAmrah, and Sāmarrāʾ primarily illustrated the life of the prince. There were official iconographic compositions, such as the monarch enthroned, or ones of pleasure and luxury, such as hunting scenes or depictions of the prince surrounded by dancers, musicians, acrobats, and unclad women. Few of these so-called princely themes were iconographic inventions of the Muslims. They usually can be traced back either to the classical world of ancient Greece and Rome or to pre-Islāmic Iran and Central Asia.

The third type of architectural decoration consists of large panels, most often in stucco, for which no meaning or



Triangle stone relief from the facade of Mshattā in Transjordan, early 8th century. In the Islamisches Museum, Staatliche Museen zu Berlin.

By courtesy of the Islamisches Museum, Staatliche Museen zu Berlin

interpretation is yet known. These panels might be called ornamental in the sense that their only apparent purpose was to beautify the buildings in which they were installed, and their relationship to the architecture is arbitrary. The Mshattā facade's decoration of a huge band of triangles is, for instance, quite independent of the building's architectural parts. Next to Mshattā, the most important series of examples of the third type of ornamentation come from Sāmarrā', although striking examples are also to be found at Khirbat al-Mafjar, Qaṣr al-Ḥayr East and West, al-Fusṭāṭ, Sīrāf, and Nīshāpūr. Two decorative motifs were predominately used on these panels: a great variety of vegetal motifs and geometric forms. At Sāmarrā' these panels eventually became so abstract that individual parts could no longer be distinguished, and the decorative design had to be viewed in terms of the relationships between line and shape, light and shade, horizontal and vertical axes, and so forth. Copied consistently from Morocco to Central Asia, the aesthetic principles of this latter type of a complex overall design influenced the development of the principle of arabesque ornamentation.

Islāmic architectural ornamentation does not lend itself easily to chronological stylistic definition. In other words, it does not seem to share consistently a cluster of formal characteristics. The reason is that in the earliest Islāmic buildings the decorative motifs were borrowed from an extraordinary variety of stylistic sources: classical themes illusionistically rendered (*e.g.*, the mosaics of the Umayyad Mosque of Damascus), hieratic Byzantine themes (*e.g.*, the Umayyad Mosque of Damascus and Qaṣr ʿAmrah), Sāsānian motifs, Central Asian motifs (especially the sculpture from Umayyad palaces), and the many regional styles of ornamentation that had developed in all parts of the pre-Islāmic world. It is the wealth of themes and motifs, therefore, that constitutes the Umayyad style of architectural decoration. The ʿAbbāsids, on the other hand, began to be more selective in their choice of ornamentation.

Decorative arts. Very little is known about early Islāmic gold and silver objects, although their existence is mentioned in many texts as well as suggested by the wealth of the Muslim princes. Except for a large number of silver plates and ewers belonging to the Sāsānian tradition, nothing has remained. These silver objects were probably made for Umayyad and ʿAbbāsid princes, although there is much controversy among scholars regarding their authenticity and date of manufacture.

Būyid silks

For entirely different reasons it is impossible to present any significant generalities about the art of textiles in the early Islāmic period. Problems of authenticity are few. Dating from the 10th century are a large number of Būyid silks, a group of funerary textiles with plant and animal motifs as well as poetic texts. Very little order has yet been made of an enormous mass of often well-dated textile fragments, and therefore, except for the Būyid silks, it is still impossible to identify any one of the textile types mentioned in early medieval literary sources. Furthermore, since it can be assumed that pre-Islāmic textile factories were taken over by the Muslims and since it is otherwise known that textiles were easily transported from one area of the Muslim world to the other or even beyond it, it is still very difficult to define Islāmic styles as opposed to Byzantine or to Coptic ones. The obvious exception lies in those fragments that are provided with inscriptions, and the main point to make is therefore that one of the characteristic features of early Islāmic textiles is their use of writing for identifying and decorative purposes. But, while true, this point in no way makes it possible to deny an Islāmic origin to fragments that are not provided with inscriptions, and thus one must await further investigations of detail before being able to define early Islāmic textiles.

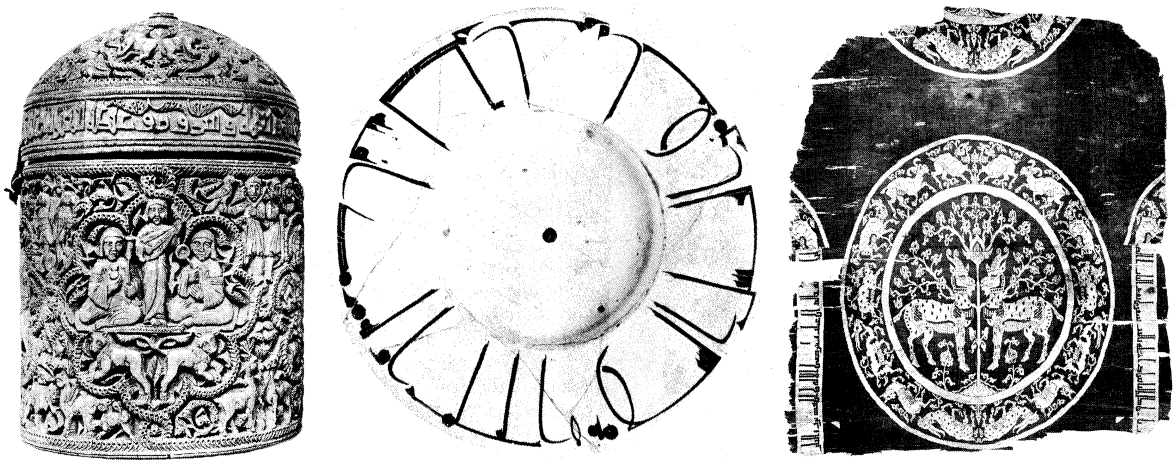
Early ceramics

The most important medium of early Islāmic decorative arts is pottery. Initially Muslims continued to sponsor whatever varieties of ceramics had existed before their arrival. Probably in the last quarter of the 8th century new and more elaborate types of glazed pottery were produced. This new development did not replace the older and simpler types of pottery but added a new dimension to the art of Islāmic ceramics. Because of the still incompletely published studies on the unfinished excavations carried out at Nīshāpūr, Sīrāf, Qaṣr al-Ḥayr East, and al-Fusṭāṭ, the scholarship on these ceramics is likely to be very much modified. Therefore, this section will treat only the most general characteristics of Islāmic ceramics, avoiding in particular the complex archaeological problems posed by the growth and spread of individual techniques.

The area of initial technical innovation seems to have been Iraq. Trade with Central Asia brought Chinese ceramics to Mesopotamia, and Islāmic ceramicists sought to imitate them. It is probably in Iraq, therefore, that the technique of lustre glazing was first developed in the Muslim world. This gave the surface of a clay object a metallic, shiny appearance. Egypt also played a leading part in the creation of the new ceramics. Since the earliest datable lustre object (a glass goblet with the name of the governor who ruled in 773, now in the Cairo Museum of Islamic Art) was Egyptian, some scholars feel that it was in Egypt and not Iraq that lustre was first used. Early pottery was also produced in northeastern Iran, where excavations at Afrāsiyāb (Samarkand) and Nīshāpūr have brought to light a new art of painted underglaze pottery. Its novelty was not so much in the technique of painting designs on the slip and covering them with a transparent glaze as in the variety of subjects employed.

While new ceramic techniques may have been sought to imitate other mediums (mostly metal) or other styles of pottery (mostly Chinese), the decorative devices rapidly became purely and unmistakably Islāmic in style. A wide variety of motifs were combined: vegetal arabesques or single flowers and trees; inscriptions, usually legible and consisting of proverbs or of good wishes; animals that were usually birds drawn from the vast folkloric past of the Near East; occasionally human figures drawn in a strikingly abstract fashion; geometric designs; all-over abstract patterns; single motifs on empty fields; and simple splashes of colour, with or without underglaze sgraffito designs (i.e., designs incised or sketched on the body or the slip of the object). All of these motifs were used on both the high-quality ceramics of Nīshāpūr and Samarkand as well as on Islāmic folk pottery.

Although ceramics has appeared to be the most character-



Early Islāmic decorative arts.
(Left) Ivory casket made for al-Mughīrah, son of ʿAbd ar-Raḥmān III, from Córdoba, Spain, 968. In the Louvre, Paris. Ht. 15 cm. (Centre) Bowl of Samarkand ware with calligraphic decoration, 10th century. In the Louvre, Paris. Diameter 37.5 cm. (Right) Fragment of a silk tomb cover with a woven design of pairs of ibex, from Iran, 998. In the Cleveland Museum of Art, Ohio. 78.1 cm × 66 cm.
By courtesy of (left, centre) the Musee du Louvre, Paris, (right) the Cleveland Museum of Art, Purchase from the J.H. Wade Fund; photographs, (left) Mansell—Giraudon from Art Resource, (centre) Cliche Musees Nationaux, Paris

Spanish ivory carving

istic medium of expression in the decorative arts during the early Islāmic period, it has only been because of the greater number of preserved objects. Glass was as important, but examples have been less well preserved. A tradition of ivory carving developed in Spain, and the objects dating from the last third of the 10th century onward attest to the high quality of this uniquely Iberian art. Many of these carved ivories certainly were made for princes; therefore it is not surprising that their decorative themes were drawn from the whole vocabulary of princely art known through Umayyad painting and sculpture of the early 8th century. These ivory carvings are also important in that they exemplify the fact that an art of sculpture in the round never totally disappeared in the Muslim world—at least in small objects.

Assessment. There are three general points that seem to characterize the art of the early Islāmic period. It can first be said that it was an art that sought self-consciously, like the culture sponsoring it, to create artistic forms that would be identifiable as being different from those produced in preceding or contemporary non-Islāmic artistic traditions. At times, as in the use of the Greco-Roman technique of mosaics or in the adoption of Persian and Roman architectural building technology, early Islāmic art simply took over whatever traditions were available. At other times, as in the development of the mosque as a building type, it recomposed into new shapes the forms that had existed before. On the other hand, in ceramics or the use of calligraphic ornamentation, the early Islāmic artist invented new techniques and a new decorative vocabulary. Whatever the nature of the phenomenon, it was almost always an attempt to identify itself visually as unique and different. Since there was initially no concept about what should constitute an Islāmic tradition in the visual arts, the early art of the Muslims often looks like only a continuation of earlier artistic styles, forms, subjects, and techniques. Many mosaics, silver plates, or textiles, therefore, were not considered to be Islāmic until recently. In order to be understood, then, as examples of the art of a new culture, these early buildings and objects have to be seen in the complete context in which they were created. When so seen they appear as conscious choices by the new Islāmic culture from its immense artistic inheritance.

Question of an early Islāmic style

A second point of definition concerns the question of whether there is an early Islāmic style or perhaps even several styles in some sort of succession. The fascinating fact is that there is a clear succession only in those artistic features that are Islāmic inventions—nonfigurative ornament and ceramics. For it is only in development of these features that one can assume to find the conscious search for form that can create a period style. Elsewhere, especially in palace art, the Muslim world sought to relate itself to an earlier and more universal tradition of princely art; its monuments, therefore, are less Islāmic than typological. In the new art of the Muslim bourgeoisie, however, uniquely Islāmic artistic phenomena began to evolve.

Finally, the geographical peculiarities of early Islāmic art must be reiterated. Its centres were Syria, Iraq, Egypt, northwestern Iran, and Spain. Of these, Iraq was probably the most originally creative, and it is from Iraq that a peculiarly Islāmic visual koine (a commonly accepted and understood system of forms) was derived and spread throughout the Islāmic world. This development, of course, is logical since the capital of the early empire and some of the first purely Muslim cities were in Iraq. In western Iran, in Afghanistan, in northern Mesopotamia, and in Morocco the more atypical and local artistic traditions were more or less affected by the centralized imperial system of Iraq. This tension between a general pan-Islāmic vocabulary and a variable number of local vocabularies was to remain a constant throughout the history of Islāmic art and is certainly one of the reasons for the difficulty, if not impossibility, one faces in trying to define an Islāmic style.

MIDDLE PERIOD

The middle period in the development of Islāmic art extends roughly from the year 1000 to 1500, when a strong central power with occasional regional political independence was replaced by a bewildering mosaic of overlapping dynasties. Ethnically this was the time of major Turkish and Mongol invasions that brought into the Muslim world new peoples and institutions. At the same time, Berbers, Kurds, and Iranians, who had been within the empire from the beginning of Islām, began to play far more effective historical and cultural roles, shortlived for the Kurds, but uniquely important for the Iranians. Besides political and ethnic confusion, there was also religious and cultural confusion during the middle period. The 10th century, for example, witnessed the transformation of the Shīʿite heterodoxy into a major political and possibly cultural phenomenon, while the extraordinary development taken by the personal and social mysticism known as Ṣūfism modified enormously the nature of Muslim piety. Culturally the most significant development was perhaps that of Persian literature as a highly original new verbal expression existing alongside the older Arabic literary tradition. Finally, the middle period was an era of expansion in all areas except Spain, which was completely lost to the Muslims in 1492 with the conquest of the Kingdom of Granada by Ferdinand II and Isabella. Anatolia and the Balkans, the Crimea, much of Central Asia and northern India, and parts of eastern Africa all became new Islāmic provinces. In some cases this expansion was the result of conquests, but in others it had been achieved through missionary work.

The immense variety of impulses that affected the Muslim world during these five centuries was one of the causes of the bewildering artistic explosion that also characterizes the middle period. Although much work has been done on individual monuments, scholarship is still in its infancy. It is particularly difficult, therefore, to decide on the appropriate means of organizing this information: by geographical or cultural areas (*e.g.,* Iran, Egypt, Morocco), by individual dynasties (*e.g.,* Seljuqs, Timurids), by periods (*e.g.,* 13th century before the Mongol invasions), or even by social categories (*e.g.,* the art of princes, the art of cities). Thus, the five following divisions of Fāṭimid, Seljuq, Western Islāmic, Mamlūk, and Mongol Iran (Il-Khanid and Timurid) art are partly arbitrary and to a large extent tentative. Their respective importance also varies, for what is known as Seljuq art certainly overwhelms almost all others in its importance.

Fāṭimid art (909–1171). The Fāṭimids were technically an Arab dynasty professing with missionary zeal the beliefs of the Ismāʿīlī sect of the Shīʿite branch of Islām. The dynasty was established in Tunisia and Sicily in 909. In 969 the Fāṭimids moved to Egypt and founded the city of Cairo. They soon controlled Syria and Palestine. In the latter part of the 11th century, however, the Fāṭimid empire began to distintegrate internally and externally; the final demise occurred in 1171. But it is not known which of the obvious components of the Fāṭimid world was more significant in influencing the development of the visual arts: its heterodoxy, its Egyptian location, its missionary relationship with almost all provinces of Islām, or the fact that during its heyday in the 11th century it was the only wealthy Islāmic centre and could thus easily gather artisans and art objects from all over the world.

Architecture. The great Fāṭimid mosques of Cairo—al-Azhar (started in 970) and al-Ḥākim (*c.* 1002–03)—were designed in the traditional hypostyle plan with axial cupolas. It is only in such architectural details as the elaborately composed facade of al-Ḥākim, with its corner towers and vaulted portal, that innovations appear, for most earlier mosques did not have large formal gates, nor was much attention previously given to the composition of the exterior facade. The Fāṭimids' architectural traditionalism was certainly a conscious attempt to perpetuate the existing aesthetic system.

Architectural traditionalism

Although much less is known about it, the Great Palace of the Fāṭimids belonged to the tradition of the enormous palace-cities typical of the ʿAbbāsids. Mediterranean rather than Iranian influences, however, played a greater part in the determination of its uses and functions. The whole city of Cairo (Arabic: al-Qāhirah, meaning "the Victorious"), on the other hand, has many symbolic and visual aspects that suggest a willful relationship to Baghdad.

The originality of Fāṭimid architecture does not lie in works sponsored by the caliphs themselves, even though Cairo's well-preserved gates and walls of the second half of the 11th century are among the best examples of early medieval military architecture. It is rather the patronage of lower officials and of the bourgeoisie, if not even of the humbler classes, that was responsible for the most interesting Fāṭimid buildings. The mosques of al-Aqmar (1125) and of aṣ-Ṣāliḥ (*c.* 1160) are among the first examples of monumental small mosques constructed to serve local needs. Even though their internal arrangement is quite traditional, their plans were adapted to the space available in the urban centre. These mosques were elaborately decorated on the exterior, exhibiting a conspicuousness absent from large hypostyle mosques.

A second innovation in Fāṭimid architecture was the tremendous development of mausoleums. This may be explained partially by Shīʿism's emphasis on the succession of holy men, but the development of these buildings in terms of both quality and quantity indicates that other influential social and religious issues were also involved. Most of the mausoleums were simple square buildings surmounted by a dome. Many of these have survived in Cairo and Aswān. Only a few, such as the *mashhad* at Aswān, are somewhat more elaborate, with side rooms. The most original of these commemorative buildings is the Juyūshī Mosque (1085) overlooking the city of Cairo. Properly speaking, it is not a mausoleum but a monument celebrating the reestablishment of Fāṭimid order after a series of popular revolts.

The Fāṭimids introduced, or developed, only two major constructional techniques: the systematization of the four-centred "keel" arch and the squinch. The latter innovation is of greater consequence because the squinch became the most common means of passing from a square to a dome, although pendentives were known as well. A peculiarly Egyptian development was the *muqarnas* squinch, which consisted of four units: a niche bracketed by two niche segments, superimposed with an additional niche. The complex profile of the *muqarnas* became an architectural element in itself used for windows, while the device of using niches and niche segments remained typical of Egyptian decorative design for centuries. It still is impossible to say whether the *muqarnas* was invented in Egypt or inspired by other architectural traditions (most likely Iranian). Fāṭimid domes were smooth or ribbed and developed a characteristic "keel" profile.

Use of the squinch

In the use of materials (brick, stone, wood) and structural concepts, Fāṭimid architecture continued earlier traditions. Occasionally local styles were incorporated, among them features of Tunisian architecture in the 10th century or of upper Mesopotamian in the late 11th century.

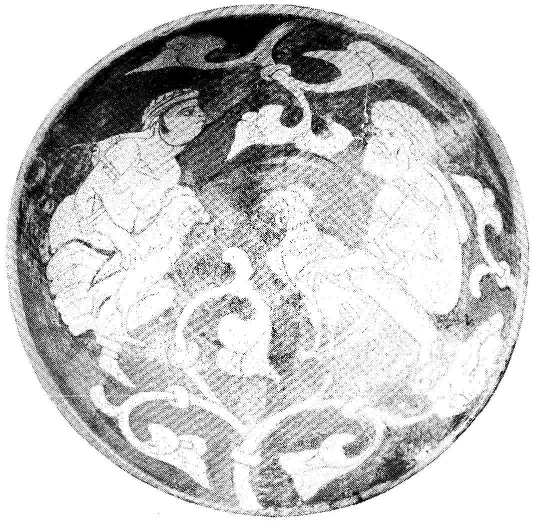
Stone sculpture, stucco work, and carved wood were utilized for architectural decorations. The Fāṭimids also employed mosaicists, who mostly worked in places like Jerusalem, where they imitated or repaired earlier mosaic murals. Many fragments of Fāṭimid wall paintings have survived in Egypt. Most of them, however, are too small to allow for making any iconographic or stylistic conclusions, with the exception of the mid-12th-century ceiling of the Cappella Palatina at Palermo. Built by the Norman kings of Sicily, the palace chapel was almost certainly decorated by Fāṭimid artists, or at least the artists adhered to Fāṭimid models. The hundreds of facets in the *muqarnas* ceiling were painted, notably with many purely ornamental vegetal and zoomorphic designs but also with scenes of daily life and many subjects that have not yet been explained. Stylistically influenced by Iraqi ʿAbbāsid art, these paintings are innovative in their more spatially aware representation of personages and of animals. Very similar tendencies appear also in the stucco and wood sculptures of Fāṭimid decoration. The stunning abstraction of the architectural decoration at Sāmarrā' tends to give way to more naturalistically conceived vegetal and animal designs; occasionally whole narrative scenes appear carved on wood. Another decorative trend is especially used on 12th-century *miḥrāb*s: explicitly complicated geometric patterns, usually based on stars, which in turn generate octagons, hexagons, triangles, and rectangles. Geometry becomes a sort of network in the midst of which small vegetal units continue to remain, often as inlaid pieces. Long inscriptions written in very elaborate calligraphies also became a typical form of architectural decoration on most of the major Fāṭimid buildings.

A clear separation must be made between the decorative arts sought by Fāṭimid princes and the arts produced within their empire. Little has been preserved of the former, notably a small number of superb ewers in rock crystal. A text has survived, however, that describes the imperial treasures looted in the middle of the 11th century by dissatisfied mercenary troops. It lists gold, silver, enamel, and porcelain objects that have all been lost, as well as textiles (perhaps the cape of the Norman king Roger II [Kunsthistorisches Museum, Vienna] is an example of the kind of textiles found in this treasure). The inventory also records that the Fāṭimids had in their possession many works of Byzantine, Chinese, and even Greco-Roman provenance. Altogether, then, it seems that the imperial art of the Fāṭimids was part of a sort of international royal taste that downplayed cultural or political differences.

Ceramics, on the other hand, were primarily produced by local urban schools and were not an imperial art. The most celebrated type of Fāṭimid wares were lustre-painted ceramics from Egypt itself. A large number of artisans' names have been preserved, thereby indicating the growing prestige of these craftsmen and the aesthetic importance

Lustreware with figural decoration

of their pottery. Most of the surviving lustre ceramics are plates on which the decoration of the main surface has been emphasized. The decorative themes used were quite varied and included all the traditional Islāmic ones: *e.g.,* calligraphy, vegetal and animal motifs, arabesques. The most distinguishing feature of these Fāṭimid ceramics, however, is the representation of the human figure. Some of these ceramics have been decorated with simplified copies of illustrations of the princely themes, but others have depictions of scenes of Egyptian daily life. The style in which these themes have been represented is simultaneously the hieratic, ornamental manner traditional to Islāmic painting combined with what can almost be called spatial illusionism. Wheel-cut rock crystal, glass, and bronze objects, especially animal-shaped aquamaniles (a type of water vessel) and ewers, are also attributed to the Fāṭimids.



Fāṭimid lustre-painted dish depicting a cockfight, from Egypt, late 11th–early 12th century. In the Edmund de Unger Collection, London. Diameter 24.1 cm.

By courtesy of the Edmund de Unger Collection; photograph, A.C. Cooper Ltd.

Book illustration. Manifestations of nonprincely Fāṭimid art also included the art of book illustration. The few remaining fragments illustrate that probably after the middle of the 11th century there developed an art of representation other than the style used to illustrate princely themes. This was a more illusionistic style that still accompanied the traditional ornamental one in the same manner as in the paintings on ceramics.

Transitional role of Fāṭimid art

In summary it would appear that Fāṭimid art was a curiously transitional one. Although much influenced by earlier Islāmic and non-Islāmic Mediterranean styles, the Fāṭimids devised new structural systems and developed a new manner of painting representational subjects, which became characteristic of all Muslim art during the 12th century. Neither documentary nor theoretical research in Islāmic art, however, has developed sufficiently to clearly establish whether the Fāṭimids were indeed innovators or whether their art was a local phenomenon that is only accidentally relatable to what followed.

Seljuq art. During the last decades of the 10th century, at the Central Asian frontiers of Islām, a migratory movement of Turkic peoples began that was to affect the whole Muslim world up to and including Egypt. The dominant political force among these Turks was the dynasty of the Seljuqs, but it was not the only one; nor can it be demonstrated, as far as the arts are concerned, that it was the major source of patronage in the period to be discussed anywhere but in Anatolia in the 12th and 13th centuries. The Seljuq empire, therefore, consisted of a succession of dynasties, and all but one (the Ayyūbids of Syria, Egypt, and northern Mesopotamia) were Turkic.

A complex feudal system was established and centred on urban areas. Cities were established or expanded, particularly in western Iran, Anatolia, and Syria. Militant Muslims, the Seljuqs also sought to revive Muslim orthodoxy. Although politically unruly and complicated in their relationships to one another, the successive and partly overlapping dynasties of the Ghaznavids, Ghūrids, the Great Seljuqs, Qarakhānids, Zangids, Ayyūbids, Seljuqs of Rūm, and Khwārezm-Shāhs (considering only the major ones) seem to have created a comparatively unified culture from India to Egypt. The art of the Seljuq period, however, is difficult to discuss coherently both because of the wealth of examples and because of the lack of synchronization between various technical and regional developments. This complex world fell apart under the impact of the Mongol invasions that, from 1220 until 1260, swept through the Muslim lands of the Near East.

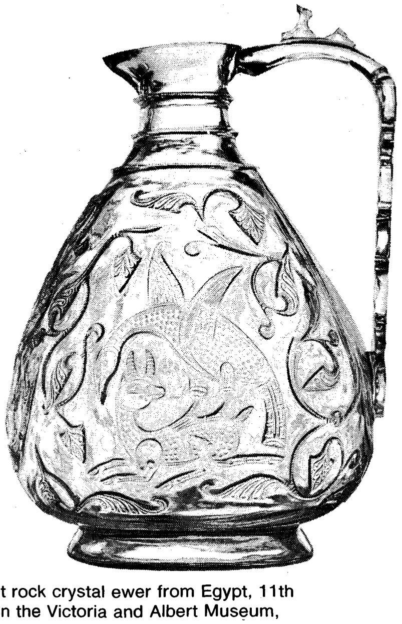
Architecture. The functions of monumental architecture in the Seljuq period were considerably modified. Large congregational mosques were still built. The earliest Seljuq examples occur in the two major new provinces of Islām—Anatolia and northwestern India—as well as in the established Muslim region of western Iran. In some areas, such as the Isfahan region, congregational mosques were rebuilt, while in other parts of Islām, such as Syria or Egypt, where there was no need for new large mosques, older ones were repaired and small ones were built. The latter were partly restricted to certain quarters or groups or were commissioned by various guilds, particularly in Damascus.

A curious side aspect of the program of building, rebuilding, or decorating mosques was the extraordinary development of minarets. Particularly in Iran, dozens of minarets are preserved from the 12th and 13th centuries, while the mosques to which they had been attached have disappeared. It is as though the visual function of the minaret was more important than the religious institution to which it was attached.

Small or large, mausoleums increased in numbers and became at this time the ubiquitous monument they appear to be. Most of the mausoleums, such as the tomb tower of Abū Yazīd al-Bisṭāmī (died 874) at Besṭām, were dedicated to holy men—both contemporary Muslim saints and all sorts of holy men dead for centuries (even pre-Islāmic holy men, especially biblical prophets, acquired a monument). The most impressive mausoleums, however, ones like the one of Sanjar at Merv, were built for royalty. Pilgrimages were organized and in many places hardly mentioned until then as holy places (*e.g.,* Meshed, Besṭām, Mosul, Aleppo); a whole monastic establishment serving as a centre for the distribution of alms was erected with hostels and kitchens for the pilgrims.

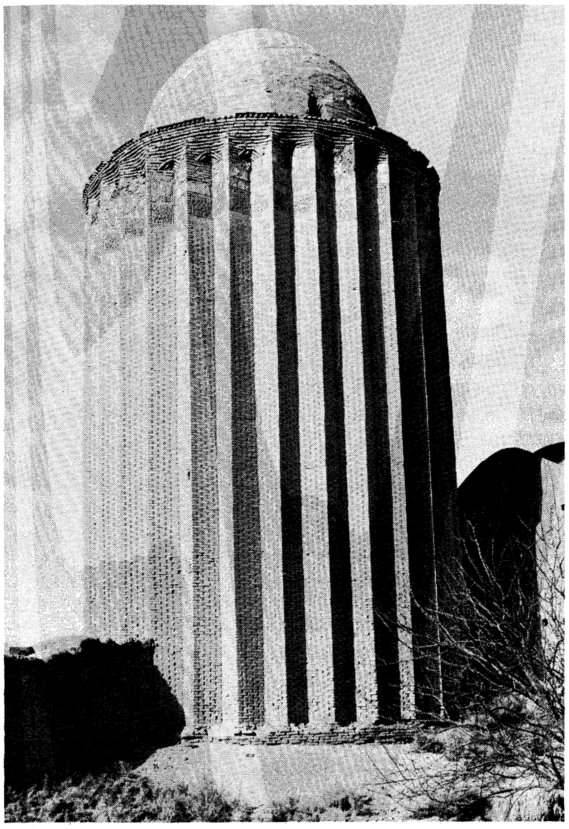
Origin and development of the *madrasah*

Although enormously expanded, mosques, minarets, and mausoleums were not new types of Islāmic architecture. The *madrasah* ("school"), however, was a new building



Wheel-cut rock crystal ewer from Egypt, 11th century. In the Victoria and Albert Museum, London. Ht. 21.5 cm.

By courtesy of the Victoria and Albert Museum, London



Tomb tower at the shrine of Abū Yazīd al-Bisṭāmī at Besṭām, Iran, 1313.

Josephine Powell, Rome

type. There is much controversy as to why and how it really developed. Although early examples have been discovered in Iran, such as the 11th century *madrasah* of Khargird in Iran and at Samarkand, it is from Anatolia, Syria, and Egypt that most of the information about the *madrasah* has been derived. In the latter regions it was usually a privately endowed establishment reserved for one or two of the schools of jurisprudence of orthodox Islām. It had to have rooms for teaching and living quarters for the students and professors. Often the tomb of the founder was attached to the *madrasah*. Later *madrasahs* were built for two or three schools of jurisprudence, and the Mustanṣirīyah in Baghdad was erected in 1233 to be a sort of ecumenical *madrasah* for the whole of Sunnī Islām.

In the Seljuq period there occurred a revival of the *ribāṭ* inside cities. *Khānqāhs*, monasteries, and various establishments of learning other than formal *madrasahs* were also built.

An impressive development of secular architecture occurred under the Seljuqs. The most characteristic building of the time was the citadel, or urban fortress, through which the new princes controlled the usually alien city they held in fief. The largest citadels, like those of Cairo and Aleppo, were whole cities with palaces, mosques, sanctuaries, and baths. Others, like the Citadel of Damascus, were simpler constructions. Occasionally, as in the Euphrates valley, single castles were built, possibly in imitation of those constructed by the Christian crusaders. Walls surrounded most cities, and all of them were built or rebuilt during the Seljuq period.

Little is known about Seljuq palaces or private residences in general. A few fragments in Konya or in Mosul are insufficient to give a coherent idea about urban palaces, and it is only in Anatolia and in Central Asia that an adequate idea of other types can be obtained. Anatolian palaces are on the whole rather small villa-like establishments; but, in Afghanistan and Soviet Central Asia, excavations at Tirmidh, Lashkarī Bāzār, and Ghaznī have brought to light a whole group of large royal palaces erected in the 11th and early 12th centuries.

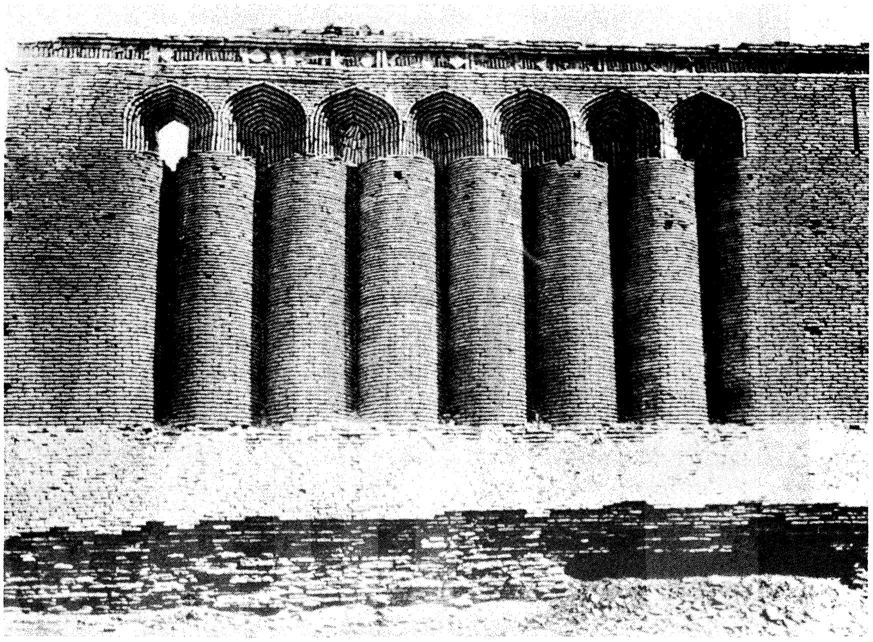
Seljuq commercial architecture

Commercial architecture became very important. Individual princes and cities probably were trying to attract business by erecting elaborate caravansaries on the main trade routes such as Rebāṭ-e Malek built between Samarkand and Bukhara in Iran. The most spectacular caravansaries were built in the 13th century in Anatolia. Equally impressive, however, although less numerous, are the caravansaries erected in eastern Iran and northern Iraq. Bridges also were rebuilt and decorated like the one at Cizre in Turkey.

The forms of architecture developed by the Seljuqs were remarkably numerous and varied considerably from region to region. Since the Iranian innovations dating from the 11th century and first half of the 12th century are the earliest and, therefore, probably influenced all other areas of the Seljuq empire, they will be discussed first.

Even though it is not entirely typical, the justly celebrated Great Mosque of Isfahan was one of the most influential of all early Seljuq religious structures. Probably completed around 1130 after a long and complicated

By courtesy of the General Direction of Museums and Historical Monuments, Ministry of Culture and Arts, Tehran, Iran



Brickwork facade of the caravansary of Rebāṭ-e Malek, 11th century.

history of rebuildings, it consisted of a large courtyard on which opened four large vaulted halls known as *eyvāns*; the *eyvāns* created the compositional axes of each side of the court. On the side of the *qiblah* the hall of the main *eyvān* was followed by a huge cupola. The area between *eyvāns* was subdivided into a large number of square bays covered by domes. The Isfahan mosque also had a unique feature: on the north side a single domed hall positioned on the main axis of the building was in all probability a formal hall for princes to change their clothes before entering into the sanctuary of the mosque.

Use of the *eyvān*

The two features of the Great Mosque at Isfahan that became characteristic of Seljuq mosques were the *eyvān* and the dome. The *eyvān* was an architectural element known already in Sāsānian architecture that had been used in residential buildings from Egypt to Central Asia before the 11th century. In fact, the use of the *eyvān* was not restricted to just mosques, but it also appears in palaces (Lashkarī Bāzār), caravansaries (Rebāṭ-e Sharaf), and in *madrasahs*. The *eyvān* was, in other words, a unit of architectural composition that had no specific use and, therefore, no meaning. In the mosques of the 12th century, four *eyvāns* were used, at least in the clearly definable architectural school of western Iran (*e.g.,* Ardestān, Zavāreh). This kind of composition had two principal effects. One was that the *eyvāns* centralized the visual effect of the mosque by making the courtyard the centre of the building. The other effect of this composition was that it broke up into four areas what had for centuries been a characteristic of the mosque: its single, unified space. The reasons for these developments are still speculative.

Whether large or small, cupolas or domes were used in mosques, caravansaries, and palaces. They were the main architectural features of almost all mausoleums, where they were set over circular or polygonal rooms.

Two characteristic Iranian architectural forms are not present in the Great Mosque of Isfahan but occur elsewhere in the city. One is the tower. Those narrow and tall (up to about 150 feet [50 metres]) were minarets, of which several dozen have been preserved all over Iran and Central Asia (such as the one at Jām). Shorter and squatter towers were mausoleums. These were particularly typical of northern Iran. The other characteristic architectural type exists only in Isfahan in a much-damaged state. It is the *pīshṭāq,* or a formal gateway that served to emphasize a building's presence and importance.

The *pīshṭāqs* of Isfahan

Domes and *eyvāns* indicate the central concern of Iranian construction during the Seljuq period: vaulting in baked brick became the main vehicle for any monumental construction (mud brick was used for secondary parts of a building, frequently for certain secular structures). A large and forcefully composed octagonal base developed the *muqarnas* squinch from a purely ornamental feature into one wherein both structural and decorative functions combined. In some later buildings, such as the mausoleum of Sanjar at Merv, a system of ribs was used to vault an octagonal zone. Seljuq architects sought to make their domes visible from afar and for this reason invented the double dome. Its outer shell was raised on a high drum, while the interior kept the traditional sequence: square base, zone of transition, and dome. Using this structural device, therefore, exterior height was achieved without making the exterior dome too heavy and without complicating the task of decorating the interior, always a problem in countries like Iran with limited supplies of wood for scaffolding. Domes along the *eyvāns* were another factor in contributing to the growing separation between the exterior and interior view of a building. There was also an emphasis on the visibility of a building from the exterior that is indicated by the construction of tall circular or polygonal minarets and high facades.

Architectural decoration was intimately tied to structure. Two mediums predominated. One was stucco, which continued to be used to cover large wall surfaces. The other was brick. Originating in the 10th-century architecture of northeastern Iran, brick came to be employed as a medium of construction as well as a medium of decoration. The complex decorative designs worked out in brick often had a rigidly geometric effect. Especially cut shapes of terracotta and brick, frequently produced in unusual sizes, served to soften these geometric patterns by modifying their tactile impact and by introducing additional curved or beveled lines to the straight lines of geometry.

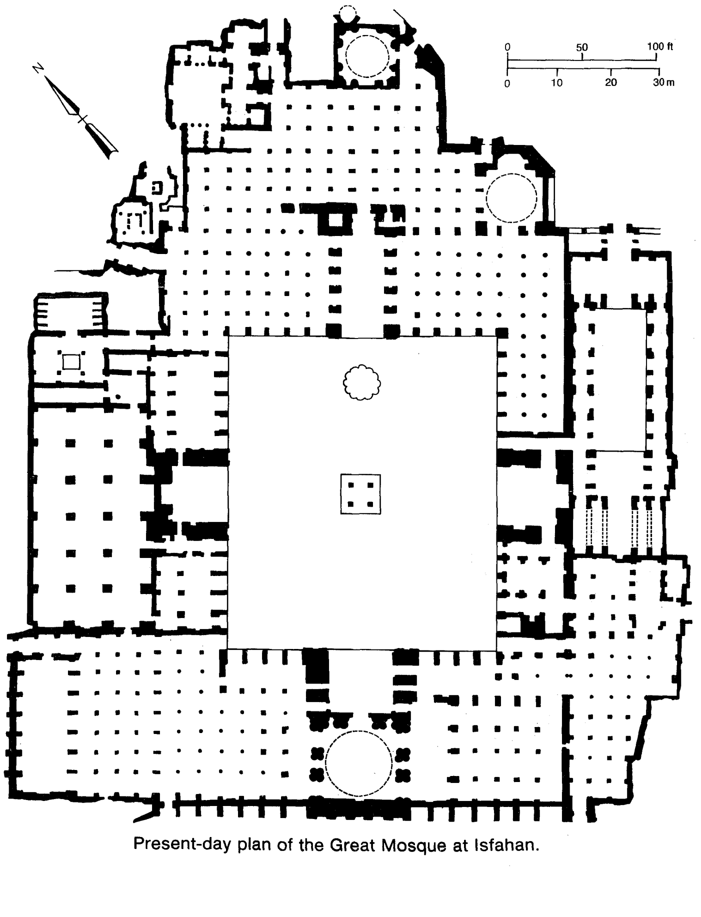
Paintings were used for architectural decoration, especially in palaces. From the second half of the 12th century coloured tiles began to be utilized to emphasize the contour of a decorative area in a structural unit; tiles were not used, however, to cover whole walls. There are also examples of architectural sculpture of animals and people.

Most of the decorative designs tended to be subordinated to geometry, and even calligraphic or vegetal patterns were affected by a seemingly mathematically controlled aesthetic. It has been suggested that these complex geometric designs were a result of an almost mystical passion for number theories that were popularized in 11th-century Iran by such persons as the scholar and scientist al-Bīrūnī or the poet-mathematician Omar Khayyam. But even if the impulses for geometric design were originally created at the highest intellectual level, the designs themselves rapidly became automatic patterns. Their quality was generally high, but a tendency toward facility can be observed in such buildings as Rebāṭ-e Sharaf.

In Iraq, northern Mesopotamia, Syria, and Egypt (after 1171), the architectural monuments do not, on the whole, appear as overwhelmingly impressive as those of Iran, largely because the taste of Umayyad and ʿAbbāsid times continued to dominate mosque architecture. It is in the construction of new building types, particularly the *madrasah,* that the most originality is apparent. The Syrian *madrasahs* in Damascus, like al-ʿĀdilīyah, aẓ-Ẓāhirīyah, or the works of Nureddin, tended also to follow a comparatively standardized plan: an elaborate facade led into a domed hallway and then into a court with at least one *eyvān.* Most of these *madrasahs* were small and were fitted into a preexisting urban pattern. The use of *eyvāns* and the construction of the many minarets found in Mosul or on the Euphrates certainly attest to the influence of Iranian Seljuq design.

The main achievement of Ayyūbid, Zangid, or Seljuq architecture in the Fertile Crescent was the translating into stone of new structural systems first developed in

From John D. Hoag, *Islamic Architecture;* copyright 1975 Electa Editrice, Milan



Present-day plan of the Great Mosque at Isfahan.

brick. The most impressive instance of this lies in the technically complex *muqarnas* domes and half domes or in the *muqarnas* pendentives of Syrian buildings. Elaborate *miḥrāb*s were also made of multicoloured stones that were carefully cut to create impressive patterns. The architecture of the Fertile Crescent, therefore, was still dominated by the sheer force of stone as a material for both construction as well as decoration, and, therefore, the architecture was more Mediterranean in effect than were the buildings of Iran.

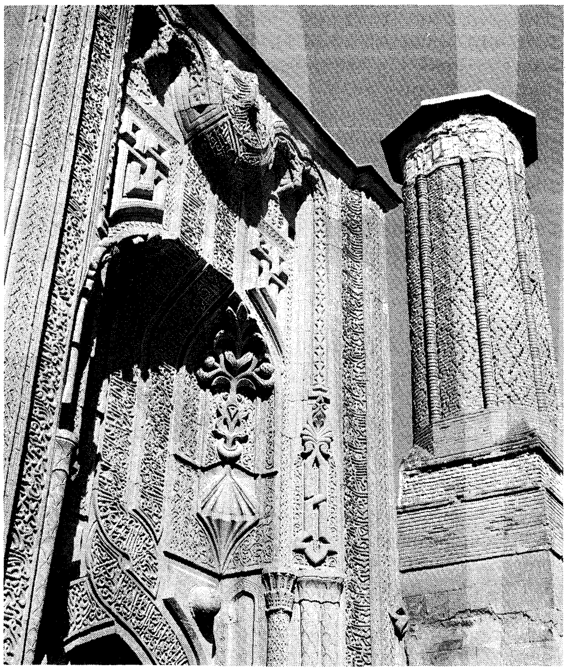
The hybrid style of Anatolia

This Mediterranean tendency was also evident in the 13th-century architecture of Seljuq Anatolia. This new province of Islām was rapidly populated with new immigrants and consequently gathered themes and motifs from throughout the Muslim world, as well as from the several native Anatolian traditions of Byzantine, Armenian, and Georgian architecture. The resulting assimilation of styles produced an overwhelmingly original architecture, for each building in Konya, Kayseri, Sivas, Divriği, Erzurum, or on the roads between them is a unique monument.

Functionally the buildings in Anatolia do not differ from those in other parts of the Muslim world. All the structural forms found in Syria and Iran can be found in Anatolia as well, although they have often been adapted to local materials. Three uniquely Anatolian architectural features, however, can be distinguished. One was limited to Konya at this time but would have an important widespread development later on. As it appears in the Ince or Karatay *medreses*, it consists of the transformation of the central courtyard into a domed space while maintaining the *eyvān*. Thus the centralized aspect of the *eyvān* plan becomes architecturally explicit. The second feature is the creation of a facade that usually consisted of a high central portal—often framed by two minarets—with an elaborately sculpted decorative composition that extended to two corner towers. The third distinguishing feature of Anatolian Seljuq architecture is the complexity of the types of funerary monuments that were constructed.

From the point of view of construction, most of Anatolian architecture is of stone. In Konya and a number of eastern Anatolian instances, brick was used. Barrel vaults, groin vaults, *muqarnas* vaults, squinch domes, pendentive domes, and the new pendentive known as "Turkish triangle" (a transformation of the curved space of the traditional pendentive into a fanlike set of long and narrow triangles built at an angle from each other) were all used by Anatolian builders, thereby initiating the great development of vault construction in Ottoman architecture (see below).

Ara Guler



Ince Minare at Konya, Tur., 1258. Detail view showing the sculptural ornamentation of the main facade portal and the decorative brickwork of the minaret.

Architectural decoration

Architectural decoration consisted primarily in the stone sculpture found on the facades of religious and secular buildings. Although influenced by Iran and Syria in many details, most Anatolian themes were original, although some exhibit Armenian and possibly Western influences. The exuberance of Anatolian architectural decoration can perhaps be best demonstrated in the facades of Sivas' Gök Medrese and of Konya's Ince Minare. In addition to the traditional geometric, epigraphic, and vegetal motifs, a decorative sculpture in the round or in high relief was created that included many representations of human figures and especially animals. Whether this sculpture is essentially a reflection of the decorative wealth of pre-Islāmic monuments in Anatolia, or whether it is the vestige of a pagan Turkish art that originated in Central Asia, is still an unsolved historical problem.

There are few examples of wall painting from Anatolia. Especially in Konya, however, a major art of painted-tile decoration did evolve, possibly developed by Iranian artists who fled from the Mongol onslaught.

In summing up the architectural development of the Seljuq period, three points seem to be particularly significant. One is the expansion of building typology and the erection of new monumental architectural forms, thus illustrating an expansion of patronage and a growing complexity of taste. The second point is that, regardless of the quality and interest of monuments in the Fertile Crescent, Egypt, and Anatolia, the most inventive and exciting architecture in the 11th and 12th centuries was that of Iran. But, far more than in the preceding period, regional needs and regional characteristics seem to predominate over synchronic and pan-Islāmic ones. Finally, there was a striking growth of architectural decoration both in sophistication of design and in variation of technique.

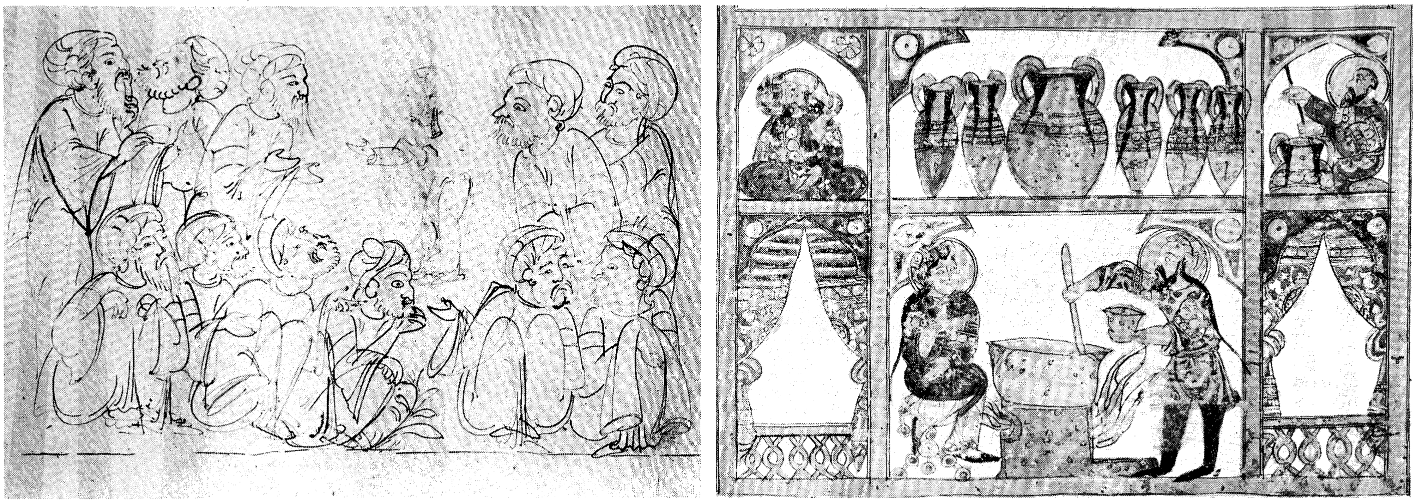
Other arts. Although probably not as varied as architecture, the other arts of the Seljuq period also underwent tremendous changes. They demonstrate an extraordinary artistic energy, a widening of the social patronage of the arts, and a hitherto unknown variety of topics and modes of expression. It was as though the Seljuq period was gathering a sort of aesthetic momentum, but this effort seems to have been curtailed by the Mongol invasion. Chronologically, almost all surviving documentation and examples of these arts date from the latter part of the period, after 1150. It is unclear whether this apparent date is merely an accidental result of what has been preserved and is known through 20th-century scholarship or whether it corresponds to some precise event or series of events.

Glass and textiles continued to be major mediums during the Seljuq period. Ceramics underwent many changes, especially in Iran, where lustre painting became widespread and where new techniques were developed for colouring pottery. Furthermore, the growth of tile decoration created a new dimension for the art of ceramics.

Inlaid metalwork became an important technique. First produced at Herāt in Iran (now in Afghanistan) in the middle of the 12th century, this type of decoration spread westward, and a series of local schools were established in various regions of the Seljuq domain. In this technique, the surfaces of utilitarian metallic objects (candlesticks, ewers, basins, kettles, and so forth) were engraved, and then silver was inlaid in the cut-out areas to make the decorative design more clearly visible.

Manuscript illustration also became an important art. Scientific books, including the medical manuals of Dioscorides and of Galen, or literary texts such as the picaresque adventures of a verbal genius known as the *Maqāmāt,* were produced with narrative illustrations throughout the text.

All of the technical novelties of the Seljuqs seem to have had one main purpose: to animate objects and books and to provide them with clearly visible and identifiable images. Even the austere art of calligraphy became occasionally animated with letters ending in human figures. The main centres for producing these arts were located in Iran and the Fertile Crescent. For reasons yet unknown, Egypt and Anatolia were far less involved. One reason may be that these two Seljuq provinces did not witness the same



Seljuq manuscript illustration.
(Left) Drawing from a manuscript of the *Maqāmāt*, 1323. In the British Museum (MS. Add 7293, f. 285v). (Right) Preparing medicine from honey, manuscript illustration from an Arabic translation of *De materia medica* of Dioscorides copied by ʿAbd Allāh ibn al-Faḍl, from Baghdad, Iraq, 13th century. In the Metropolitan Museum of Art, New York City. 12.5 cm × 17.5 cm.

By courtesy of (left) the trustees of the British Museum, and (right) the Metropolitan Museum of Art, New York, the Cora Timken Burnett collection of Persian Miniatures and other Persian art objects; bequest of Cora Timken Burnett, 1957

Urban middle-class patronage

rise of an urban middle class as did Iran, Iraq, or Syria. It would seem from a large number of art objects whose patrons are known that the main market for these works of art was the mercantile bourgeoisie of the big cities and not, as has often been believed, the princes. Seljuq decorative arts and book illustration, therefore, reflect an urban taste.

The themes and motifs used were particularly numerous. In books they tend to be illustrations of the text, even if a manuscript such as the Schefer *Maqāmāt* (1237; Bibliothèque Nationale, Paris) sought to combine a strict narrative with a fairly naturalistic panorama of contemporary life. Narrative scenes taken from books or reflecting folk stories are also common on Persian ceramics. In all mediums, however, the predominant vocabulary of images is the one provided by the older art of princes; but its meaning is no longer that of illustrating the actual life of princes but rather that of symbolizing a good and happy life. The motifs, therefore, do not have to be taken literally. Next to princely and narrative themes, there are depictions of scenes of daily life, astronomical motifs, and a myriad of topics that can be described but not understood.

While it is possible within certain limits to generalize about the subject matter of Seljuq art, regional stylistic definitions tend to be more valid. Thus the bronzes produced in northeastern Iran in the 12th century are characterized by simple decorative compositions rather than by the very elaborate ones created by the so-called school of Mosul in Iraq during the 13th century. In general, the art of metalwork exhibits a consistently growing intricacy in composition and in details to the point that individual subjects are at times lost in overlapping planes of arabesques. Ceramic pieces of Iran have usually been classified according to a more or less fictitious provenance. Kāshān ware exhibits a perfection of line in the depiction of moon-faced personages with heavily patterned clothes, while Rayy ceramic work is less sophisticated in design and execution but more vividly coloured. Sāveh and Gurgān are still other Iranian varieties of pottery. With the exception of Kāshān ware, where dynasties of ceramicists are known, all these types of Iranian pottery were contemporary with each other. In Syria, Raqqah pottery imitated Iranian ceramic wares but with a far more limited vocabulary of designs.

The Baghdad, or Arab, school

The main identifiable group of miniature painters was the so-called Baghdad school of the first half of the 13th century. The group should be called the Arab school because the subject matter and style employed could have been identified with any one of the major artistic centres of Egypt and the Fertile Crescent, and very little evidence currently exists to limit this school to one city. The miniatures painted by these artists are characterized by the colourful and often humorous way in which the urbanized Arab is depicted. The compositions, often lacking in any strong aesthetic intent, are documentary caricatures in which the artist has recorded the telling and recognizable gesture or a known and common setting or activity. In many images or compositional devices one can recognize the impact of the richer Christian Mediterranean tradition of manuscript illumination. A greater attention to aesthetic considerations is apparent in the illustrated manuscript of the Persian epic *Varqeh o-Golshāh* (Topkapı Saray Museum, Istanbul), unique in the Seljuq period.

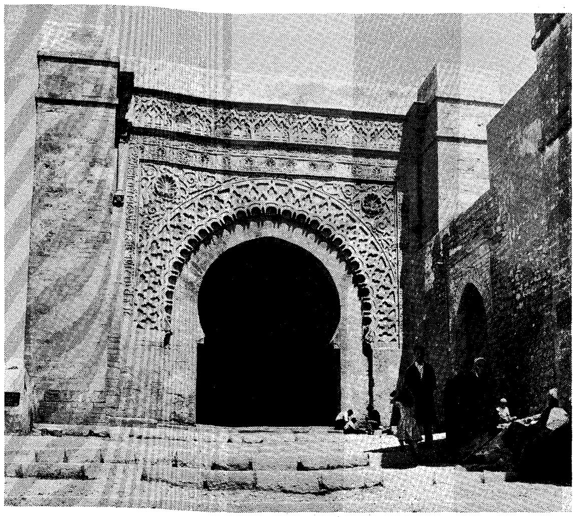
Western Islāmic art: Moorish. The 11th to 13th centuries were not peaceful in the Maghrib. Berber dynasties overthrew each other in Morocco and the Iberian Peninsula. The Christian reconquest gradually diminished Muslim holdings in Spain and Portugal, and Tunisia was ruined during the Hilālī invasion when Bedouin tribes were sent by the Fāṭimids to prevent local independence.

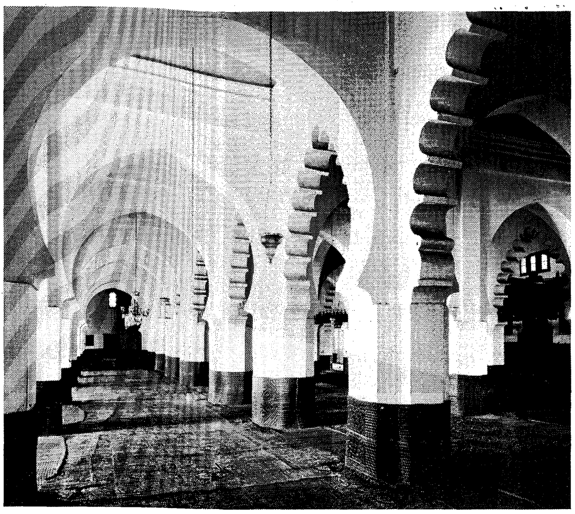
Austerity of North African architecture

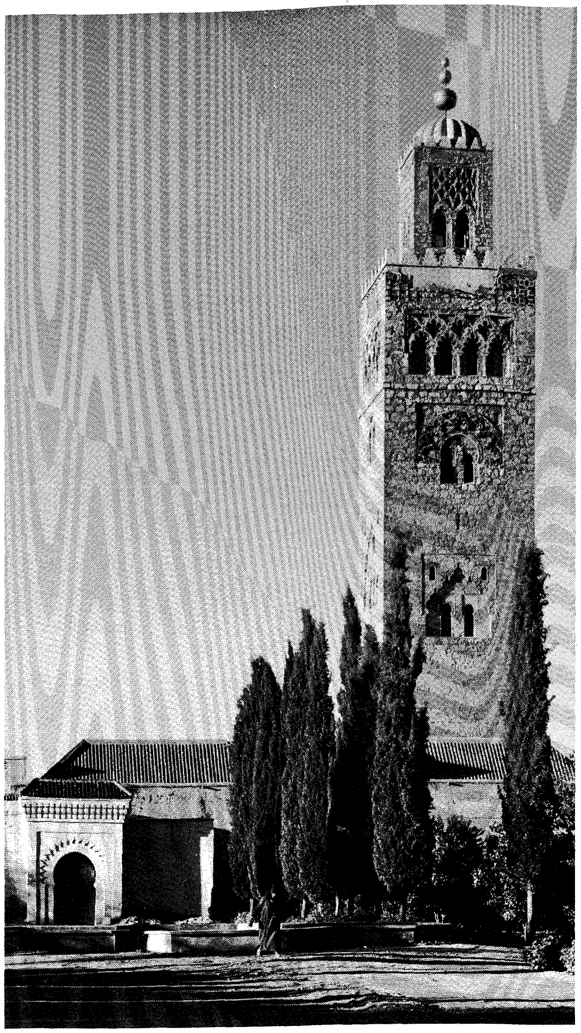
Two types of structures characterize the Almoravid (1056–1147) and Almohad (1130–1269) periods in Morocco and Spain. One comprises the large, severely designed Moroccan mosques such as those of Tinmel, of Ḥasan in Rabat, or of the Kutubīyah in Marrakech. They are all austere hypostyles with tall, massive, square minarets. The other distinctive type of architecture was that built for military purposes, including fortifications and, especially, massive city gates with low-slung horseshoe arches, such as the Oudaia Gate at Rabat (12th century) or the Rabat Gate at Marrakech (12th century). Palaces built in central Algeria by minor dynasties such as the Zīrids were more in the Fāṭimid tradition of Egypt than in the Almoravid and Almohad traditions of western Islām. Almost nothing is known or has been studied about North African arts other than architecture because the puritanical world of the Berber dynasties did not foster the arts of luxury.

In North Africa the artistic milieu did not change much in the 14th and 15th centuries. Hypostyle mosques such as the Great Mosque of Algiers continued to be built; *madrasahs* were constructed with more elaborate plans; the Bū ʿInānīyah *madrasah* at Fès is one of the few monumental buildings of the period. A few mausoleums were erected such as the so-called Marīnid tombs near Fès (second half of the 14th century) or the complex of Chella at Rabat (mostly 14th century). Architectural decoration in stucco or sculpted stone was usually limited to elaborate geometric patterns, epigraphic themes, and a few vegetal motifs.

A stunning exception to the austerity of North African architecture exists in Spain in the Alhambra palace com-







Almoravid and Almohad architecture in North Africa.
(Top left) The Rabat Gate, Marrakech, Mor., late 12th century, Almoravid period. (Bottom left) Interior of the Great Mosque of Algiers, Alg., 12th century, Almohad period. (Right) The square minaret of the Kutubīyah Mosque in Marrakech, Mor., Almoravid period.

(Top left) Josephine Powell, Rome, (bottom left) H. Roger-Viollet, (right) GEKS

plex at Granada. The hill site of the Alhambra had been occupied by a citadel and possibly by a palace since the 11th century, but little of these earlier constructions has remained. In the 14th century two successive princes, Yūsuf I and Muḥammad V, transformed the hill into their official residence. Outside of a number of gates built like triumphal arches and several ruined forecourts, only three parts of the palace remain intact. First there is the long Court of the Myrtles leading to the huge Hall of Ambassadors located in one of the exterior towers. This was the part of the Alhambra built by Yūsuf I. Then there is the Court of the Lions, with its celebrated lion fountain in the centre. Numerous rooms open off this court, including the elaborately decorated Hall of the Two Sisters and the Hall of the Abencerrajes. The third part, slightly earlier than the first two, is the Generalife; it is a summer residence built higher up the hill and surrounded by gardens with fountains, pavilions, and portico walks.

Importance of the Alhambra

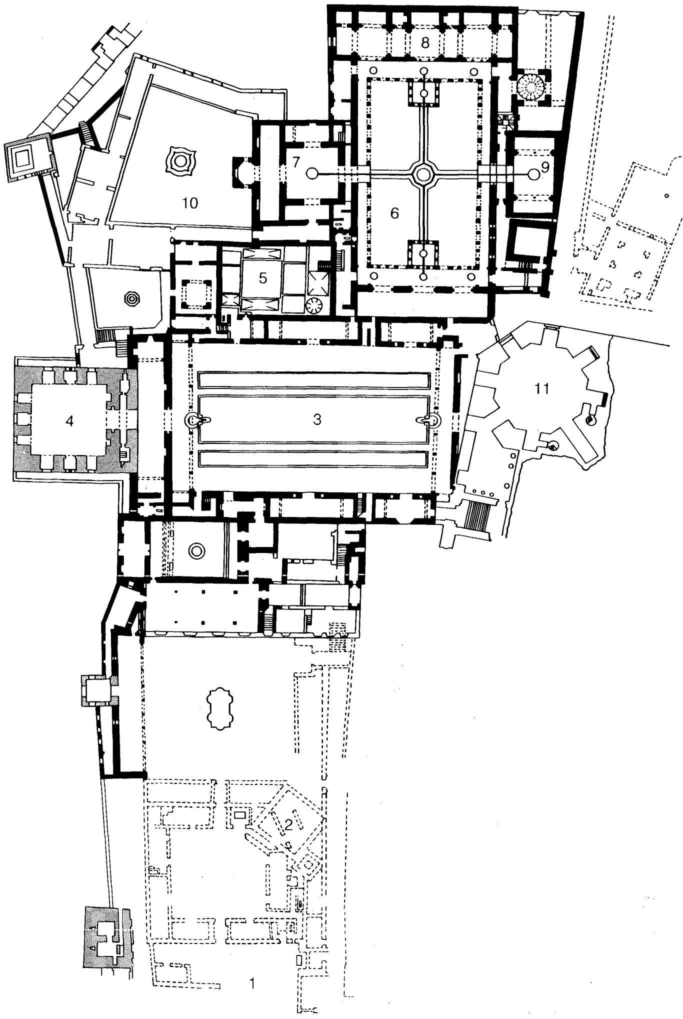
The Alhambra is especially important because it is one of the few palaces to have survived from medieval Islāmic times. It illustrates superbly a number of architectural concerns occasionally documented in literary references: the contrast between an unassuming exterior and a richly decorated interior to achieve an effect of secluded or private brilliance; the constant presence of water, either as a single, static basin or as a dynamic fountain; the inclusion of oratories and baths; the lack of an overall plan (the units are simply attached to each other).

The architectural decoration of the Alhambra was mostly of stucco. Some of it is flat, but the extraordinarily complex cupolas of *muqarnas,* such as in the Hall of the Two Sisters, appear as huge multifaceted diadems. The decoration of the Alhambra becomes a sort of paradox as well as a tour de force. Weighty, elaborately decorated ceilings, for example, are supported by frail columns or by walls pierced with many windows (light permeates almost every part of the large, domed halls). Much of the design and decoration of the Alhambra is symbolically oriented. The poems that adorn the Alhambra as calligraphic ornamentation celebrate its cupolas as domes of heaven rotating around the prince sitting under them.

Islāmic art as such ceased to be produced in Spain after 1492, when Granada, the last Moorish kingdom in Spain, fell to the Christians; but the Islāmic tradition continued in North Africa, which remained Muslim. In Morocco the so-called Sharīfian dynasties from the 16th century onward ornamentally developed the artistic forms created in the 14th century.

Most of the best known monuments of western Islāmic art are buildings, although a very original calligraphy was developed. The other arts cannot be compared in wealth and importance either with what occurred elsewhere in Islām at the same time or with earlier objects created in Spain. There are some important examples of metalwork, wood inlaid with ivory, and a lustre-glaze pottery known as Hispano-Moresque ware. The fact that the latter was made in Valencia or Málaga after the termination of Muslim rule demonstrates that Islāmic traditions in the decorative arts continued to be adhered to, if only partially. The term Mudéjar, therefore, is used to refer to all the things made in a Muslim style but under Christian rule. Numerous examples of Mudéjar art exist in ceramics and

Mudéjar and Mozarabic art



Plan of the Alhambra in Granada, Spain.
1. Alcazaba; 2. ruins of the mosque; 3. Court of the Myrtles; 4. Tower of Comares and Hall of Ambassadors; 5. baths; 6. Court of the Lions; 7. Hall of the Two Sisters; 8. Hall of the Kings; 9. Hall of the Abencerrajes; 10. Garden of Daraxa; 11. Palace of Charles V.
From G. Marcais, *L'Architecture Musalmane d'Occident*

textiles, as well as in architectural monuments such as the synagogues of Toledo and the Alcazba in Seville, where even the name of the ruling Christian prince, Don Pedro, was written in Arabic letters. The Mudėjar spirit, in fact, permeated most of Spanish architectural ornament and decorative arts for centuries, and its influence can even be found in Spanish America.

Mudėjar art must be carefully distinguished from Mozarabic art: the art of Christians under Muslim rule. Mozarabic art primarily flourished in Spain during the earlier periods of Muslim rule. Its major manifestations are architectural decorations, decorative objects, and illuminated manuscripts. Dating mostly from the 10th and 11th centuries, the celebrated illuminations for the commentary on the Revelation to John by an 8th-century Spanish abbot, Beatus of Liėbana, are purely Christian subjects treated in styles possibly influenced by Muslim miniature painting or book illustration. The most celebrated example, known as the "Saint-Sever Apocalypse," is in the collection of the Bibliothėque Nationale in Paris.

Mamlūk art. The Mamlūks were originally white male slaves, chiefly Turks and Circassians from the Caucasus and Central Asia who formed the mercenary army of the various feudal states of Syria and Egypt. During the 13th century the importance of this military caste grew as the older feudal order weakened and military commanders took over power generally as nonhereditary sultans. They succeeded in arresting the Mongol onslaught in 1260 and, through a judicious but complicated system of alliance with the urban elite class, managed to maintain themselves in power in Egypt, Palestine, and Syria until 1517.

During the Mamlūk period Egypt and Syria were rich commercial emporiums. This wealth explains the quality and quantity of Mamlūk art. Most of the existing monuments in the old quarters of Cairo, Damascus, Tripoli, and Aleppo are Mamlūk; in Jerusalem almost everything visible on the Ḥaram ash-Sharīf, outside the Dome of the Rock, is Mamlūk. Museum collections of Islāmic art generally abound with Mamlūk metalwork and glass. Some of the oldest remaining carpets are Mamlūk. This creativity required, of course, more than wealth; it also required a certain will to transform wealth into art. This will was in part the desire of parvenu rulers and their cohorts to be remembered. Furthermore, architectural patronage flourished because of the institutionalization of the *waqf,* an economic system in which investments made for holy purposes were inalienable. This law allowed the wealthy to avoid confiscation of their properties at the whim of the caliph by investing their funds in religious institutions. In the Mamlūk period, therefore, there was a multiplication of *madrasah*s, *khānqāh*s, *ribāṭ*s, and *masjid*s, often with tombs of founders attached to them. The Mamlūk establishment also repaired and kept up all the institutions, religious or secular, that had been inherited by them, as can be demonstrated by the well-documented repairs carried on in Jerusalem and Damascus.

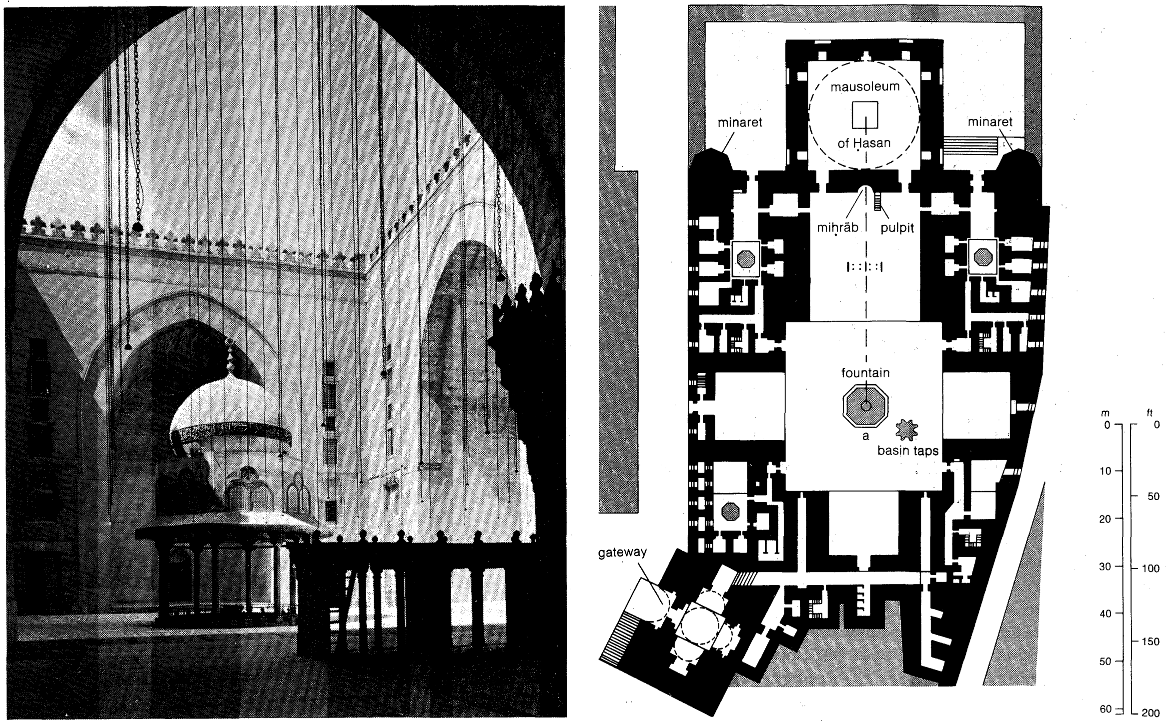
The influential role of the *waqf*

Architecture. The Mamlūks created a monumental setting for Syria and Egypt that lasted until the 20th century. It was at its most remarkable in architecture, and nearly 3,000 major monuments have been preserved or are known from texts in cities from the Euphrates to Cairo. No new architectural types came into being, although many more urban commercial buildings and private houses have been preserved than from previous centuries. The hypostyle form continued to be used for mosques and oratories, as in the Cairene mosques of Baybars I (1262–63), Nāṣir (1335), and Mu'ayyad Shaykh (1415–20). *Madrasah*s used *eyvān*s, and the justly celebrated *madrasah* of Sulṭan Ḥasan in Cairo (1356–62) is one of the few perfect four-*eyvān madrasah*s in the Islāmic world. Mausoleums were squares or polygons covered with domes. In other words, there were only minor modifications in the typology of architecture, and even the 15th-century buildings with interiors totally covered with ornamentation have possible prototypes in the architecture of the Seljuqs. Yet there are formal and functional features that do distinguish Mamlūk buildings. One is the tendency to build structures of different functions in a complex or cluster. Thus the Qalā'ūn mosque (1284–85) in Cairo has a mausoleum, a *madrasah,* and a hospital erected as one architectural unit. Another characteristic is the tendency of Mamlūk patrons to build their major monuments near each other. As a result, certain streets of Cairo, such as Bayn al-Qaṣrayn, became galleries of architectural masterpieces. The plans of these buildings may have had to be adapted to the exigencies of the city, but their spectacular facades and minarets competed with each other for effect. From the second half of the 14th century onward, building space for mausoleums began to be limited in Cairo, and a vast complex of commemorative monuments was created in the city's western cemetery. In Aleppo and Damascus similar phenomena can be observed.

Architectural changes in the Mamlūk period

Although Mamlūk architecture was essentially conservative in its development of building types, more originality is evident in the constructional systems used, although traditional structural features continued to be employed—*e.g.,* cupolas raised on squinches or more commonly pendentives, barrel and groin vaults, and wooden ceilings covering large areas supported by columns and piers. The main innovations are of three kinds. First, minarets became particularly elaborate and, toward the end of the period, almost absurd in their ornamentation. Facades were huge, with overwhelming portals 25 to 35 feet high.

A second characteristically Mamlūk feature was technical virtuosity in stone construction. At times this led to a superb purity of form, as in the Gate of the Cotton Merchants in Jerusalem or the complex of the Barqūq mosque in Cairo. At other times, as in the Mamlūk ar-



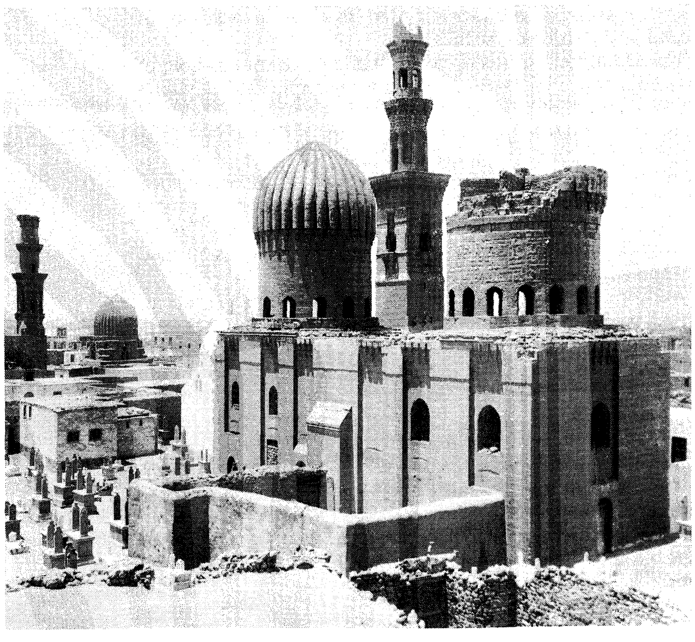
The madrasah of Sultan Ḥasan, Cairo, 1356–62.
(Left) Courtyard and (right) plan.
(Left) GEKS, (right) from B. Fletcher, *A History of Architecture on the Comparative Method;* Athlone Press of the University of London, London

chitecture of Baybars and Qā'it Bāy, there was an almost wild playfulness with forms. Another aspect of Mamlūk masonry was the alternation of stones of different colours to provide variations on the surfaces of buildings.

The third element of change in Mamlūk art was perhaps the most important: almost all formal artistic achievements rapidly became part of the common vocabulary of the whole culture, thus ensuring high quality of construction and decorative technique throughout the period.

With the exception of portals and *qiblah* walls, architectural decoration was usually subordinated to the architectural elements of the design. Generally the material of construction (usually stone) was carved with ornamental motifs. Stucco decoration was primarily used in early Mamlūk architecture, while coloured tile was a late decorative device that was rarely employed.

H. Roger-Viollet



Mamlūk tombs, Cairo, 14th–15th centuries.

Other arts. Like architecture, the other arts of the Mamlūk period achieved a high level of technical perfection but were often lacking in originality. The so-called "Baptistère de St. Louis" (*c.* 1310, Louvre) is the most impressive example of inlaid metalwork preserved from this period. Several Mamlūk illustrated manuscripts, such as the *Maqāmāt* (1334) in the Nationalbibliothek, Vienna, display an amazing ornamental sense in the use of colour on gold backgrounds. Mamlūk mosque lamps provide some of the finest examples of medieval glass. The wooden objects made by Mamlūk craftsmen were widely celebrated for the quality of their painted, inlaid, or carved designs. And the bold inscriptions that decorate the hundreds of remaining bronzes testify to the Mamlūks' mastery of calligraphy. None of these examples, however, exhibits much inventiveness of design.

Mongol Iran: Il-Khanid and Timurid periods. Seen from the vantage point of contemporary or later chronicles, the 13th century in Iran was a period of destructive wars and invasions. Such cities as Balkh, Nīshāpūr, or Rayy, which had been centres of Islāmic culture for nearly six centuries, were eradicated as the Mongol army swept through Iran. The turning point toward some sort of stability took place in 1295 with the accession of Maḥmūd Ghāzān to the Mongol throne. Under him and his successors (the Il-Khan dynasty), order was reestablished throughout Iran, and cities in northeastern Iran, especially Tabriz and Solṭānīyeh, became the main creative centres of the new Mongol regime. At Tabriz, for example, the Rashīdīyeh (a sort of academy of sciences and arts to which books, scholars, and ideas from all over the world were collected) was established in the early 14th century.

Existing under the Mongol rulers were a number of secondary dynasties that flourished in various provinces of Iran: the Jalāyirid dynasty, centred in Baghdad, controlled most of western Iran; the Moẓaffarid dynasty of southwestern Iran contained the cities of Isfahan, Yazd, and Shīrāz; and the Karts reigned in Khorāsān. Until the last decade of the 14th century, however, all the major cultural centres were in western Iran. Under Timur (1336–1405; the Timurid dynasty) and his successors, however, northeastern Iran, especially the cities of Samarkand and Herāt, became focal points of artistic and intellectual activity. But Timurid culture affected the whole of Iran either directly or through minor local dynasties. Many

The "Baptistère de St. Louis," copper inlaid with gold and silver, from Egypt, c. 1310, Mamlūk period. In the Louvre, Paris. Ht. 22.8 cm.

By courtesy of the Musee du Louvre, Paris; photograph, Cliche Musees Nationaux, Paris

Timurid monuments, therefore, are found in western or southern Iran.

Il-Khanid refinement and taste for the grandiose

Architecture. Stylistically, Il-Khanid architecture is defined best by buildings such as the mosque of Varāmīn (1322–26) and the mausoleums at Sarakhs, Merv, Rād-Kān, and Marāgheh. In all of these examples, the elements of architectural composition, decoration, and construction that had been developed earlier were refined by Il-Khanid architects. *Eyvān*s were shallower but better integrated with the courts; facades were more thoughtfully composed; the *muqarnas* became more linear and varied; and coloured tiles were used to enhance the building's character.

The architectural masterpiece of the Il-Khanid period is the mausoleum of Öljeitü at Solṭānīyeh. With its double system of galleries, eight minarets, large blue-tiled dome, and an interior measuring 80 feet (25 metres), it is clear that the building was intended to be imposing. Il-Khanid attention to impressiveness of scale also accounted for the ʿAlī Shāh mosque in Tabriz, whose *eyvān* measuring 150 by 80 by 100 feet (45 by 25 by 30 metres) was meant to be the largest ever built. The *eyvān* vault collapsed almost immediately after it had been constructed, but its walls, 35 feet (10 metres) thick, remain as a symbol of the grandiose taste of the Il-Khanids. In the regions of Isfahan and Yazd numerous smaller mosques (often with unusual plans) and less pretentious mausoleums, as well as palaces with elaborate gardens, were built in the 14th century. These buildings were constructed to provide a monumental setting for the Islāmic faith and for the authority of the state. The study of these buildings began only in the mid-20th century, and therefore no definitive conclusions have been reached as to whether regional or pan-Iranian stylistic and formal features predominated.

The Timurid period began architecturally in 1390 with the sanctuary of Aḥmad Yasavī in Turkistan. Between 1390 and the last works of Sultan Ḥusayn Bayqara almost a century later, hundreds of buildings were constructed at Herāt, many of which have been preserved, although few have been studied except by Soviet scholars. The most spectacular examples of Timurid architecture are found in Samarkand, Herāt, Meshed, Khargird, Tayābād, Baku, and Tabriz, although important Timurid structures were also erected in southern Iran.

Timurid concern with a commemorative architecture

Architectural projects were well patronized by the Timurids as a means to commemorate their respective reigns. Every ruler or local governor constructed his own sanctuaries, mosques, and, especially, memorial buildings dedicated to holy men of the past. While the Shāh-e Zendah in Samarkand—a long street of mausoleums comparable to the Mamlūk cemetery of Cairo—is perhaps the most accessible of the sites of Timurid commemorative architecture, more spectacular ones are to be seen at Meshed, Torbat-e Sheykh Jām, and Mazār-e Sharīf. The Timurid princes also erected mausoleums for themselves, such as the Gūr-e Amīr and the ʿIshrat-Khāneh in Samarkand.

Major Timurid buildings, such as the so-called mosque of Bībī Khānom, the Gūr-e Amīr mausoleum, the mosque of Gowhar Shād in Meshed, or the *madrasah*s at Khargird and Herāt, are all characterized by strong axial symmetry. Often the facade on the inner court repeats the design of the outer facade, and minarets are used to frame the composition. Changes took place in the technique of dome construction. The *muqarnas* was not entirely abandoned but was often replaced by a geometrically rigorous net of intersecting arches that could be adapted to various shapes by modifying the width or span of the dome. The Khargird *madrasah* and the ʿIshrat-Khāneh mausoleum in Samarkand are particularly striking examples of this structural development. The Timurids also made use of double domes on high drums.

In the Timurid period the use of colour in architecture reached a high point. Every architectural unit was divided, on both the exterior and interior, into panels of brilliantly coloured tiles that sometimes were mixed with stucco or terra-cotta architectural decorations.

Painting. A new period of Persian painting began in the Mongol era, and, even though here and there one can recognize the impact of Seljuq painting, on the whole it is a limited one. Although the new style was primarily expressed in miniature painting, it is known from literary sources that mural painting flourished as well. Masterpieces of Persian literature were illustrated: first the *Shāh-nāmeh* ("Book of Kings") by the 11th-century poet Ferdowsī and then, from the second half of the 14th century, lyrical and mystical works, primarily those by the 12th-century poet Neẓāmī. Historical texts or chronicles such as the *Jāmiʿ at-tawārīkh* ("Universal History of Rashīd ad-Dīn") were also illustrated, especially in the early Mongol period.

The first major monument of Persian painting in the Mongol period is a group of manuscripts of the *Jāmiʿ at-tawārīkh* (British Museum, London; University Library, Edinburgh; and Topkapı Saray Museum, Istanbul). The miniatures are historical narrative scenes. Stylistically they are related to Chinese painting—an influence introduced by the Mongols during the Il-Khanid period.

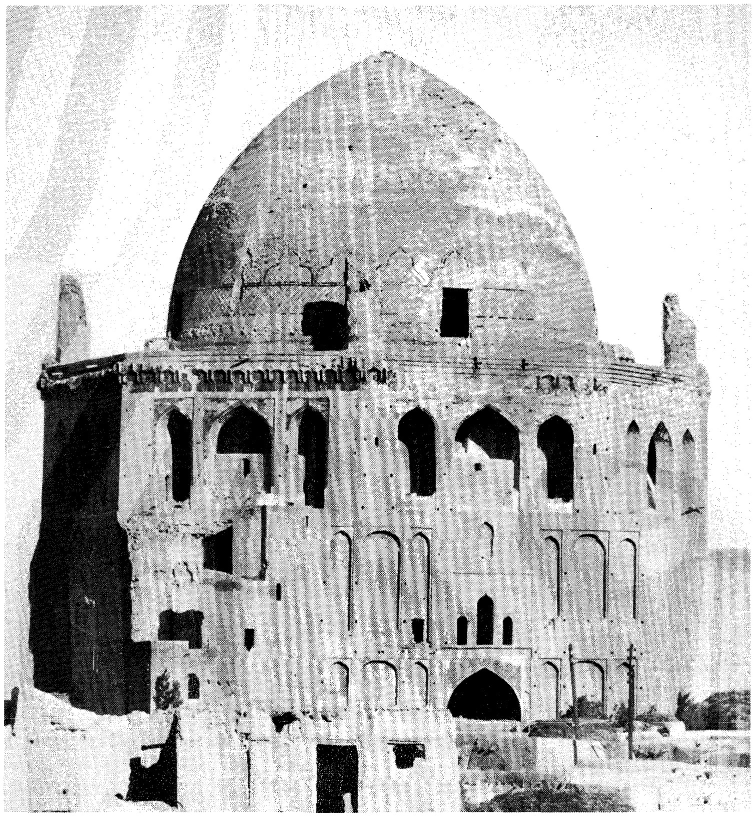
The expressive importance of the *Shāh-nāmeh* miniatures

Chinese influence can still be discovered in the masterpiece of 14th-century Persian painting, the so-called Demotte *Shāh-nāmeh.* Illustrated between 1320 and 1360, its 56 preserved miniatures have been dispersed all over the world. The compositional complexity of these paintings can be attributed to the fact that several painters probably were involved in the illustration of this manuscript and that these artists drew from a wide variety of different stylistic sources (*e.g.,* Chinese, European, local Iranian traditions). Its main importance lies in the fact that it is the earliest known illustrative work that sought to depict in a strikingly dramatic fashion the meaning of the Iranian epic. Its battle scenes, its descriptions of fights with monsters, its enthronement scenes are all powerful representations of the colourful and often cruel legend of Iranian kingship. The artists also tried to express the powerlessness of man confronted by fate in a series of mourning and death scenes.

The Demotte *Shāh-nāmeh* is but the most remarkable of a whole series of 14th-century manuscripts, all of which suggest an art of painting in search of a coherent style. At the very end of the period a manuscript such as that of the poems of Sultan Aḥmad (Freer Gallery of Art, Washington, D.C.) still exhibits an effective variety of established themes, while some of the miniatures in the Deutsche Staatsbibliothek, East Berlin, and in the Topkapı Saray, Istanbul, illustrate the astounding variety of styles studied or copied by Persian masters.

The lyric style

A more organized and stylistically coherent period in Persian painting began around 1396 with the Khwāju Kermānī manuscript (British Museum) and culminated between 1420 and 1440 in the paintings produced by the Herāt school, where the emperor Baysunqur created an academy in which classical Iranian literature was codified, copied, and illustrated. Although several *Shāh-nāmeh*s are known from this time, the mood of these manuscripts is no longer epic but lyrical. Puppet-like figures almost unemotionally engage in a variety of activities always set in an idealized garden or palace depicted against a rich gold background. It is a world of sensuous pleasure that also embodies the themes of a mystically interpreted lyrical poetry, for what is represented is not the real world but a divine paradise in the guise of a royal palace or garden. These miniatures easily became clichés, for later artists



Mausoleum of Öljeitü at Solṭānīyeh, 1305–13, Il-Khanid period.
Josephine Powell, Rome

endlessly repeated stereotyped formulas. But at its best, as in the Metropolitan Museum Neẓāmī, this style of Persian painting succeeds in defining something more than mere ornamental colourfulness. It expresses in its controlled lyricism a fascinating search for the divine, similar to the search of such epic characters as Neẓāmī, Rūmī, or Ḥāfeẓ—at times earthly and vulgar, at other times quite ambiguous and hermetic, but often providing a language for the ways in which human beings can talk about God.

Behzād and his school

Another major change in Persian painting occurred during the second half of the 15th century at Herāt under Ḥusayn Bayqara. This change is associated with the first major painter of Islāmic art, Behzād. Many problems of attribution are still posed about Behzād's art, and, in the examples that follow, works by his school, as well as images by the master's own hand, are included. In the Garrett *Ẓafar-nāmeh* (*c.* 1490), the Egyptian Cairo National Library's Būstān (1488), or the British Museum's Neẓāmī (1493–94), the stereotyped formulas of the earlier lyric style were endowed with new vitality. Behzād's interest in observing his environment resulted in the introduction of more realistic poses and the introduction of numerous details of daily life or genre elements. His works also reflect a concern for a psychological interpretation of the scenes and events depicted. It is thus not by chance that portraits have been attributed to Behzād.

Persian art of the Mongol period differs in a very important way from any of the other traditions of the middle period of Islāmic art. Even though Iran, like all other areas at that time, was not ethnically homogeneous, its art tended to be uniquely "national." In architecture nationalism was mostly a matter of function, for during this period the Shīʿites grew in importance, and new monumental settings were required for their holy places. Iranian individualism is especially apparent in painting, in which Chinese and other foreign styles were consistently adapted to express intensely Iranian subjects, thereby creating a uniquely Persian style.

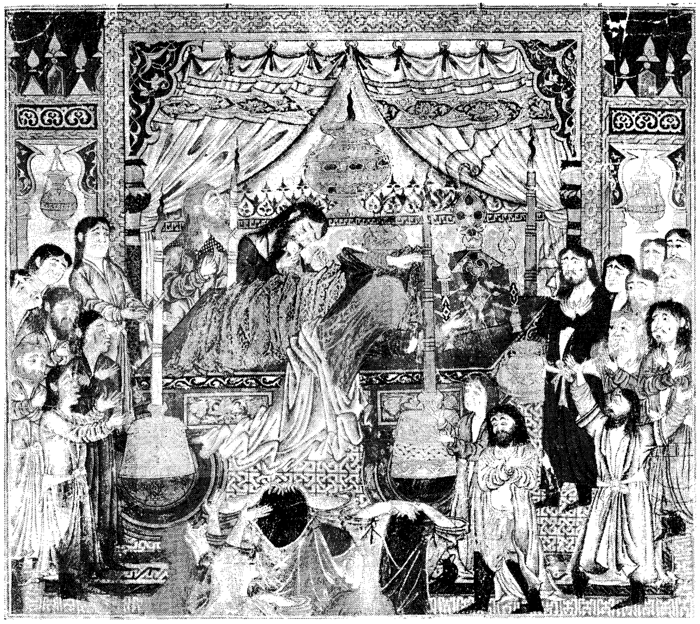
LATE PERIOD

The last period of an Islāmic artistic expression created within a context of political and intellectual independence was centred in the Ottoman, Ṣafavid, and Mughal empires. Although culturally very different from each other, these three imperial states shared a common past, a common consciousness of the nature of their ancestry and of the artistic forms associated with it. Painters and architects moved from one empire to the other, especially from Iran to India; Ottoman princes wrote Persian poetry, and Ṣafavid rulers spoke Turkish. But most of all, they were aware of the fact that they were much closer to each other than to any non-Islāmic cultural entity. However different their individual artistic forms may have been, they collected each other's works, exchanged gifts, and felt that they belonged to the same world.

Ottoman art. The Ottomans were originally only one of the small Turkmen principalities (*beyliks*) that sprang up in Anatolia around 1300 after the collapse of Seljuq rule. In many ways, all the *beyliks* shared the same culture, but it was the extraordinary political and social attributes of the Ottomans that led them eventually to swallow up the other kingdoms, to conquer the Balkans, to take Constantinople in 1453, and to control almost the whole of the Arab world by 1520. Only in the 19th century did this complex empire begin to crumble. Thus, while Ottoman art, especially architecture, is best known through the monuments in Turkey, there is, in fact, evidence of Ottoman art extending from Algiers to Cairo in North Africa, to Damascus in the Levant, and in the Balkans from Sarajevo, Yugos., to Sofia, Bulg.

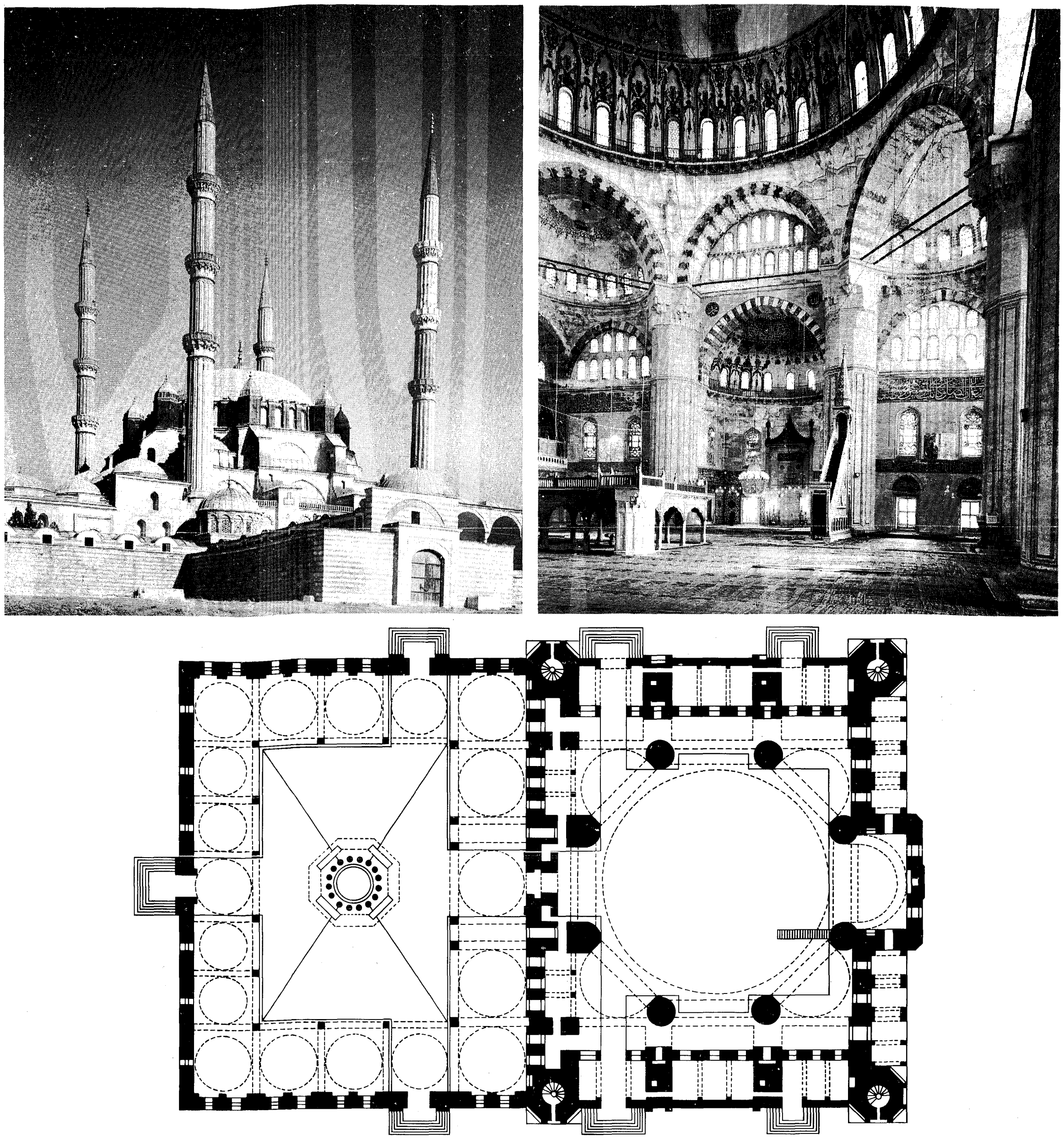
Origin of the *tekke* and *külliye*

Architecture. The grand tradition of Ottoman architecture, established in the 16th century, was derived from two main sources. One was the rather complex development of new architectural forms that occurred all over Anatolia, especially at Manisa, İznik, Bursa, and Selçuk in the 14th and early 15th centuries. In addition to the usual mosques, mausoleums, and *madrasahs*, a number of buildings called *tekkes* were constructed to house dervishes (members of mystical fraternities) and other holy men who lived communally. The *tekke* (or *zeviye*) was often joined to a mosque or mausoleum. The entire complex was then called a *külliye.* All these buildings continued to develop the domed, central-plan structure, constructed by the Seljuqs in Anatolia. The other source of Ottoman architecture is Christian art. The Byzantine tradition, especially as embodied in Hagia Sophia, became a major source of inspiration. Byzantine influence appears in such features as stone and brick used together or in the use of pendentive dome construction. Also artistically influential were the contacts that the early Ottomans had with



By courtesy of the Smithsonian Institution, Freer Gallery of Art, Washington, D.C.

Mourning scene at the bier of Alexander the Great, miniature from the Demotte *Shāh-nāmeh* ("Book of Kings") of Ferdowsī, colour and gold on paper, Tabriz school, 14th century. In the Freer Gallery of Art, Washington, D.C. 25 cm × 28 cm.



Selim Mosque at Edirne, Tur., designed by Sinan, 1569–75. (Top left) Exterior; (top right) interior; (bottom) plan.

(Top) K. Scholz—Shostal/EB Inc., (bottom) from G. Goodwin, *A History of Ottoman Architecture;* Johns Hopkins University Press

Italy. Thus, in several mosques at Bursa, Tur., there are stylistic parallels in the designs of the exterior facade and of windows, gates, and roofs to features found in Italian architecture. A distinctive feature of Ottoman architecture is that it drew from both Islāmic and European artistic traditions and was, therefore, a part of both.

The apogee of Ottoman architecture was achieved in the great series of *külliye*s and mosques that still dominate the Istanbul skyline: the Fatih *külliye* (1463–70), the Bayezid Mosque (after 1491), the Selim Mosque (1522), the Şehzade *külliye* (1548), and the Süleyman *külliye* (after 1550). The Şehzade and Süleyman *külliye*s were built by Sinan, the greatest Ottoman architect, whose masterpiece is the Selim Mosque at Edirne, Tur. (1569–75). All of these buildings exhibit total clarity and logic in both plan and elevation; every part has been considered in relation to the whole, and each architectural element has acquired a hierarchic function in the total composition. Whatever is unnecessary has been eliminated. This simplicity of design in the late 15th and 16th centuries has often been attributed to the fact that Sinan and many Ottoman architects were first trained as military engineers. Everything in these buildings was subordinated to an imposing central dome. A sort of cascade of descending half domes, vaults, and ascending buttresses leads the eye up and down the building's exterior. Minarets, slender and numerous, frame the exterior composition, while the open space of the surrounding courts prevents the building from being swallowed by the surrounding city. These masterpieces of Ottoman architecture seem to be the final perfection of two great traditions: a stylistic and aesthetic tradition that had been indigenous to Istanbul since the construction of the Byzantine church of Hagia Sophia in the 6th century and the other Islāmic tradition of domical construction dating to the 10th century.

The tragedy of Ottoman architecture is that it never managed to renew its 16th-century brilliance. Later buildings, such as the impressive Sultan Ahmed mosque in Istanbul, were mostly variations on Sinan's architecture, and sometimes there were revivals of older building types, especially in the provinces. Occasionally, as in the early 18th-century Nûruosman mosque in Istanbul, interesting new variants appear illustrating the little-known Turkish Baroque style. The latter, however, is more visible in ornamental details or in smaller buildings, especially the numerous fountains built in Istanbul in the 18th century.

The sources of the Turkish Baroque are probably to be sought in the Baroque architecture of Vienna and the bordering Austro-Hungarian states. Throughout the 18th and 19th centuries, a consistent Europeanization of a local tradition occurs in the Ottoman empire.

While mosques and *külliye*s are the most characteristic monuments of Ottoman architecture, important secular buildings were also built: baths, caravansaries, and especially the huge palace complex of Topkapı Saray at Istanbul, in which 300 years of royal architecture are preserved in its elaborate pavilions, halls, and fountains.

Other arts. Architectural decoration was generally subordinated to the structural forms or architectonic features of the building. A wide variety of themes and techniques originating from many different sources were used. One decorative device, the Ottoman version of colour-tile decoration, deserves particular mention, for it succeeds in transforming smaller buildings such as the mosque of Rüstem Paşa in Istanbul into a visual spectacle of brilliant colours. The history and development of this type of ceramic decoration is intimately tied to the complex and much controverted problem of the growth of several distinctive Ottoman schools of pottery: İznik, Rhodian, and Damascus ware. Both in technique and in design, Ottoman ceramics are the only major examples of pottery produced in the late Islāmic period.

Wide-spread use of tile decoration

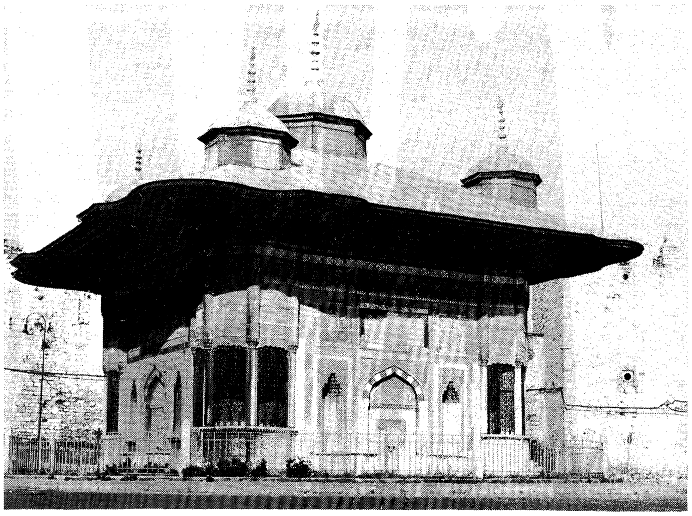
Ottoman miniature painting does not compare in quality with Persian painting, which originally influenced the Turkish school. Yet Ottoman miniatures do have a character of their own, either in the almost folk art effect of religious images or in the precise depictions of such daily events as military expeditions or great festivals. Among the finest examples of the latter is the manuscript *Surname-i Vehbi* (Topkapı Saray Museum, Istanbul) painted by Levnî in the early 18th century.

The production of metalwork, wood inlaid with ivory, Usak carpets, and textiles flourished under the Ottomans, both in Istanbul workshops sponsored by the sultan and in numerous provincial centres. The influence of these ornamental objects on European decorative arts from the 16th through the 19th century was considerable.

Ṣafavid art. The Ṣafavid dynasty was founded by Esmāʿīl I (1501–24). The art of this dynasty reached its zenith during the reigns of Ṭahmāsp (1524–76) and of ʿAbbās I (1588–1629). This phase of the Ṣafavid period also marked the last significant development of Islāmic art in Iran, for after the middle of the 17th century original creativity disappeared in all mediums. Rugs and objects in silver, gold, and enamel continued to be made and exhibited a considerable technical virtuosity, even when they were lacking in inventiveness.

The Ṣafavids abandoned Central Asia and northeastern Iran to a new Uzbek dynasty that maintained the Timurid style in many buildings (especially at Bukhara) and briefly sponsored a minor and derivative school of painting. Only the great sanctuary of Meshed was being kept up and built-up, but, like many of the other religious sanctuaries of the time—Qom, an-Najaf, Karbalāʾ, it is still far too little known to lend itself to coherent analysis. For this is the time when Shīʿism became a state religion and for the first time in Islām there appeared an organized ecclesiastical system rather than the more or less loose spiritual and practical leadership of old. The main centres of the Ṣafavid empire were Tabriz and Ardabīl in the northwest, with Kazvin in the central region, and, especially, Isfahan in the west. The Ṣafavid period, like the Ottoman era, was an imperial age, and therefore there is hardly a part of Iran where either Ṣafavid buildings or major Ṣafavid restorations cannot be found. The dynasty spent much money and effort on the building of bridges, roads, and caravansaries to encourage trade.

Josephine Powell, Rome



Turkish Baroque style exemplified by the Fountain of Ahmed III, Istanbul, 1728.

Architecture. The best known Ṣafavid monuments are located at Isfahan, where ʿAbbās I built a whole new city. According to one description, it contained 162 mosques, 48 *madrasah*s, 1,802 commercial buildings, and 283 baths. Most of these buildings no longer survive, but what has remained constitutes some of the finest monuments of Islāmic architecture.

At the centre of Isfahan is the Meydān-e Shāh, a large open space, about 1,670 by 520 feet (510 by 158 metres), originally surrounded by trees. Used for polo games and parades, it could be illuminated with 50,000 lamps. Each side of the *meydān* was provided with the monumental facade of a building. On one of the smaller sides was the entrance to a large mosque, the celebrated Masjed-e Shāh. On the other side was the entrance into the bazaar or marketplace. On the longer sides were the small funerary mosque of Sheykh Loṭfollāh and, facing it, the ʿAlī Qāpū, the "high gate," the first unit of a succession of palaces and gardens that extended beyond the *meydān,* most of which have now disappeared except for the Chehel Sotūn, the palace of the "Forty Columns." The ʿAlī Qāpū was, in its lower floors, a semipublic place to which petitions could be brought, while its upper floors are a world of pure fantasy—a succession of rooms, halls, and balconies overlooking the city, which were purely for the prince's pleasure.

The Meydān-e Shāh

The Meydān-e Shāh unites in a single composition all the concerns of medieval Islāmic architecture: prayer, commemoration, princely pleasure, trade, and spatial effect. None of the hundreds of other remaining Ṣafavid monuments can match its historical importance, and in it also are found the major traits of Ṣafavid construction and decoration. The forms are traditional, for the most part, and even in vaulting techniques and the use of coloured tiles it is to Timurid art that the Ṣafavids looked for their models. The Persian architects of the early 17th century sought to achieve a monumentality in exterior spatial composition (an interesting parallel to the interior spaciousness created at the same time by the Ottomans); a logical precision in vaulting that was successful in the Masjed-e Shāh but rapidly led to cheap effects or to stucco imitations; and a coloristic brilliance that has made the domes and portals of Isfahan justly famous.

Painting. In the 16th and 17th centuries, possibly for the first time in Islāmic art, painters were conscious of historical styles—even self-conscious. Miniatures from the past were collected, copied, and imitated. Patronage, however, was fickle. A royal whim would gather painters together or exile them. Many names of painters have been preserved, and there is little doubt that the whim of patrons was being countered by the artists' will to be socially and economically independent as well as individually recognized for their artistic talents. Too many different impulses, therefore, existed in Ṣafavid Iran for painting to follow any clear line of development.

Importance of historical styles of painting

Three major painting styles, or schools (excluding a number of interesting provincial schools), existed in the Ṣafavid period. One school of miniature painting is exemplified by such masterpieces as the Houghton *Shāhnāmeh* (completed in 1537), the Jāmī *Haft owrang* (1556–1665; Freer Gallery of Art, Washington, D.C.), or the illustrations to stories from Ḥāfeẓ which have not been identified in detail (Fogg Art Museum, Cambridge, Mass., and in a private collection). However different they are from each other, these large, colourful miniatures all were executed in a grand manner. Their compositions are complex, individual faces appear in crowded masses, there is much diversification in landscape, and, despite a few ferocious details of monsters or of strongly caricaturized poses and expressions, these book illustrations are concerned with an idealized vision of life. The sources of this school lie with the Timurid academy. Behzād, Sulṭān Muḥammad, Sheykhzādeh, Mīr Sayyid ʿAlī, Āqā Mīrak, and Maḥmūd Muṣavvīr continued and modified, each in his own way, the ideal of a balance between an overall composition and precise rendering of details.

The miniatures of the second tradition of Ṣafavid painting seem at first to be like a detail out of the work of the previously discussed school. The same purity of colour, elegance of poses, interest in details, and assertion of the individual figure is found. Āqā Reẕā and Reẕā ʿAbbāsī (both active around 1600) excelled in these extraordinary portrayals of poets, musicians, courtiers, and aristocratic life in general.

Portraiture and genre painting

In both traditions of painting, the beautiful personages depicted frequently are satirized; this note of satirical criticism is even more pronounced in portraiture of the time. But it is in pen or brush drawings, mostly dating from the 17th century, that the third aspect of Ṣafavid painting appeared: an interest in genre, or the depiction of minor events of daily life (*e.g.,* a washerwoman at work, a tailor sewing, an animal). With stunning precision Ṣafavid artists showed a whole society falling apart with a cruel sympathy totally absent from the literary documents of the time.

While architecture and painting were the main artistic vehicles of the Ṣafavids, the making of textiles and carpets was also of great importance. It is in the 16th century that a hitherto primarily nomadic and folk medium of the decorative arts was transformed into an expression of royal and urban tasks by the creation of court workshops. The predominantly geometric themes of earlier Iranian carpets were not abandoned entirely but tended to be replaced by vegetal, animal, and even occasional human motifs. Great schools of carpetmaking developed particularly at Tabriz, Kāshān, and Kermān.

Mughal art. Since the culture of the Mughals was intimately connected to the indigenous Hindu traditions of the Indian subcontinent, their art will be treated only synoptically in this article. (For a more detailed account, the reader should see the section on Mughal art in the visual arts portion of the article SOUTH ASIAN ARTS).

The art of the Mughals was similar to that of the Ottomans in that it was a late imperial art of Muslim princes. Both styles were rooted in several centuries (at least from the 13th century onward) of adaptation of Islāmic functions to indigenous forms. It was in the 14th-century architecture of South Asian sites such as Tughluqābād, Gaur, and Ahmadābād that a uniquely Indian type of Islāmic hypostyle mosque was created, with a triple axial nave, corner towers, axial minarets, and cupolas. It was also during these centuries that the first mausoleums set in scenically spectacular locations were built. By then the conquering Muslims had fully learned how to utilize local methods of construction, and they adapted South Asian decorative techniques and motifs.

Mughal art was in continuous contact with Iran or, rather, with the Timurid world of the second half of the 15th century. The models and the memories were in Herāt or Samarkand, but the artists were raided from Ṣafavid Iran, and the continuous flow of painters from Iran to the Mughal empire is a key factor in understanding Mughal painting.

The mausoleum of Humāyūn in Delhi (1565–69), the city of Fateḥpūr Sīkri (from 1569 onward), and the Tāj Mahal at Āgra (1631–53) summarize the development of Mughal architecture. In all three examples it can be seen that what Mughal architecture brought to the Islāmic tradition (other than traditional Indian themes, especially in decoration) was technical perfection in the use of red sandstone or marble as building and decorative materials.

Mughal painting

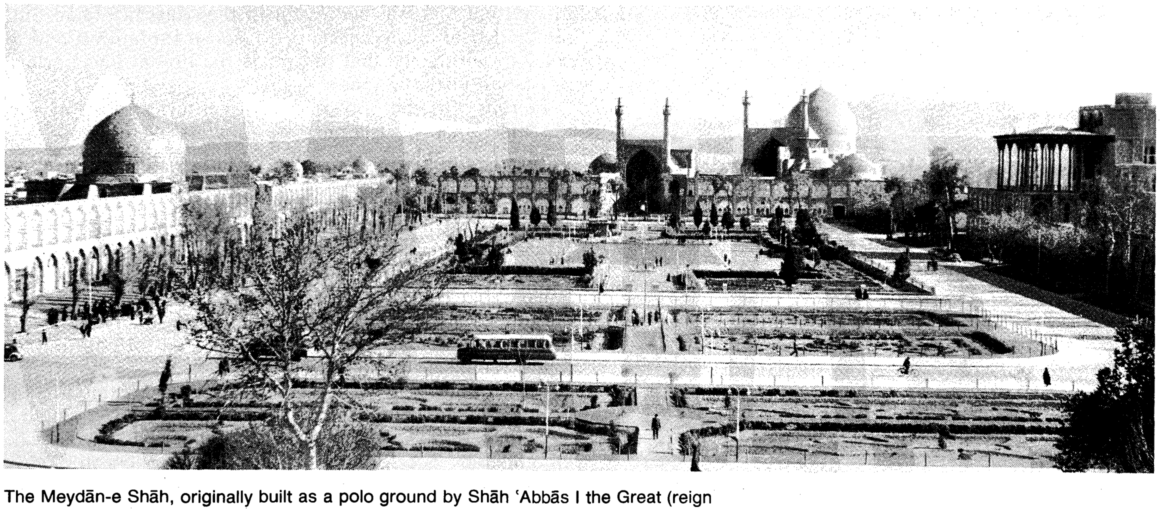
In Mughal painting the kind of subject that tended to be illustrated was remarkably close to those used in Ṣafavid history books—legendary stories, local events, portraits, genre scenes. What evolved quickly was a new manner of execution, and this style can be seen as early as about 1567, when the celebrated manuscript *Dāstān-e Amīr Ḥamzeh* ("Stories of Amīr Ḥamzeh") was painted (some 200 miniatures remain and are found in most major collections of Indian miniatures, especially at the Freer Gallery of Art, Washington, D.C.). Traditional Iranian themes—battles, receptions, feasts—acquired monumentality, not only because of the inordinate size of the images but also because almost all of the objects and figures depicted were seen in terms of mass rather than line. Something of the colourfulness of Iranian painting was lost, but instead images acquired a greater expressive power. Mughal portraiture gave more of a sense of the individual than did the portraits of the Ṣafavids. As in a celebrated representation of a dying courtier in the Boston Museum of Fine Arts, Mughal drawings could be poignantly naturalistic. Mood was important to the Mughal artist—in many paintings of animals there is a playful mood; a sensuous mood is evident in the first Muslim images to glorify the female body and the erotic.

In summary it can be said that the Mughals produced an art of extraordinary stylistic contrasts that reflected the complexities of its origins and of its aristocratic patronage.

ISLĀMIC ART UNDER EUROPEAN INFLUENCE AND CONTEMPORARY TRENDS

It is extremely difficult to decide when, how, and to what extent European art began to affect the art of the

Inge Morath—Magnum



The Meydān-e Shāh, originally built as a polo ground by Shāh ʿAbbās I the Great (reign 1588–1629), at Isfahan, Iran. Facing the square on the left is the mosque of Sheykh Loṭfollāh, in the centre the Masjed-e Shāh, and at the right the palace of ʿAlī Qāpū.

traditional Muslim world. Ottoman architecture was from the beginning affected by Western influences. In Mughal India, European landscapes and Western spatial concerns influenced painting in the 18th century; and Persian painting has exhibited constant Western influence since the 17th century. Thus, Islāmic art began to be affected by European traditions before Europe began (in the 18th and 19th centuries) its conquests of most of the Muslim world. Since the Ottomans ruled North Africa (except Morocco), Egypt, Syria, Palestine, as well as the Balkans, much of the Muslim world was first introduced to "modern" European art through its adaptation in Istanbul or in other major Ottoman cities like Smyrna or Alexandria.

European influence tended to have been mostly limited to architecture. Nineteenth-century European engineers and architects, for example, adapted modern structural technology and decorative styles to local Islāmic needs or idioms: the Sūq al-Ḥamīdīyah bazaar in Damascus was built with steel roofing; the Hejaz railway station at Damascus was decorated in a sort of Oriental Art Nouveau style.

Revival of the decorative arts

During actual European occupation of Muslim territory, there was a conscious revival of traditional decorative arts, but new techniques were often employed. This especially occurred in India and Morocco, where the retail success of an art object depended less on the local tradition than on the taste of the Europeans. What was romantic to a European, therefore, was no longer part of the world of the newly enriched and Europeanized Muslim. Much of the Europeanized architecture was drab and pretentious. The only real artistic accomplishment of this period was in the preservation and encouragement of the traditional techniques and designs of the decorative arts. The latter often had to be maintained artificially through government subsidies, for the local market, except in Morocco or India, was more easily seduced by second-rate European objects.

During the period of occupation it was questioned whether alien techniques necessarily brought with them new forms. This mood was clearly expressed in literature but less so in the visual arts, since the quality of Muslim art had deteriorated so much in the decades preceding European arrival that there was no longer a lively creative force to maintain. As various schools based on the École des Beaux-Arts in Paris were formed, however, the faculties and the students suffered from constant uncertainty as to whether they should preserve an art that was mostly artisanal or revolutionize it altogether.

20th-century currents

It is much more difficult to define in broad terms the characteristics of art in Muslim countries after the formation of independent countries in the 1940s and '50s. Extensive planning programs and building projects have been undertaken in even the poorest countries; and the wealthy Arab states, as well as pre-revolutionary Iran, transformed their traditional cities and countryside with spectacular modern complexes ranging from housing projects to universities. Many of these buildings were planned and constructed by Western firms and architects, and some are mere copies of European and American models, ill-adapted to the physical conditions and visual traditions of the Muslim world. Others are interesting and even sensitive projects: spectacular and technically innovative, such as the Intercontinental Hotel in Mecca (Frei and Otto) and the Haj Terminal of the King Abdul Aziz International Airport at Jidda, Saudi Arabia (the U.S. firm of Skidmore, Owings & Merrill); or intelligent and imaginative, such as the government buildings of Dhākā, Bangladesh (designed by the late Louis Kahn of the United States), or in the numerous buildings designed by the Frenchman André Ravereau in Mali or Algeria. Furthermore, within the Muslim world emerged several schools of architects that adopted modes of an international language to suit local conditions. The oldest of these schools are in Turkey, where architects such as Eldhem and Cansever, among many others, built highly successful works of art. Other major Muslim contributors to a contemporary Islāmic architecture are the Iranians Nader Ardalan and Kemzan Diba, the Iraqis Rifat Chaderji and Muhammad Makkiya, the Jordanian Rassem Badran, or the Bangladeshi Mazhar ul-Islam. Finally, a unique message was being transmitted by the visionary Egyptian architect Hassan Fathy, who, in eloquent and prophetic terms, urged that the traditional forms and techniques of vernacular architecture be studied and adapted to contemporary needs. Directly or indirectly, his work has inspired many young architects in the Muslim world and has led to a host of fascinating private houses, mosques, and educational facilities. The Aga Khan Award for Architecture was instituted to encourage genuine and contemporary architectural innovation in Muslim lands.

The results of dozens of new art schools and of a more enlightened patronage than during the 19th century are perhaps less spectacular in the other arts, and especially in painting. In spite of several interesting attempts to deal with calligraphy, with geometric designs, or with local folk arts, successes so far have not been clearly identified. But Turkey, Jordan, Egypt, Morocco, Iraq, Pakistan, and Indonesia all have produced talented artists.

EVALUATION

In order to evaluate and to understand a millenary artistic tradition spread over an area extending from Spain to India, the emphasis of this article has had to be on those features that relate the monuments to each other rather than on the myriad of characteristics that differentiate them. A few words about the latter are essential, however, for very soon after the formation of Islāmic culture (certainly by 1000), it seems clear that the nature of aesthetic impulses and of visual expectations began to vary. The question is one of determining what may be called the break-off points: the areas, moments, or forces that led to differentiations. One such point is the early 14th century, for almost everywhere in Islām artistic functions, forms, and techniques were renewed. And it is quite easy to separate the arts that followed the turn of the century from those that preceded it.

Importance of Iranian artistic traditions

Next to this chronological break-off, there are cultural ones, one might almost say ethnic ones, even though their ethnic association is often debatable. The clearest instance is that of Iran, whose artists and craftsmen, almost from the time of the first groups of Nīshāpūr ceramics, used distinctive techniques, styles, and especially subjects, many of which can be traced to pre-Islāmic times. The existence of a forceful Iranian personality in Islāmic art is self-evident, and its impact is found in almost all other subdivisions of the culture. Although it was not a single or even (until the 16th century) a politically or socially unified personality, it found uniqueness, possibly because it soon became (as early as in the 9th century) strongly conscious of its ancient past. The fact of that consciousness seems more important than the individual and on the whole scarce motifs it picked up from the past. A more curious example is that of the Ottomans and of the Arabs. For their ethnic past, in Central Asia and Arabia, respectively, played only a minor part in the formation of their art and was often intellectually rejected. At the same time and with notable exceptions, neither entity consistently sought models and ideas in the pre-Islāmic art of the region they had occupied. If they succeeded in creating an original artistic expression, it is in large part because of their success in creating a viable social order: the Ottoman imperial system of the 15th century, the urban order supported by military feudalism of Egypt, Syria, and North Africa. In these areas it is less a land than a society that provided the visual arts with their own distinctiveness, and it is only in recent years that Ottoman art began to be seen as Turkish and Mamlūk art as Arab. The case of India lies somewhere between the Iranian and Ottoman instances. Created by an imperial overlay on a powerful alien culture, it never entirely escaped the forms of the latter.

Thus, one can distinguish the following large cultural entities within Islāmic art: Ottoman, western Islāmic, Egypt and Fertile Crescent, Iran, India. They were all distinctive by the early 14th century. Detailed studies, of course, manage to find many additional subdivisions in time and space, and much mid-20th-century scholarship tended to work in those directions.

The unity of functions

Among the features that appear to unite these various traditions and especially to separate them collectively from other large artistic and cultural units is the unity of functions. There was created, in other words, an Islāmic religious and social function that is unique to Muslim lands. It was a diversified function, and its monuments are not alike in their forms. But they are alike in the human activity for which they were built. Limited in symbolic forms (*miḥrāb,* minaret, calligraphy as decoration), the Muslim function could be adapted to any architectural or ornamental tradition; and it was, not only in the cultures examined above, but in China, Indonesia, Africa—wherever Islām spread. The key concept here is that of a community of attitudes and of the uses of forms rather than of the making of forms.

There is a corollary to this conclusion that leads to the second level of an attempt to identify Islāmic visual arts as a whole; namely, that, as Islām limited its system of religious visual symbols, it developed a set of secular values. From the very beginning there occurred a major art of trade and of the city, as well as an art of the palace. More than any other culture and certainly earlier than any other, the Muslim world created a number of secular tastes and sponsored techniques of secular beautification. The result lies, on the one hand, in a striking succession of palaces from Khirbat al-Mafjar to the Alhambra or to Fatehpur Sīkri. It lies also in the impetus given to techniques of ceramics, textiles, and metalwork. These all tended to be the techniques of the artisan, and their importance lies not so much in the manufacture of an occasional object of art as in the raising of the level of quality of all industrial or decorative arts. This particular feature of the Islāmic tradition survived all political misfortunes. Remarkably beautiful objects were made as late as the early 19th century, and the techniques and traditions have often been revived in the 20th century with considerable success. Historically, Islāmic art became a sort of secular consciousness of artistic traditions elsewhere. Renaissance madonnas, for instance, were provided sometimes with halos containing Arabic inscriptions; bodies of saints were buried in Muslim cloth; Christian princes collected objects of Islāmic art; and *turquerie,* or Turkish themes, lay behind one of the styles of European decorative arts in the Baroque period of the 17th and 18th centuries. All this was possible also because the themes of Islāmic art almost never possessed the specificity of meaning that would make them unsuitable for use by others. Ambiguous in their abstraction of subjects and of styles, works of Islāmic art tended at times to the facile multiplication of known formulas. Yet again at this level, it was the user who determined the value of the form used.

The Islāmic visual vocabulary

All this is not to say that Islāmic art did not develop an internal visual vocabulary with a depth of its own. From the mosaics of the Umayyad Mosque of Damascus to the Alhambra or to certain Persian ceramics, one can determine the existence of concrete symbolic systems, royal and religious. It is even possible to see in the abstract arabesque or in certain uses of calligraphy attempts to express an early Muslim vision of the divine, while the glorious colour of Iranian mosques may reflect the more complex mystical thought of Shīʿism. There is no doubt that further research will provide many more examples of a meaningful visual symbolic system in the Muslim world. But in most instances that have already been studied—in particular Umayyad and Seljuq art—the remarkable point has been that such symbols did not last and that they were soon misunderstood or ignored. This refusal to be committed to visual symbols is reflected in the little that is known about Islāmic writing on art. It is only very incidentally that references are made to the value or meaning of visual expression; there are no theories on art, and even the religious injunctions against representations are a minute and almost incidental aspect of religious literature. Much more is known about individuals—ceramicists and metalworkers in early times, painters and architects in later times. The emphasis has always been on their technical skill, on their ability to do visual tricks, or on the speed and efficiency with which they created. The artist was regarded not as a prophet or a genius but as a technically equipped individual who succeeds in beautifying the surroundings of all men. It is in this manner that one can perhaps best define the Muslim artistic tradition: it avoided the conscious search for a unique masterpiece, and it did not build monuments for the eternal glory of God. It sought instead to please man and to make every moment of his life as attractive and enjoyable as possible. There is a hedonistic element in Islāmic art, therefore, but this hedonism is intellectually and emotionally mitigated by the conscious knowledge of the perishable character of all things human. In this fashion, Islāmic art seen as a whole is a curious paradox, for as it softened and embellished life's activities, it was created with destructible materials, thereby reiterating Islām's conviction that only God remains. (O.Gr.)

BIBLIOGRAPHY

Literature: JAMES KRITZECK (comp.), *Modern Islamic Literature: From 1800 to the Present* (1970), is a useful anthology of poetry and prose from different parts of the Muslim world. (*Arabic literature*): CARL BROCKELMANN, *Geschichte der arabischen Litteratur,* 2nd ed. (1943–49, and suppl., 1937–42), is the standard reference work containing information about almost every Arabic writer from pre-Islāmic to modern times. This work has been enlarged by FUAT SEZGĪN, *Geschichte des arabischen Schrifttums* (1967–), who has included many hitherto unknown books and manuscripts. R.A. NICHOLSON, *A Literary History of the Arabs,* 2nd ed. (1930, reprinted 1969), emphasizes poetry in the classical age; his *Studies in Islamic Poetry* (1921, reprinted 1963) contains the best analysis of al-Maʿarri's poetry. H.A.R. GIBB, *Arabic Literature,* 2nd rev. ed. (1963), is concise and informative; the German translation, *Arabische Literaturgeschichte* (1968), of Gibb's book has been enlarged by a section on modern Arabic literature by JACOB M. LANDAU and has an extensive bibliography on works of Islāmic literature translated into western European languages. GOTTHOLD WEIL, *Grundriss und System der altarabischen Metren* (1958), is an introduction to Arabic prosody. JOHANN FÜECK, *Arabiya: Untersuchungen zur arabischen Sprach- und Stilgeschichte* (1950), is an indispensable study of the development of a High Arabic style. GUSTAVE E. VON GRUNEBAUM, *Kritik und Dichtkunst: Studien zur arabischen Literaturgeschichte* (1955), contains essays on Arabic literature especially of the ʿAbbāsid period. Von Grunebaum's "Spirit of Islam as Shown in Its Literature," in his *Islam: Essays on the Nature and Growth of a Cultural Tradition,* 2nd ed. (1961, reprinted 1981), is an important essay mainly concerned with Ḥarīrī's *Maqāmāt,* and his "Acculturation as a Theme in Contemporary Arab Literature," in *Diogenes,* 39: 84–118 (1962), is a study on the problem of westernization in modern Arabic literature. ʿABDALQĀHIR AL-JURJĀNĪ, *Die Geheimnisse der Wortkunst* (*Asrār al-balāġa*) . . . (1959), is a German translation by HELLMUT RITTER of this classic on Arabic rhetoric. It is available also in the English translation, *The Mysteries of Eloquence,* ed. by HELLMUT RITTER (1954). ADOLF F. VON SCHACK, *Poesie und Kunst der Araber in Spanien und Sicilien,* 2nd ed., 2 vol. (1877), though often superseded by modern research, remains a charming introduction to the culture and art of Moorish Spain. U.M. DAUDPOTA, *The Influence of Arabic Poetry on the Development of Persian Poetry* (1934), attempts to show the formal influences of Arabic on early Persian poetry. ROGER ALLEN, *The Arabic Novel: An Historical and Critical Introduction* (1982), covers 1938–80. WOLFHART HEINRICHS, *Arabische Dichtung und griechische Poetik* (1969), is an important introduction to the literary criticism of classical Arabic literature.

(*Persian literature*): JAN RYPKA, *History of Iranian Literature* (1968; originally published in Czech, 1956), the standard work on Persian literature from its origins to the 20th century, includes folk literature and Tajik and Indo-Persian literature. See also CHARLES A. STOREY, *Persian Literature: A Bio-Bibliographical Survey,* 2 vol. in 4 (1970–72). EDWARD G. BROWNE, *A Literary History of Persia,* 4 vol. (1902–24, reprinted 1969–78), is an informative classic. A.J. ARBERRY, *Classical Persian Literature* (1958, reprinted 1967), is a classic work by one of the most prolific translators of Arabic and Persian poetry into English. ANTONINO PAGLIARO and ALESSANDRO BAUSANI, *Storia della letteratura Persiana* (1960), contains many interesting and unusual viewpoints. HERMANN ETHÉ, "Neupersische Literatur," and THEODOR NÖLDEKE, "Das iranische Nationalepos," in WILHELM GEIGER and ERNST KUHN (eds.), *Grundriss der iranischen Philologie,* vol. 2 (1896–1904, reprinted 1974), provides a masterly survey of classical Persian literature, including Indo-Persian. FRITZ M. MEIER (ed. and trans.), *Die schöne Mahsatī,* vol. 1 (1963), an immensely learned work centring around the poet Mahsatī, deals with the development of the *rubāʾī* and other forms of Persian poetry. HELLMUT RITTER, *Über die Bildersprache Nizāmīs* (1927), is the classic work on the imagery

in Neẓāmī's poetry. ANNEMARIE SCHIMMEL, *Stern und Blume* (1984), deals with imagery in Persian poetry. CHARLES H. DE FOUCHÉCOUR, *La Description de la nature dans la poésie lyrique persane du XI^e^ siècle* (1969), is a study of nature imagery in particular in early Persian poetry. FRIEDRICH RUCKERT, *Grammatik, Poetik und Rhetorik der Perser,* ed. by WILHELM PERTSCH (1874, reprinted 1966), a translation and commentary of a late Indo-Persian manual of rhetoric, is noted for its acute observations and amusing details. FINN THIESEN, *A Manual of Classical Persian Prosody* (1982), is an introduction to problems of Persian, as well as Turkish and Urdu, prosody.

(*Turkish literature*): E.J.W. GIBB, *A History of Ottoman Poetry,* 6 vol. (1900–09, reprinted 1958–67), the classical study of the historical developments of Turkish literature from its beginnings to 1900, includes many translations of poems. OTTO SPIES, *Die Türkische Prosa-literatur der Gegenwart* (1943), deals with Turkish prose after the revolution.

Music: "Bibliography of Asiatic Musics," in the MUSIC LIBRARY ASSOCIATION, *Notes,* 2nd series, vol. 5–6 (1947–49), an extensive bibliography compiled by five scholars, includes an important section on Islāmic music, with 592 references divided into categories dealing with music among Muslims in general, Arabic-speaking peoples, Turkic peoples, and Iranians and others. HENRY GEORGE FARMER, *A History of Arabian Music to the XIIIth Century* (1929, reprinted 1967), is still regarded as a key historical study. His "Music of Islam," in *The New Oxford History of Music,* vol. 1, pp. 421–77 (1957, reprinted 1966), is a good concise survey, as is PETER CROSSLEY-HOLLAND, "The Arabic World," in the *Pelican History of Music,* vol. 1, pp. 118–36 (1960, reprinted 1978). RODOLPHE VON ERLANGER (ed. and trans.), *La Musique Arabe,* 6 vol. (1930–59), includes French translations of the Arabic treatises by al-Fārābī, Avicenna, Ṣafī od-Dīn, and others (vol. 1–4) and devotes the last two volumes to an analytical study of contemporary Arabian music. CURT SACHS, *The Rise of Music in the Ancient World, East and West* (1943), has a large section on Arabic music in the context of an intercultural study. MEHDI BARKECHLI (ed.), *La Musique traditionnelle de l'Iran* (1963), gives a comprehensive musical transcription of the *Radīf* (modal systems of the Iranian traditional music). ADNAN SAYGUN, "La Musique Turque," in the *Encyclopédie de la Pléiade,* vol. 9, pp. 573–617 (1960); and ALEXIS CHOTTIN, *Tableau de la musique marocaine* (1939), discuss regional and local particularities and have useful bibliographies. AMNON SHILOAH, *Caractéristiques de l'art vocal arabe au moyen-âge* (1963), is an important essay on medieval Islāmic vocal music, and his *Theory of Music in Arabic Writings (c. 900–1900)* (1979), is an extensive analytical catalog of manuscripts and published sources on Arabic music. See also O. WRIGHT, *The Model System of Arab and Persian Music, A.D. 1250–1300* (1978), an analytical presentation of the system based on Persian and Arabic medieval treatises; KURT REINHARD and URSULA REINHARD, *Turquie* (1969), a comprehensive presentation of Turkish music in its diverse aspects; and ELLA ZONIS, *Classical Persian Music* (1973), a comprehensive historical study.

Dance and theatre: The classic work on the shadow play in the Middle East is still GEORG JACOB, *Geschichte des Schattentheaters im Morgan- und Abendland,* 2nd ed. (1925, reprinted 1972). METIN AND, *A History of Theatre and Popular Entertainment in Turkey* (1963–64), is a perceptive, scholarly account of the Turkish theatre in all its manifestations, and his *Pictorial History of Turkish Dancing* (1976) is an excellent study on the subject. CHRISTA URSULA SPULER, *Das türkische Drama der Gegenwart* (1968), treats in more detail 20th-century Turkish playwrights and theatrical literature. NICHOLAS N. MARTINOVICH, *The Turkish Theatre* (1933, reprinted 1968); and HELLMUT RITTER, *Karagös,* 3 vol. (1924–53), comprise translations of Turkish shadow plays into English and German, respectively. IGNACZ KUNOS, *Das türkische Volksschauspiel Orta ojnu* (1908), is an introduction to the *ortaoyunu* popular shows, with samples translated into German.

As for the Persian theatre and dance (mainly the latter), the most up-to-date book is MEDJID REZVANI, *Le Théâtre et la danse en Iran* (1962). PETER J. CHELKOWSKI (ed.), *Taʿziyeh: Ritual and Drama in Iran* (1979), is a collection of scholarly writings on the subject. CHARLES VIROLLEAUD, *Le Théâtre persan, ou le drama de Kerbéla* (1950), is a good sampling of *taʿziyah*s in French translation. The Arab theatre and dance (chiefly the former) are discussed in JACOB M. LANDAU, *Studies in the Arab Theater and Cinema* (1958), which also includes a detailed list of Arabic plays.

Visual arts: Among the numerous works dealing with Islāmic art as a whole, only one can be recommended as having a text of considerable merit—KATHARINA OTTO-DORN, *Kunst des Islam* (1964). An important, though partial, interpretation is found in TITUS BURCKHARDT, *Art of Islam: Language and Meaning* (1976). ALEXANDRE PAPADOPOULO, *Islam and Muslim Art* (1979, originally published in French, 1976), presents excellent photographic surveys. See also MILO CLEVELAND BEACH, *The Imperial Image: Paintings from the Mughal Court* (1981). K.A.C. CRESWELL, *A Bibliography of the Architecture, Arts and Crafts of Islam* (1961), with a supplement covering 1960–72 (1973), is a good bibliographical source; current literature on Islāmic art in all languages is surveyed in the *Abstracta Islamica,* published as an annual supplement to the *Revue des Études Islamiques* in Paris. Textual information about the arts has never been properly gathered. For a typical text on painters, see QĀDĪ AḤMAD, *Calligraphers and Painters,* trans. from the Persian by VLADIMIR MINORSKY (1959). The only lists of artists have been collected by LEO A. MAYER in several books, of which the most important are *Islamic Metalworkers and Their Works* (1959) and *Islamic Architects and Their Works* (1956). The vast majority of material on Islāmic art is to be found in periodicals rather than in books. The three publications that have dealt or deal systematically with all aspects of Islāmic art are *Ars Islamica* (irregular, 1934–51; reprinted in 16 vol., 1968), *Ars Orientalis* (irregular from 1954), and *Kunst des Orients* (annual from 1950). Articles are published in English, French, and German.

Area surveys: (*Spain*): MANUEL GÓMEZ-MORENO, *El arte árabe español hasta los almohades y arte mozárabe* (1951); and LEOPOLDO TORRES BALBÁS, *Arte almohade; arte nazarí; arte mudéjar* (1949). (*North Africa*): There is no recent general work dealing with all the arts; for architecture the indispensable manual is that of GEORGES MARÇAIS, *L'Architecture musulmane d'Occident* (1955), which deals also with Spain. (*Egypt*): DIETRICH BRANDENBURG, *Islamische Baukunst in Ägypten* (1966), is a convenient summary but does not supersede the exhaustive work of K.A.C. CRESWELL, *Muslim Architecture of Egypt,* 2 vol. (1952–59, reprinted 1979), going only up to the middle of the 14th century; and LOUIS HAUTECOEUR and GASTON WIET, *Les Mosquées du Caire,* 2 vol. (1932). A useful periodical is the *Journal of the American Research Center in Egypt.* (*Palestine, Syria*): There are no coherent works dealing with the whole area; JEAN SAUVAGET, *Alep* (1941), is a model (in French) of what can be done with a single city over the centuries; key journals are *Syria* (quarterly), *Levant* (annual), and *Quarterly of the Department of Antiquities in Palestine* (until 1950, when it was superseded in part by the annual publication of the Department of Antiquities of Jordan), and *Annales Archéologiques de Syrie* (annual). (*Iraq and upper Mesopotamia*): The main archaeological source is still FRIEDRICH SARRE and ERNST HERZFELD, *Archäologische Reise im Euphrat- und Tigris-Gebiet,* 4 vol. (1911–20); a model of archaeological history is ROBERT M. ADAMS, *Land Behind Baghdad* (1965). The main journals are *Sumer* (annual) and *Iraq* (semiannual). (*Anatolia*): ESIN ATIL (ed.), *Turkish Art* (1980), covers all fields evenly and has a good bibliography; see also YANNI PETSOPOULOS (ed.), *Tulips, Arabesques and Turbans: Decorative Arts from the Ottoman Empire* (1982). EKREM AKURGAL (ed.), *The Art and Architecture of Turkey* (1980), is a historical treatment of major and minor arts. The principal journals are *Anatolica* (annual) and *Anatolian Studies* (annual). (*Iran*): Nothing has superseded ARTHUR UPHAM POPE and PHYLLIS ACKERMAN (eds.), *A Survey of Persian Art from Prehistoric Times to the Present,* 3rd ed. (1977–). One may also consult ANDRÉ GODARD, *The Art of Iran* (1965, originally published in French, 1962); and the chapters by OLEG GRABAR in *The Cambridge History of Iran,* "The Visual Arts," vol. 4, pp. 329–63 (1975), and "The Visual Arts, 1050–1350," vol. 5, pp. 626–58 (1968). Important information is to be found in HANS E. WULFF, *The Traditional Crafts of Persia* (1966); and in several works by ARTHUR UPHAM POPE, such as *Persian Architecture* (1965). Periodicals of importance are the defunct *Athār-é Īrān* (1936–49), the *Bulletin of the American Institute for Persian Art and Archaeology,* and *Iran* (annual), the active journal of the British School. (*India*): Among several architectural surveys, PERCY BROWN, *Indian Architecture,* vol. 2, *The Islamic Period,* 6th ed. (1971), is the best. See also R. NATH, *History of Sultanate Architecture* (1978), and *History of Mughal Architecture* (1982); WAYNE E. BEGLEY, "Myth of the Taj Mahal and a New Theory of Its Symbolic Meaning," *The Art Bulletin,* 61:7–37 (March 1979); and ELIZABETH B. MOYNIHAN, *Paradise as a Garden* (1979).

Techniques: (*Architecture*): JOHN D. HOAG, *Islamic Architecture* (1976); GEORGE MICHELL (ed.), *Architecture of the Islamic World: Its History and Social Meaning* (1978); and NADER ARDALAN and LALEH BAKHTIAR, *The Sense of Unity: The Sufi Tradition in Persian Architecture* (1973, reprinted 1979). (*Painting*): RICHARD ETTINGHAUSEN, *Arab Painting* (1962); BASIL GRAY, *Persian Painting from Miniatures of the XIII–XVI Centuries* (1947); and DOUGLAS E. BARRETT and BASIL GRAY, *Painting of India* (1963, reissued 1978 as *Indian Painting*). See also OLEG GRABAR, *The Illustrations of the Maqamat* (1984). (*Metalwork*): EVA BAER, *Metalwork in Medieval Islamic Art* (1983); ASSADULLAH SOUREN MELIKIAN-CHIRVANI, *Islamic Metalwork from the Iranian World, 8–18th Centuries* (1982); and

JAMES W. ALLAN, *Islamic Metalwork: The Nuhad Es-Said Collection* (1982). The field owes much to the work of the late D.S. RICE: "Studies in Islamic Metalwork," *Bulletin of the School of Oriental and African Studies, University of London,* vol. 14–16 (1952–57); "Inlaid Brasses from the Workshop of Aḥmad al-Dhakī al-Mawṣilī," *Ars Orientalis,* 2:283–326 (1957); *The Wade Cup in the Cleveland Museum of Art* (1955); and *Le Baptistère de Saint Louis* (1951). See also RICHARD ETTINGHAUSEN, "The Wade Cup . . . ," *Ars Orientalis,* 2:327–366 (1957). (*Ceramics*): The key studies are ARTHUR LANE, *Early Islamic Pottery* (1947, reprinted 1965), and *Later Islamic Pottery,* 2nd ed. (1971). (*Carpets*): Among scholarly studies on carpets are those of KURT ERDMANN, especially *Oriental Carpets* (1960, reissued 1976; originally published in German, 2nd ed., 1960), and *Seven Hundred Years of Oriental Carpets* (1970; originally published in German, 1966). (*Ivories*): JOHN BECKWITH, *Caskets from Cordoba* (1960), is a scholarly study of Moorish ivory work. In addition, see ERNST KÜHNEL, *Die islamischen Elfenbeinskulpturen, VIII.–XIII. Jahrhundert* (1971).

Historical works: (*Early period*): Most of the problems are summarized in OLEG GRABAR, *The Formation of Islamic Art* (1973). For architecture the main books are K.A.C. CRESWELL, *Early Muslim Architecture* (vol. 1, 2nd ed., 1969; vol. 2, 1941; reissued 1979, 2 vol. in 3); R.W. HAMILTON, *Khirbat al-Mafjar* (1959); and JEAN SAUVAGET, *La Mosquée omeyyade de Médine* (1947). (*Middle period*): On the Faṭimids, see RICHARD ETTINGHAUSEN, "Painting in the Fatimid Period," *Ars Islamica,* 9:112–124 (1942). On the Seljuqs, for Iran, in addition to vol. 4 of *The Cambridge History of Iran,* see RICHARD ETTINGHAUSEN, "Some Comments on Medieval Iranian Art," *Artibus Asiae,* 31:276–300 (1969); for Syria and Egypt, one should consult ERNST HERZFELD, "Damascus," *Ars Islamica,* vol. 9–12 (1942–51); and JEAN SAUVAGET *et al., Les Monuments Ayyoubides de Damas,* 4 vol. (1938–50); and for Anatolia, KURT ERDMANN, *Das anatolische Karavansaray des 13 Jahrhunderts,* 3 vol. (1961–76). Major monuments are discussed by RICHARD ETTINGHAUSEN in "The Bobrinski Kettle," *Gazette des Beaux-Arts,* 24:193–208 (1943); "The Iconography of a Kāshān *Luster* Plate," *Ars Orientalis,* 4:25–64 (1961); "The Flowering of Seljuq Art," *Metropolitan Museum Journal,* 3:113–131 (1970); and ASSADULLAH SOUREN MELIKIAN-CHIRVANI (ed.), *Le Roman de "Varge et Golšâh"* (1970). Newer interpretations of the Alhambra are based on FREDERICK P. BARGEBUHR, *The Alhambra* (1968; originally published in Spanish, 1966); see also OLEG GRABAR, *The Alhambra* (1978). ESIN ATIL, *Renaissance of Islam* (1981), is a comprehensive study of Mamlūk art; see also SALEH L. MOSTAFA, *Kloster und Mausoleum des Farağ ibn Barqūq in Kairo* (1968). For Mongol architecture, see DONALD N. WILBER, *The Architecture of Islamic Iran: The Il Khānid Period* (1955, reprinted 1969); LISA GOLOMBEK, *The Timurid Shrine at Gazur Gah* (1969); and various accounts in the annual *Iran.* For painting, see ERNST J. GRUBE, *The Classical Style in Islamic Painting* (1968), but especially the rich volume of IVAN STCHOUKINE, *Les Peintures des manuscrits tîmûrides* (1954); and M.S. IPSIROGLU, *Painting and Culture of the Mongols* (1966; originally published in German, 1965). See also OLEG GRABAR and SHEILA BLAIR, *Epic Images and Contemporary History: The Illustrations of the Great Mongol Shahnama* (1980). (*Late period*): For Ottoman architecture, GODFREY GOODWIN, *A History of Ottoman Architecture* (1971); and APTULLAH KURAN, *The Mosque in Early Ottoman Architecture* (1968), supersede all previous work. Painting is covered in NURHAN ATASOY and FILIZ ÇAĞMAN, *Turkish Miniature Painting* (1974). For ceramics, see ARTHUR LANE, "The Ottoman Pottery of Isnik," *Ars Orientalis,* 2:247–282 (1957). For Ṣafavid architecture, see DONALD N. WILBER, *Persian Gardens and Garden Pavilions,* 2nd ed. (1979); RENATA HOLOD (ed.), *Studies on Isfahan* (1974); EUGENIO GALDIERI, *Eṣfahān, ʿAli Qāpū: An Architectural Survey* (1979); MARTIN BERNARD DICKSON and STUART C. WELCH (eds.), *The Houghton Shahnameh* (1981); and ANTHONY WELCH, *Artists for the Shah: Late Sixteenth-Century Painting at the Imperial Court of Iran* (1976). The most important publications on painting are both by IVAN STCHOUKINE, *Les Peintures des manuscrits Safavīs de 1502 à 1587* (1959), and *Les Peintures des manuscrits de Shāh ʿAbbās I^{er} à la fin des Safavīs* (1964). The principal work on India is STUART C. WELCH, *The Art of Mughal India* (1963, reprinted 1976). For contemporary architecture see RENATA HOLOD (ed.), *Architecture and Community* (1983), and the quarterly journal *Mimar,* published in Singapore.

(An.Sc./A.Sh./J.M.L./O.Gr.)

The Islāmic World

Adherence to Islām is a global phenomenon: Muslims predominate in some 30 to 40 countries, from the Atlantic to the Pacific and along a belt that stretches across northern Africa to the southern borders of the Soviet Union and the northern regions of the Indian subcontinent. Arabs account for fewer than one-fifth of all Muslims, more than half of whom live east of Karāchi, Pak. Despite the absence of large-scale Islāmic political entities, the Islāmic faith continues to expand, by some estimates faster than any other major religion.

The Muslim religion and the life of the Prophet Muḥammad are treated specifically in the article ISLĀM, MUḤAMMAD AND THE RELIGION OF. The literature, music, dance, and visual arts of Muslim peoples are treated in the article ISLĀMIC ARTS. Islām is also discussed in articles on individual countries or on regions in which the religion is a factor, such as EGYPT, IRAN, ARABIA, and NORTH AFRICA. Articles on individual branches or sects and concepts are found in the *Micropædia.* See, for example, AMERICAN MUSLIM MISSION; SUNNĪ; ḤADĪTH.

A very broad perspective is required to explain the history of today's Islāmic world. This approach must enlarge upon conventional political or dynastic divisions to draw a comprehensive picture of the stages by which successive Muslim communities, throughout Islām's 14 centuries, encountered and incorporated new peoples so as to produce an international religion and civilization.

In general, events in this article are dated according to the Gregorian calendar and eras are designated BCE (before the Common Era or Christian Era) and CE (Common Era or Christian Era), equivalent to BC (before Christ) and AD (Latin anno Domini). In some cases the Muslim reckoning of the Islāmic era is used, indicated by AH (Latin anno Hegirae). The Islāmic era begins with the date of Muḥammad's emigration (*hijrah*) to Medina, which corresponds to July 16, 622, in the Gregorian calendar. The term Islāmic refers to Islām as a religion. The term Islāmicate refers to the social and cultural complex historically associated with Islām and the Muslims, even when found among non-Muslims. Islāmdom refers to that complex of societies in which the Muslims and their faith have been prevalent and socially dominant.

The article is divided into the following sections:

Prehistory (c. 3000 BCE–CE 500)

The prehistory of Islāmdom is the history of central Afro-Eurasia from Hammurabi of Babylon to the Achaemenid Cyrus II in Persia to Alexander the Great to the Sāsānian emperor Nūshīrvān to Muḥammad in Arabia; or, in a Muslim view, from Adam to Noah to Abraham to Moses to Jesus to Muḥammad. The potential for Muslim empire building was established with the rise of the earliest civilizations in western Asia. It was refined with the emergence and spread of what have been called the region's Axial Age religions—Abrahamic, centred on the Hebrew patriarch Abraham, and Mazdean, focused on the Iranian deity Ahura Mazdāh—and their later relative, Christianity. It was facilitated by the expansion of trade from eastern Asia to the Mediterranean, and by the political changes thus effected. The Muslims were heirs to the ancient Egyptians, Babylonians, Persians, Hebrews, even the Greeks and Indians; the societies they created bridged time and space, from ancient to modern and from east to west.

THE RISE OF AGRARIAN-BASED CITIED SOCIETIES

In the 7th century CE a coalition of Arab groups, some sedentary and some migratory, inside and outside the Arabian Peninsula, seized political and fiscal control in western Asia, specifically of the lands between the Nile and Oxus (Amu Darya) rivers—territory formerly controlled by the Byzantines in the west and the Sāsānians in the east. The factors that surrounded and directed their accomplishment had begun to coalesce long before, with the emergence of agrarian-based citied societies in western Asia in the 4th millennium BCE. The rise of complex agrarian-based societies, such as Sumer, out of a subsistence agricultural and pastoralist environment, involved the founding of cities, the extension of citied power over surrounding villages, and the interaction of both with pastoralists.

This type of social organization offered new possibilities. Agricultural production and intercity trading, particularly in luxury goods, increased. Some individuals were able to take advantage of the manual labour of others to amass enough wealth to patronize a wide range of arts and crafts; of these, a few were able to establish territorial monarchies and foster religious institutions with wider appeal. Gradually the familiar troika of court, temple, and market emerged. The new ruling groups cultivated skills for administering and integrating non-kin-related groups. They benefited from the increased use of writing and, in many cases, from the adoption of a single writing system, such as the cuneiform, for administrative use. New institutions, such as coinage, territorial deities, royal priesthoods, and standing armies, further enhanced their power.

Court, temple, and market

In such town-and-country complexes the pace of change quickened enough so that a well-placed individual might see the effects of his actions in his own lifetime and be stimulated to self-criticism and moral reflection of an unprecedented sort. The religion of these new social entities reflected and supported the new social environments. Unlike the religions of small groups, the religions of complex societies focused on deities, such as Marduk, Isis, or Mithra, whose appeal was not limited to one small area or group and whose powers were much less fragmented. The relationship of earthly existence to the afterlife became more problematic, as evidenced by the elaborate death rites of Pharaonic Egypt. Individual religious action began to compete with communal worship and ritual; sometimes it promised spiritual transformation and transcendence of a new sort, as illustrated in the pan-Mediterranean mystery religions. Yet large-scale organization had introduced social and economic injustices that rulers and religions could address but not resolve. To many, an absolute ruler uniting a plurality of ethnic, religious, and interest groups offered the best hope of justice.

CULTURAL CORE AREAS OF THE SETTLED WORLD

By the middle of the 1st millennium BCE the settled world had crystallized into four cultural core areas: Mediterranean, Nile-to-Oxus, Indic, and East Asian. The Nile-to-Oxus, the future core of Islāmdom, was the least cohesive and the most complicated. Whereas each of the other regions developed a single language of high culture—Greek, Sanskrit, and Chinese, respectively—the Nile-to-Oxus region was a linguistic palimpsest of Irano-Semitic languages of several sorts: Aramaic, Syriac (eastern or Iranian Aramaic), and Middle Persian (the language of eastern Iran).

The Nile-to-Oxus region. The Nile-to-Oxus region differed in climate and ecology, too. It lay at the centre of a vast arid zone stretching across Afro-Eurasia from the Sahara to the Gobi; it favoured those who could deal with aridity—not only states that could control flooding (as in Egypt), or maintain irrigation (as in Mesopotamia), but also pastoralists and oasis dwellers. Although its agricultural potential was severely limited, its commercial possibilities were virtually unlimited. Located at the crossroads of the trans-Asian trade and blessed with numerous natural transit points, the region offered special social and economic prominence to its merchants.

The period from 800 to 200 BCE has been called the Axial Age because of its pivotal importance for the history of religion and culture. The world's first religions of salvation developed in the four core areas. From these traditions, for example, Judaism, Mazdeism, Buddhism, and Confucianism, derived all later forms of high religion, including Christianity and Islām. Unlike the religions that surrounded their formation, the Axial Age religions concentrated transcendent power into one locus, be it symbolized theistically or nontheistically. Their radically dualistic cosmology posited another realm, totally unlike the earthly realm and capable of challenging and replacing ordinary earthly values. The individual was challenged to adopt the right relationship with that "other" realm, so as to transcend mortality by earning a final resting place, or to escape the immortality guaranteed by rebirth by achieving annihilation of earthly attachment.

The Abrahamic and Mazdean traditions

In the Nile-to-Oxus region two major traditions arose during the Axial Age: the Abrahamic in the west and the Mazdean in the east. Because they required exclusive allegiance through an individual confession of faith in a just and judging deity, they are called confessional religions. The god of these religions was a unique all-powerful creator who remained active in history; and each event in the life of every individual was meaningful in terms of the judgment of God at the end of time. The universally applicable truth of these new religions was expressed in sacred writings. The traditions reflected the mercantile environment in which they were formed in their special concern for fairness, honesty, covenant keeping, moderation, law and order, accountability, and the rights of ordinary human beings. These values were always potentially incompatible with the elitism and absolutism of courtly circles. Most often, as for example in the case of the Achaemenid Empire, the conflict was expressed in rebellion against the crown or was adjudicated by viewing kingship as the guarantor of divine justice.

Although modern Western historiography has projected an East–West dichotomy onto ancient times, Afro-Eurasian continuities and interactions were well established by the Axial Age and persisted throughout premodern times. The history of Islāmdom cannot be understood without reference to them. Through Alexander's conquests in the 4th century BCE in three of the four core areas, the Irano-Semitic cultures of the Nile-to-Oxus region were permanently overlaid with Hellenistic elements, and a link was forged between the Indian subcontinent and Iran. By the 3rd century CE, crosscutting movements like Gnosticism and Manichaeism integrated individuals from disparate cultures. Similarly organized large, land-based empires with official religions existed in all parts of the settled world. The Christian Roman Empire was locked in conflict with its counterpart to the east, the Zoroastrian–Mazdean Sāsānian Empire. Another Christian empire in East Africa, the Abyssinian, was involved alternately with each of the others. In the context of these regional interrelationships inhabitants of Arabia made their fateful entrance into international political, religious, and economic life.

The Arabian Peninsula. The Arabian Peninsula consists of a large central arid zone punctuated by oases, wells, and small seasonal streams and bounded in the south by well-watered lands that are generally thin, sometimes mountainous coastal strips. To the north of the peninsula are the irrigated agricultural areas of Syria and Iraq, the site of large-scale states from the 4th millennium BCE. As early as the beginning of the 1st millennium BCE the southwest corner of Arabia, the Yemen, also was divided into settled kingdoms. Their language was a South Arabian Semitic dialect and their culture bore some affinity to Semitic societies in the Fertile Crescent. By the beginning of the Common Era (the 1st century AD in the Christian calendar) the major occupants of the habitable parts of the arid centre were known as Arabs. They were Semitic-speaking tribes of settled, semi-settled, and fully migratory peoples who drew their name and apparently their identity from what the camel-herding Bedouin pastoralists among them called themselves: *ʿarab.*

The Arab tribes

Until the beginning of the 3rd century of the Common Era the greatest economic and political power in the peninsula rested in the relatively independent kingdoms of the Yemen. The Yemenis, with a knowledge of the monsoon winds, had evolved an exceptionally long and profitable trade route from East Africa across the Red Sea and from India across the Indian Ocean up through the peninsula into Iraq and Syria, where it joined older Phoenician routes across the Mediterranean and into the Iberian Peninsula. Their power depended on their ability to protect islands discovered in the Indian Ocean and to control the straits of Hormuz and Aden as well as the Bedouin caravanners who guided and protected the caravans that carried the trade northward to Arab entrepôts like Petra and Palmyra. Participation in this trade was in turn an important source of power for tribal Arabs, whose livelihood otherwise depended on a combination of intergroup raiding, agriculture, and animal husbandry.

By the 3rd century, however, external developments began to impinge. In 226 Ardashīr I founded the Sāsānian Empire in Fars; within 70 years the Sāsānian state was at war with Rome, a conflict that was to last up to Islāmic times. The reorganization of the Roman Empire under Constantine the Great, with the adoption of a new faith, Christianity, and a new capital, Constantinople, exacerbated the competition with the Sāsānian Empire and resulted in the spreading of Christianity into Egypt and Abyssinia and the encouraging of missionizing in Arabia itself. There Christians encountered Jews who had been settling since the 1st century, as well as Arabs who had converted to Judaism. By the beginning of the 4th century the rulers of Abyssinia and Ptolemaic Egypt were interfering in the Red Sea area and carrying their aggression into the Yemen proper. In the first quarter of the 6th century the proselytizing efforts of a Jewish Yemeni ruler resulted in a massacre of Christians in the major Christian centre of Najrān. This event invited Abyssinian Christian reprisal and occupation, which put a virtual end to indigenous control of the Yemen. In conflict with the Byzantines, the Zoroastrian–Mazdean Sāsānians invaded Yemen toward the end of the 6th century, further expanding the religious and cultural horizons of Arabia, where membership in a religious community could not be apolitical and could even have international ramifications. The connection between communal affiliation and political orientations would be expressed in the early Muslim community and in fact has continued to function to the present day.

The long-term result of Arabia's entry into international politics was paradoxical: it enhanced the power of the tribal Arabs at the expense of the "superpowers." Living in an ecological environment that favoured tribal independence and small-group loyalties, the Arabs had never established lasting large-scale states, only transient tribal confederations. By the 5th century, however, the settled powers needed their hinterlands enough to foster client states: the Byzantines oversaw the Ghassānid kingdom; the Persians oversaw the Lakhmid; and the Yemenis (prior to the Abyssinian invasion) had Kindah. These relationships increased Arab awareness of other cultures and religions; and the awareness seems to have stimulated internal Arab cultural activity, especially the classical Arabic, or *muḍarī,* poetry, for which the pre-Islāmic Arabs are so famous. In the north, Arabic speakers were drawn into the imperial administrations of the Romans and Sāsānians; soon certain settled and semi-settled Arabs spoke and wrote Aramaic or Persian as well as Arabic, and some Persian or Aramaic speakers could speak and write Arabic. The prosperity of the 5th and 6th centuries, as well as the intensification of imperial rivalries in the late 6th century, seems to have brought the Arabs of the interior permanently into the wider network of communication that fostered the rise of the Muslim community at Mecca and Medina.

Formation and orientation (c. 500–634)

THE CITY OF MECCA: CENTRE OF TRADE AND RELIGION

Although the 6th-century client states were the largest Arab polities of their day, it was not from them that a permanently significant Arab state arose. Rather, it emerged among independent Arabs living in Mecca (Makkah) at the junction of major north–south and west–east routes, in one of the less naturally favoured Arab settlements of the Hejaz (al-Ḥijāz). The development of a trading town into a city-state was not unusual; but unlike many other western Arabian settlements, Mecca was not centred on an oasis or located in the hinterland of any non-Arab power. Although it had enough well water and springwater to provide for large numbers of camels, it did not have enough for agriculture; its economy depended on long-distance as well as short-distance trade.

Mecca under the Quraysh clans. Around the year 400 CE Mecca had come under the control of a group of Arabs who were in the process of becoming sedentary; they were known as Quraysh and were led by a man remembered as Quṣayy. During the generations before Muḥammad's birth in about 570, the several clans of the Quraysh fostered a development in Mecca that seems to have been occurring in a few other Arab towns as well. They used their trading connections and their relationships with their Bedouin cousins to make their town a regional centre whose influence radiated in many directions. They designated Mecca as a quarterly *ḥaram,* a safe haven from the intertribal warfare and raiding that was endemic among the Bedouin. Thus Mecca became an attractive site for large trade fairs that coincided with pilgrimage (*ḥajj*) to a local shrine, the Kaʿbah. The Kaʿbah housed the deities of visitors as well as the Meccans' supra-tribal creator and covenant-guaranteeing deity, called Allāh. Most Arabs probably viewed this deity as one among many, possessing powers not specific to a particular tribe; others may have identified this figure with the God of the Jews and Christians.

Quarterly haven at Mecca

The building activities of the Quraysh threatened one non-Arab power enough to invite direct interference: the Abyssinians are said to have invaded Mecca in the year of Muḥammad's birth. But the Byzantines and Sāsānians were distracted by internal reorganization and renewed conflict; simultaneously the Yemeni kingdoms were declining. Furthermore, these shifts in the international balance of power may have dislocated existing tribal connections enough to make Mecca an attractive new focus for supra-tribal organization, just as Mecca's equidistance from the major powers protected its independence and neutrality.

The Meccan link between shrine and market has a broader significance in the history of religion. It is reminiscent of changes that had taken place with the emergence of complex societies across the settled world several

Changes in religion: the sharing of deities

millennia earlier. Much of the religious life of the tribal Arabs had the characteristics of small-group, or "primitive," religion, including the sacralization of group-specific natural objects and phenomena and the multifarious presence of spirit beings, known among the Arabs as *jinn.* Where more complex settlement patterns had developed, however, widely shared deities had already emerged, such as the "trinity" of Allāh's "daughters" known as al-Lāt, Manāt, and al-ʿUzzāh. Such qualified simplification and inclusivity, wherever they have occurred in human history, seem to have been associated with other fundamental changes—increased settlement, extension and intensification of trade, and the emergence of lingua francas and other cultural commonalties, all of which had been occurring in central Arabia for several centuries.

New social patterns among the Meccans and their neighbours. The sedentation of the Quraysh and their efforts to create an expanding network of cooperative Arabs generated social stresses that demanded new patterns of behaviour. The ability of the Quraysh to solve their problems was affected by an ambiguous relationship between sedentary and migratory Arabs. Tribal Arabs could go in and out of sedentation easily, and kinship ties often transcended life-styles. The sedentation of the Quraysh did not involve the destruction of their ties with the Bedouin or their idealization of Bedouin life. Thus, for example, did wealthy Meccans, thinking Mecca unhealthy, often send their infants to Bedouin foster mothers. Yet the settling of the Quraysh at Mecca was no ordinary instance of sedentation. Their commercial success produced a society unlike that of the Bedouin and unlike that of many other sedentary Arabs. Whereas stratification was minimal among the Bedouin, a hierarchy based on wealth appeared among the Quraysh. Although a Bedouin group might include a small number of outsiders, such as prisoners of war, Meccan society was markedly diverse, including non-Arabs as well as Arabs, slave as well as free. Among the Bedouin, lines of protection for in-group members were clearly drawn; in Mecca, sedentation and socioeconomic stratification had begun to blur family responsibilities and foster the growth of an oligarchy whose economic objectives could easily supersede other motivations and values. Whereas the Bedouin acted in and through groups, and even regularized intergroup raiding and warfare as a way of life, Meccans needed to act in their own interest and to minimize conflict by institutionalizing new, broader social alliances and interrelationships. The market–shrine complex encouraged surrounding tribes to put aside their conflicts periodically and to visit and worship the deities of the Kaʿbah; but such worship, as in most complex societies, could not replace either the particularistic worship of small groups or the competing religious practices of other regional centres, such as aṭ-Ṭā'if.

Very little in the Arabian environment favoured the formation of stable, large-scale states. Therefore, Meccan efforts at centralization and unification might well have been transient, especially because they were not reinforced by any stronger power and because they depended almost entirely on the prosperity of a trade route that had been formerly controlled at its southern terminus and could be controlled elsewhere in the future, or exclude Mecca entirely. The rise of the Meccan system also coincided with the spread of the confessional religions, through immigration, missionization, conversion, and foreign interference. Alongside members of the confessional religions, unaffiliated monotheists, known as *ḥanīf*s, distanced themselves from the Meccan religious system by repudiating the old gods but embracing neither Judaism nor Christianity. Eventually in Mecca and elsewhere a few individuals came to envision the possibility of effecting supra-tribal association through a leadership role common to the confessional religions, that is, prophethood or messengership. The only such individual who succeeded in effecting broad social changes was a member of the Hāshim (Hāshem) clan of Quraysh named Muḥammad ibn ʿAbd Allāh ibn ʿAbd al-Muṭṭalib. One of their own, he accomplished what the Quraysh had started, first by working against them, later by working with them. When he was born, around 570, the potential for pan-Arab unification seemed nil; but after he died, in 632, the first generation of his followers were able not only to maintain pan-Arab unification but to expand far beyond the peninsula.

Concept of social unity through shared deity

THE PROPHET MUḤAMMAD

Muḥammad's years in Mecca. *Spiritual awakening.* Any explanation of such an unprecedented development must include an analysis not only of Muḥammad's individual genius but also of his ability to articulate an ideology capable of appealing to multiple constituencies. His approach to the role of prophet allowed a variety of groups to conceptualize and form a single community. Muḥammad was, according to many students of social behaviour, particularly well placed to lead such a social movement; in both ascribed and acquired characteristics he was unusual. Although he was a member of a high-status tribe, he belonged to one of its less well-placed clans. He was fatherless at birth; his mother and grandfather died when he was young, leaving him under the protection of an uncle. Although he possessed certain admirable personality traits to an unusual degree, his commercial success derived not from his own status but from his marriage to a much older woman, a wealthy widow named Khadījah. During the years of his marriage, his personal habits grew increasingly atypical; he began to absent himself in the hills outside Mecca to engage in the solitary spiritual activity of the *ḥanīf*s. At age 40, while on retreat, he saw a figure, whom he later identified as the angel Gabriel, who forced him to repeat these words: "Recite: In the name of God, the Merciful and Compassionate. Recite: And your Lord is Most Generous. He teaches by the pen, teaches man what he knew not." Although a few individuals, including his wife Khadījah, recognized his experience as that of a messenger of God, the contemporary religious life of most of the Meccans and the surrounding Arabs did not prepare them to share in this recognition easily.

Muḥammad's first recitation

Arabs did recognize several other types of intermediaries with the sacred. Some of the kings of the Yemen are said to have had priestly functions; and tribal leaders, *shaykh*s, in protecting their tribes' hallowed custom (*sunnah*), had a spiritual dimension. Tribal Arabs also had their *kāhin*s, religious specialists who delivered oracles in ecstatic rhymed prose (*sajʿ*) and read omens. They also had their *shā'ir*s, professionally trained oral poets who defended the group's honour, expressed its identity, and engaged in verbal duels with the poets of other groups. The power of the recited word was well established; the poets' words were even likened to arrows that could wound the unprotected enemy. Because Muḥammad's utterances seemed similar, at least in form, to those of the *kāhin*s, many of his hearers naturally assumed that he was one of the figures with whom they were more familiar. Indeed, Muḥammad might not even have attracted attention had he not sounded like other holy men; but by eschewing any source other than the one supreme being, whom he identified as Allāh ("the god") and whose message he regarded as cosmically significant and binding, he was gradually able to distinguish himself from all other intermediaries. Like many successful leaders Muḥammad broke through existing restraints by what might be called transformative conservatism. By combining familiar leadership roles with a less familiar one, he expanded his authority; by giving existing practices a new history, he reoriented them; by assigning a new cause to existing problems, he resolved them. His personal characteristics fit his historical circumstances perfectly.

Early reactions to Muḥammad's preachings

Public recitations. Muḥammad's first vision was followed by a brief lull, after which he began to hear messages frequently, entering a special physical state to receive them and returning to normalcy to deliver them orally. Soon he began publicly to recite warnings of an imminent reckoning by Allāh that disturbed the Meccan leaders. Muḥammad was one of their own, a man respected for his personal qualities. Yet weakening kinship ties and increasing social diversity were helping him attract followers from many different clans and also from among tribeless persons, giving all of them a new and potentially disruptive affiliation. The fundamentals of his message, delivered often in the vicinity of the Kaʿbah itself, questioned the very reasons for which so many people gathered there. If visitors to the Kaʿbah assumed, as so many Arabs did, that the deities represented by its idols were all useful and accessible in that place, Muḥammad spoke, as had Axial Age figures before, of a placeless and timeless deity that not only had created human beings, making them dependent on him, but would also bring them to account at an apocalypse of his own making. In place of time or chance, which the Arabs assumed to govern their destiny, Muḥammad installed a final reward or punishment based on individual actions. Such individual accountability to an unseen power that took no account whatsoever of kin relationships and operated beyond the Meccan system could, if taken seriously, undermine any authority the Quraysh had acquired. Muḥammad's insistence on the protection of the weak, which echoed Bedouin values, threatened the unbridled amassing of wealth so important to the Meccan oligarchy.

Efforts to reform Meccan society. Yet Muḥammad also appealed to the town dweller by describing the human being as a member of a polis (city-state) and by suggesting ways to overcome the inequities that such an environment breeds. By insisting that an event of cosmic significance was occurring in Mecca, he made the town the rival of all the greater cities with which the Meccans traded. To Meccans who believed that what went on in their town and at their shrine was hallowed by tribal custom, *sunnah,* Muḥammad replied that their activities in fact were a corrupt form of a practice that had a very long history with the god of whom he spoke. In Muḥammad's view, the Kaʿbah had been dedicated to the aniconic worship of the one God (Allāh) by Abraham, who fathered the ancestor of the Israelites, Isḥāq (Isaac), as well as the ancestor of the Arabs, Ismāʿīl (Ishmael). Muḥammad asked his hearers not to embrace something new, but to abandon the traditional in favour of the original. He appealed to his fellow Quraysh not to reject the *sunnah* of their ancestors, but rather to appreciate and fulfill its true nature. God should be worshiped not through offerings but through prayer and recitation of his messages, and his house should be emptied of its useless idols.

In their initial rejection of his appeal, Muḥammad's Meccan opponents took the first step toward accepting the new idea: they attacked it. For it was their rejection of him, as well as his subsequent rejection by many Jews and Christians, that helped to forge Muḥammad's followers into a community with an identity of its own and capable of ultimately incorporating its opponents. Muḥammad's disparate following was exceptionally vulnerable, bound together not by kinship ties but by a "generic" monotheism that involved being faithful (*muʾmin*) to the message God was sending through their leader. Their vulnerability was mitigated by the absence of formal municipal discipline; but their opponents within Quraysh could apply informal pressures ranging from harassment and violence against the weakest to a boycott against Muḥammad's clan, who were persuaded by his uncle Ṭālib to remain loyal even though most of them were not his followers. Meanwhile Muḥammad and his closest associates were thinking about reconstituting themselves as a separate community in a less hostile environment. In about 612 some 80 of his followers made an emigration (*hijrah*) to Abyssinia, perhaps assuming that they would be welcome in a place that had a history of hostility to the Meccan oligarchy and that worshiped the same god who had sent Muḥammad to them; but they eventually returned without establishing a permanent community. During the next decade, continued rejection intensified the group's identity and its search for another home. Although the boycott against Muḥammad's clan began to disintegrate, the deaths of his wife and his uncle, in about 619, removed an important source of psychological and social support. Muḥammad had already begun to preach and attract followers at market gatherings outside Mecca; now he intensified his search for a more hospitable environment. In 620 he met with a delegation of followers from Yathrib, an oasis about 200 miles to the northeast; in the next two years their support grew into an offer of protection.

Muḥammad's emigration to Yathrib (Medina). Like Mecca, Yathrib was experiencing demographic problems: several tribal groups coexisted, descendants of its Arab Jewish founders as well as a number of pagan Arab immigrants divided into two tribes, the Aws and the Khazraj. Unable to resolve their conflicts, the Yathribis invited Muḥammad to perform the well-established role of neutral outside arbiter (*ḥakam*). In 622, having sent his followers ahead, he and one companion, Abū Bakr, completed the community's second and final emigration, barely avoiding Quraysh attempts to prevent his departure by force. By the time of the emigration a new label had begun to appear in Muḥammad's recitations to describe his followers; in addition to being described in terms of their faithfulness (*īmān*) to God and his messenger, they were also described in terms of their undivided attention, that is, as *muslim*s, individuals who assumed the right relationship to God by surrendering (*islām*) to his will. Although the label *muslim,* derived from *islām,* eventually became a proper name for a specific historical community, at this point it appears to have expressed commonalty with other monotheists: like the others, *muslim*s faced Jerusalem to pray; Muḥammad was believed to have been transported from Jerusalem to the heavens to talk with God; and Abraham, Noah, Moses, David, and Jesus, as well as Muḥammad, all were considered to be prophets (*nabī*s) and messengers of the same God. In Yathrib, however, conflicts between other monotheists and the *muslim*s sharpened their distinctiveness.

Use of the terms *islām* and *muslim*

The forging of Muḥammad's community. As an autonomous community *muslim*s might have become a tribal unit like those with whom they had affiliated, especially because the terms of their immigration gave them no special status. Yet under Muḥammad's leadership they developed a social organization that could absorb or challenge everyone around them. They became Muḥammad's *ummah* ("community") because they had recognized and supported God's emissary (*rasūl Allāh*). The *ummah*'s members differed from one another not by wealth or genealogical superiority but by the degree of their faith and piety; and membership in the community was itself an expression of faith. Anyone could join, regardless of origin, by following Muḥammad's lead, and the nature of members' support could vary. In the concept of *ummah,* Muḥammad supplied the missing ingredient in the Meccan system: a powerful abstract principle for defining, justifying, and stimulating membership in a single community.

Muḥammad made the concept of *ummah* work by expanding his role as arbiter so as to become the sole spokesman for all residents of Yathrib, also known as Medina. Even though the agreement under which Muḥammad had emigrated did not obligate non-Muslims to follow him except in his arbitration, they necessarily became involved in the fortunes of his community. By protecting him from his Meccan enemies, the residents of Medina identified with his fate. Those who supported him as Muslims received special designations: the Medinans were called *anṣār* ("helpers"), and his fellow emigrants were distinguished as *muhājirūn* ("emigrants"). He was often able to use revelation to arbitrate. Because the terms of his emigration did not provide adequate financial support, he began to provide for his community through caravan raiding, a tactic familiar to tribal Arabs. By thus inviting hostility, he required all the Medinans to take sides. Initial failure was followed by success, first at Nakhlah, where the Muslims defied Meccan custom by violating one of the truce months so essential to Meccan prosperity and prestige. Their most memorable victory occurred in 624 at Badr, against a large Meccan force; they continued to succeed, with only one serious setback, at Uḥud in 625. From that time on, "conversion" to Islām involved joining an established polity, the successes of which were tied to its proper spiritual orientation, regardless of whether the convert shared that orientation completely. During the early years in Medina a major motif of Islāmic history emerged: the connection between material success and divine favour, which had also been prominent in the history of the Israelites.

The relation of material success to divine favour

The ummah's allies and enemies. During these years, Muḥammad used his outstanding knowledge of tribal relations to act as a great tribal leader, or *shaykh,* further expanding his authority beyond the role that the Medinans had given him. He developed a network of alliances between his *ummah* and neighbouring tribes, and so competed with the Meccans at their own game. He managed and distributed the booty from raiding, keeping one-fifth for the *ummah*'s overall needs and distributing the rest among its members. In return, members gave a portion of their wealth as *zakat,* to help the needy and to demonstrate their awareness of their dependence on God for all of their material benefits. Like other *shaykh*s, Muḥammad contracted numerous, often strategically motivated, marriage alliances. He was also more able to harass and discipline Medinans, Muslim and non-Muslim alike, who did not support his activities fully; he agitated in particular against the Jews, one of whose clans, the Banū Qaynuqa, he expelled.

Increasingly estranged from nonresponsive Jews and Christians, he reoriented his followers' direction of prayer from Jerusalem to Mecca. He formally instituted the *ḥajj* to Mecca and fasting during the month of Ramaḍān as distinctive cultic acts, in recognition of the fact that *islām,* a generic act of surrender to God, had become Islām, a proper-name identity distinguished not only from paganism but from other forms of monotheism as well. As more and more of Medina was absorbed into the Muslim community, and as the Meccans weakened, Muḥammad's authority expanded. He continued to lead a three-pronged campaign, against nonsupporters in Medina, against the Quraysh in Mecca, and against surrounding tribes; he even ordered raids into southern Syria. Eventually, Muḥammad became powerful enough to punish nonsupporters severely, especially those who leaned toward Mecca. For example, he had the men of the Qurayẓah clan of Jews in Medina executed after they failed to help him against the Meccan forces at the Battle of the Ditch in 627. But he also used force and diplomacy to bring in other Jewish and Christian groups. Because they were seen, unlike pagans, to have formed *ummah*s of their own around a revelation from God, Jews and Christians were entitled to pay for protection (*dhimmah*). Muḥammad thus set a precedent for another major characteristic of Islāmicate civilization, that of qualified religious pluralism under Muslim authority.

Muḥammad's later recitations. During these years of warfare and consolidation, Muḥammad continued to transmit revealed recitations, though their nature began to change. Some commented on Muḥammad's situation, consoled and encouraged his community, explained the continuing resistance of the Meccans, and urged appropriate responses. Some told stories about figures familiar to Jews and Christians, cast in an Islāmic framework. Though still delivered in the form of God's direct speech, the messages became longer and less ecstatic, less urgent in their warnings if more earnest in their guidance. Eventually they focused on interpersonal regulations in areas of particular importance for a new community, such as sexuality, marriage, divorce, and inheritance. By this time certain Muslims had begun to write down what Muḥammad uttered or to recite passages for cultic worship (*ṣalāt*) and private devotion. The recited word, so important among the Arab tribes, had found a greatly enlarged significance. A competitor for Muḥammad's status as God's messenger even declared himself among a nonmember tribe; he was Maslamah of Yamāmah, who claimed to convey revelations from God. He managed to attract numerous Bedouin Arabs but failed to speak as successfully as Muḥammad to the various available constituencies.

The importance of the recited word

Activism in the name of God, nonmilitary as well as military, would become a permanent strand in Muslim piety. Given the environment in which Muḥammad operated, his *ummah* was unlikely to survive without it; to compete as leader of a community he had to exhibit military prowess. (Like most successful leaders, however, Muḥammad was a moderate and a compromiser; some of his followers were more militant and aggressive than he, and some were less so.) Circumstantial necessity had ideological ramifications, too. Because Muḥammad as messenger was also, by divine providence, leader of an established community, he could easily define the whole realm of social action as an expression of faith. Thus Muslims were able to identify messengership with worldly leadership to an extent almost unparalleled in the history of religion. There had been activist prophets before Muḥammad, and there were activist prophets after him, but in no other religious tradition does the image of the activist prophet, and by extension the activist follower, have such a comprehensive and coherent justification in the formative period.

ISLĀM AT MUḤAMMAD'S DEATH

Muḥammad's continuing success gradually impinged on the Quraysh in Mecca. Some defected and joined his community. His marriage to a Quraysh woman provided him with a useful go-between. In 628 he and his followers tried to make an Islāmized *ḥajj* but were forestalled by the Meccans. At al-Ḥudaybiyah, outside Mecca, Muḥammad granted a 10-year truce on the condition that the Meccans would allow a Muslim pilgrimage the next year. Even at this point, however, Muḥammad's control over his followers had its limits; his more zealous followers agreed to the pact only after much persuasion. As in all instances of charismatic leadership, persisting loyalty was correlated with continuing success. In the next year the Meccans allowed a Muslim *ḥajj;* and in the next, 630, the Muslims occupied Mecca without a struggle. Muḥammad began to receive deputations from many parts of Arabia. By his death in 632 he was ruler of virtually all of it.

The Meccan Quraysh were allowed to become Muslims without shame. In fact, they quickly became assimilated to the actual *muhājirūn,* even though they had not emigrated to Yathrib themselves. Ironically, in defeat they had accomplished much more than they would have in victory: the centralization of all of Arabia around their polity and their shrine, the Kaʿbah, which had been emptied of its idols to be filled with an infinitely greater invisible power.

Because intergroup conflict was banned to all members of the *ummah* on the basis of their shared loyalty to the emissary of a single higher authority, the limitations of the Meccan concept of *ḥaram,* according to which the city quarterly became a safe haven, could be overcome. The broader solidarity that Muḥammad had begun to build was stabilized only after his death; and this was achieved, paradoxically, by some of the same people who had initially opposed him. In the next two years one of his most significant legacies became apparent: the willingness and ability of his closest supporters to sustain the ideal and the reality of one Muslim community under one leader, even in the face of significant opposition. When Muḥammad died, two vital sources of his authority ended—ongoing revelation and his unique ability to exemplify his messages on a daily basis. A leader capable of keeping revelation alive might have had the best chance of inheriting his movement; but no Muslim claimed messengership, nor had Muḥammad unequivocally designated any other type of successor. The *anṣār,* his early supporters in Medina, moved to elect their own leader, leaving the *muhājirūn* to choose theirs; but a small number of *muhājirūn* managed to impose one of their own over the whole. That man was Abū Bakr, one of Muḥammad's earliest followers and the father of his favourite wife, ʿĀ'ishah. The title Abū Bakr took, *khalīfah* (caliph), meaning deputy or successor, echoed revealed references to those who assist major leaders and even God himself. To *khalīfah* he appended *rasūl Allāh,* so that his authority was based on his assistance to Muḥammad as messenger of God.

ABŪ BAKR'S SUCCESSION

Abū Bakr soon confronted two new threats: the secession of many of the tribes that had joined the *ummah* after 630 and the appearance among them of other prophet figures who claimed continuing guidance from God. In withdrawing, the tribes appear to have been able to distinguish loyalty to Muḥammad from full acceptance of the uniqueness and permanence of his message. The appearance of other prophets illustrates a general phenomenon in the history of religion: the volatility of revelation as a source of authority. When successfully claimed, it has almost no competitor; once opened, it is difficult to close;

and, if it cannot be contained and focused at the appropriate moment, its power disperses. Jews and Christians had responded to this dilemma in their own ways; now it was the turn of the Muslims, whose future was dramatically affected by Abū Bakr's response. He put an end to revelation with a combination of military force and coherent rhetoric. He defined withdrawal from Muḥammad's coalition as ingratitude to or denial of God (the concept of *kufr*); thus he gave secession (*riddah*) cosmic significance as an act of apostasy punishable, according to God's revealed messages to Muḥammad, by death. He declared that the secessionists had become Muslims, and thus servants of God, by joining Muḥammad; they were not free *not* to be Muslims, nor could they be Muslims, and thus loyal to God, under any leader whose legitimacy did not derive from Muḥammad. Finally, he declared Muḥammad to be the last prophet God would send, relying on a reference to Muḥammad in one of the revealed messages as *khatm al-anbiyāʾ* ("Seal of the Prophets"). In his ability to interpret the events of his reign from the perspective of Islām, Abū Bakr demonstrated the power of the new conceptual vocabulary Muḥammad had introduced.

Ban on revelation and secession

Had Abū Bakr not asserted the independence and uniqueness of Islām, the movement he had inherited could have been splintered or absorbed by other monotheistic communities or by new Islām-like movements led by other tribal figures. Moreover, had he not quickly made the ban on secession and intergroup conflict yield material success, his chances for survival would have been very slim, because Arabia's resources could not support his state. To provide an adequate fiscal base, Abū Bakr enlarged impulses present in pre-Islāmic Mecca and in the *ummah.* At his death he was beginning to turn his followers to raiding non-Muslims in the only direction where that was possible, the north. Migration into Syria and Iraq already had a long history; and Arabs, both migratory and settled, were already present there. Indeed some of them were already launching raids when ʿUmar I, Abū Bakr's acknowledged successor, assumed the caliphate in 634. The ability of the Medinan state to absorb random action into a relatively centralized movement of expansion testifies to the strength of the new ideological and administrative patterns inherent in the concept of *ummah.*

Fusion of community and faith

The fusion of two once separable phenomena, membership in Muḥammad's community and faith in Islām—the mundane and the spiritual—would become one of Islām's most distinctive features. Becoming and being Muslim always involved *doing* more than it involved *believing.* On balance, Muslims have always favoured orthopraxy (correctness of practice) over orthodoxy (correctness of doctrine). Being Muslim has always meant making a commitment to a set of behavioral patterns because they reflect the right orientation to God. Where choices were later posed, they were posed not in terms of religion and politics, or church and state, but between living in the world the right way or the wrong way. Just as classical Islāmicate languages developed no equivalents for the words *religion* and *politics,* modern European languages have developed no adequate terms to capture the choices as Muslims have posed them.

Conversion and crystallization (634–870)

SOCIAL AND CULTURAL TRANSFORMATIONS

The Arab conquests are often viewed as a discrete period. The end of the conquests appears to be a convenient dividing line because it coincides with a conventional watershed, the overthrow of the Umayyad caliphs by the ʿAbbāsids. To illustrate their role in broader social and cultural change, however, the military conquests should be included in a period more than twice as long, during which the conquest of the hearts and minds of the majority of the subject population also occurred. Between 634 and 870 Islām was transformed from the badge of a small Arab ruling class to the dominant faith of a vast empire that stretched from the western Mediterranean into Central Asia. As a result of this long and gradual period of conversion, Arab cultures intermingled with the indigenous cultures of the conquered peoples to produce Islām's fundamental orientations and identities. The Arabic language became a vehicle for the transmission of high culture, even though the Arabs remained a minority; for the first time in the history of the Nile-to-Oxus region, a new language of high culture, carrying a great cultural florescence, replaced all previous languages of high culture. Trade and taxation replaced booty as the fiscal basis of the Muslim state; a nontribal army replaced a tribal one; and a centralized empire became a nominal confederation, with all of the social dislocation and rivalries those changes imply. Yet despite continuous internal dissension, virtually no Muslim raised the possibility of there being more than one legitimate leader. Furthermore, the impulse toward solidarity, inherited from Muḥammad and Abū Bakr, may have actually been encouraged by persisting minority status. While Muslims were a minority, they naturally formed a conception of Islāmic dominance as territorial rather than religious; and of unconverted non-Muslim communities as secondary members. In one important respect the Islāmic faith differed from all other major religious traditions: the formative period of the faith coincided with its political domination of a rich complex of old cultures. Thus, during the formative period of their civilization, the Muslims could both introduce new elements and reorient old ones in creative ways.

Arabic as a language of high culture

Just as Muḥammad fulfilled and redirected ongoing tendencies in Arabia, the builders of early Islāmicate civilization carried forth and transformed developments in the Roman and Sāsānian territories in which they first dominated. While Muḥammad was emerging as a leader in the Hejaz, the Byzantine and Sāsānian emperors were ruling states that resembled what the Islāmicate empire was to become. Byzantine rule stretched from North Africa into Syria and sometimes Iraq; the Sāsānians competed with the Byzantines in Syria and Iraq and extended their sway, at its furthest, across the Oxus River. Among their subjects were speakers and writers of several major languages—various forms of Aramaic such as Mandaean and Syriac; Greek; Arabic; and Middle Persian. In fact, a significant number of persons were probably bilingual or trilingual. Each empire had its official religion, Christianity and Zoroastrian–Mazdaism, respectively. The Sāsānian Empire in the early 7th century was ruled by a religion-backed centralized monarchy with an elaborate bureaucratic structure that was reproduced on a smaller scale at the provincial courts of its appointed governors. Its religious demography was complex—Christians of many persuasions, Monophysites, Nestorians, Orthodox, and others; pagans; gnostics; Jews; Mazdeans. Minority religious communities were becoming more clearly organized and isolated. The population included priests; traders and merchants; landlords (*dihqans*), sometimes living not on the land but as absentees in the cities; pastoralists; and large numbers of peasant agriculturalists. In southern Iraq, especially in and around towns like al-Ḥīrah, it included migratory and settled Arabs as well. Both empires relied on standing armies for their defense and on agriculture, taxation, conquest, and trade for their resources. When the Muslim conquests began, the Byzantines and Sāsānians had been in conflict for a century; in the most recent exchanges, the Sāsānians had established direct rule in al-Ḥīrah, further exposing its many Arabs to their administration. When the Arab conquests began, representatives of Byzantine and Sāsānian rule on Arabia's northern borders were not strong enough to resist.

ʿUMAR I'S SUCCESSION

ʿUmar I, successor to Abū Bakr

The spirit of conquest under ʿUmar I. Abū Bakr's successor in Medina, ʿUmar I (ruled 634–644), had not so much to stimulate conquest as to organize and channel it. As leaders he chose skillful managers experienced in trade and commerce as well as warfare and imbued with an ideology that provided their activities with a cosmic significance. The total numbers involved in the initial conquests may have been relatively small, perhaps less than 50,000, divided into numerous shifting groups. Yet few actions took place without any sanction from the Medinan government or one of its appointed commanders. The fighters, or *muqātilah,* could generally accomplish much more with Medina's support than without. ʿUmar, one of Muḥammad's earliest and staunchest supporters, had quickly developed an administrative system of manifestly superior effectiveness. He defined the *ummah* as a continually expansive polity managed by a new ruling elite, which included successful military commanders like Khālid ibn al-Walīd. Even after the conquests ended, this sense of expansiveness continued to be expressed in the way Muslims divided the world into their own zone, the Dār al-Islām, and the zone into which they could and should expand, the Dār al-Ḥarb, the abode of war. The norms of ʿUmar's new elite were supplied by Islām as it was then understood. Taken together, Muḥammad's revelations from God and his *sunnah* (precedent-setting example) defined the cultic and personal practices that distinguished Muslims from others: prayer, fasting, pilgrimage, charity, avoidance of pork and intoxicants, membership in one community centred at Mecca, and activism (*jihād*) in the community's behalf.

Distribution of booty among the *ummah*

Origin of the garrison mosque

Forging the link of activism with faithfulness. ʿUmar symbolized this conception of the *ummah* in two ways. He assumed an additional title, *amīr al-muʾminīn* ("commander of the faithful"), which linked organized activism with faithfulness (*īmān*), the earliest defining feature of the Muslim. He also adopted a lunar calendar that began with the emigration (*hijrah*), the moment at which a group of individual followers of Muḥammad had become an active social presence. Because booty was the *ummah*'s major resource, ʿUmar concentrated on ways to distribute and sustain it. He established a *dīwān,* or register, to pay all members of the ruling elite and the conquering forces, from Muḥammad's family on down, in order of entry into the *ummah.* The immovable booty was kept for the state. After the government's fifth-share of the movable booty was reserved, the rest was distributed according to the *dīwān.* The *muqātilah* he stationed as an occupying army in garrisons (*amṣār*) constructed in locations strategic to further conquest: al-Fusṭāṭ in Egypt, Damascus in Syria, Kūfah and Basra in Iraq. The garrisons attracted indigenous population and initiated significant demographic changes, such as a population shift from northern to southern Iraq. They also inaugurated the rudiments of an "Islāmic" daily life; each garrison was commanded by a caliphal appointee, responsible for setting aside an area for prayer, a mosque (*masjid*), named for the prostrations (*sujūd*) that had become a characteristic element in the five daily worship sessions (*ṣalāt*s). There the fighters could hear God's revelations to Muḥammad recited by men trained in that emerging art. The most pious might commit the whole to memory. There, too, the Friday midday *ṣalāt* could be performed communally, accompanied by an important educational device, the sermon (*khuṭbah*), through which the fighters could be instructed in the principles of the faith. The mosque fused the practical and the spiritual in a special way: because the Friday prayer included an expression of loyalty to the ruler, it could also provide an opportunity to declare rebellion.

The series of ongoing conquests that fueled this system had their most extensive phase under ʿUmar and his successor ʿUthmān ibn ʿAffān (ruled 644–656). Within 25 years, Muslim Arab forces created the first empire permanently to link western Asia with the Mediterranean. Within another century, Muslim conquerors surpassed the achievement of Alexander the Great, not only in the durability of their accomplishment but in its scope as well, reaching from the Iberian Peninsula to Central Asia. Resistance was generally slight and nondestructive, and conquest through capitulation was preferred to conquest by force. After Sāsānian al-Ḥīrah fell in 633, a large Byzantine force was defeated in Syria, opening the way to the final conquest of Damascus in 636. The next year, further gains were made in Sāsānian territory, especially at the Battle of al-Qādisīyah; in the next, the focus returned to Syria and the taking of Jerusalem. By 640, Roman control in Syria was over; by 641, the Sāsānians had lost all of their territory west of Zagros. During the years 642 to 646 Egypt was taken under the leadership of ʿAmr ibn al-ʿĀṣ, who soon began raids into what the Muslims called the Maghrib, the lands west of Egypt. Shortly thereafter, in the east, Persepolis fell; in 651 the defeat and assassination of the last Sāsānian emperor, Yazdegerd III, marked the end of the 400-year-old Sāsānian Empire.

ʿUTHMĀN'S SUCCESSION AND POLICIES

ʿUthmān's standardization of the Qurʾān

Discontent in ʿUthmān's reign. This phase of conquest ended under ʿUthmān and ramified widely. ʿUthmān may even have sent an emissary to China in 651; by the end of the 7th century Arab Muslims were trading there. The fiscal strain of such expansion and the growing independence of local Arabs outside the peninsula underlay the persisting discontents that surfaced toward the end of ʿUthmān's reign. The very way in which he was made caliph had already signaled the potential for competition over leadership and resources. Perceived as pliable and docile, he was the choice of the small committee charged by the dying ʿUmar with selecting one of their own number. Once in office, however, ʿUthmān acted to establish the power of Medina over and against some of the powerful Quraysh families at Mecca and local notables outside Arabia. He was accused of nepotism for relying on his own family, the Banū Umayyah, whose talents ʿUmar had already recognized. Among his many other "objectionable" acts was his call for the production of a single standard collection of Muḥammad's messages from God, which was known simply as the Qurʾān ("Recitation" or "Recitations"). Simultaneously he ordered the destruction of any other collections. Although they might have differed only in minor respects, they represented the independence of local communities. Above all, ʿUthmān was the natural target of anyone dissatisfied with the distribution of the conquest's wealth, since he represented and defended a system that defined all income as Medina's to distribute.

The difficulties of ʿUthmān's reign took more than a century to resolve. They were the inevitable result not just of the actions of individuals but of the whole process initiated by Muḥammad's achievements. His coalition had been fragile. He had disturbed existing social arrangements without being able to reconstruct and stabilize new ones quickly. Into a society organized along family lines, he had introduced the supremacy of trans-kinship ties. Yet he had been forced to make use of kinship ties himself; and, despite his egalitarian message, he had introduced new inequities by granting privileges to the earliest and most intensely devoted followers of his cause. Furthermore, personal rivalries were stimulated by his charisma; individuals like his wife ʿĀʾishah, his daughter Fāṭimah, and her husband ʿAlī frequently vied for his affection. ʿUmar's *dīwān* had, then, reinforced old inequities by extending privileges to wealthy high-placed Meccans, and it had introduced new tensions by assigning a lower status to those, indigenous or immigrant to the provinces, who joined the cause later (but who felt themselves to be making an equivalent or greater contribution). Other tensions resulted from conditions in the conquered lands: the initial isolation of Arab Muslims, and even Arab Christians who fought with them, from the indigenous non-Arab population; the discouragement of non-Arab converts, except as clients (*mawālī*) of Arab tribes; the administrative dependence of peninsular Arabs on local Arabs and non-Arabs; and the development of a tax system that discriminated against non-Muslims.

Intra-Muslim conflicts. The ensuing conflicts were played out in a series of intra-Muslim disputes that began with ʿUthmān's assassination and continued to the end of the period under discussion. The importance of kinship ties persisted, but they were gradually replaced by the identities of a new social order. These new identities resulted from Muslim responses to anti-Muslim activity as well as from Muslim participation in a series of controversies focused on the issue of leadership. Because the *ummah,* unified under one leader, was seen as an earthly expression of God's favour, and because God was seen as the controller of all aspects of human existence, the identities formed in the course of the *ummah*'s early history could fuse dimensions that secular modern observers are able to distinguish—religious, social, political, and economic. Furthermore, intra-Muslim rivalries changed during the conversion period; the meaningfulness of the new iden-

Unifying of the Muslim tradition

tities expanded as non-Muslims contributed to Islām's formation, through opposition or through conversion, and the key issues broadened as the participating constituencies enlarged. At first the disputes were coterminous with intra-Arab, indeed even intra-Quraysh, rivalries; only later did they involve persons of other backgrounds. Thus the faith of Islām was formed in conjunction with the crises that attended the establishment of rule by Muslims. Muslims might have produced an extremely localized and exclusivistic religion; but in spite of, and perhaps because of, their willingness to engage in continuing internal conflicts, they produced one of the most unified religious traditions in human history.

THE FOUR FITNAHS

By the end of the period of conversion and crystallization, Muslim historians would retrospectively identify four discrete periods of conflict and label them *fitnah*s, trials or temptations to test the unity of the *ummah.* Many historians also came to view some identities formed during the *fitnah*s as authentic and others as deviant. This retrospective interpretation may be anachronistic and misleading. The entire period between 656 and the last quarter of the 9th century was conflict-ridden, and the *fitnah*s merely mark periods of intensification; yet the most striking characteristic of the period was the pursuit of unity.

The first fitnah. In the first two *fitnah*s the claimants to the caliphate relied on their high standing among the Quraysh and their local support in either Arabia, Iraq, or Syria. Competition for the caliphate thus reflected rivalries among the leading Arab families as well as regional interests. The first *fitnah* occurred between ʿUthmān's assassination in 656 and the accession of his kinsman Muʿāwiyah I in 661 and included the caliphate of ʿAlī, the cousin and son-in-law of Muḥammad. It involved a three-way contest between ʿAlī's party in Iraq; a coalition of important Quraysh families in Mecca, including Muḥammad's wife ʿĀʾishah and Ṭalḥah and Zubayr; and the party of Muʿāwiyah, the governor of Syria and member of ʿUthmān's clan, the Banū Umayyah. Ostensibly the conflict focused on whether ʿUthmān had been assassinated justly, whether ʿAlī had been involved, and whether ʿUthmān's death should be avenged by Muʿāwiyah or by the leading Meccans. ʿAlī and his party (*shīʿah*) at first gained power over the representatives of the other leading Meccan families, then lost it permanently to Muʿāwiyah, who elevated Damascus, which had been his provincial capital, to the status of imperial capital. Disappointed at the Battle of Ṣiffīn (657) with ʿAlī's failure to insist on his right to rule, a segment of his partisans withdrew, calling themselves accordingly Khawārij (Kharijites; "seceders"). Their spiritual heirs would come to recognize any pious Muslim as leader. Meanwhile, another segment of ʿAlī's party intensified their loyalty to him as a just and heroic leader who was one of Muḥammad's dearest intimates and the father of his only male descendants.

The second fitnah. The second *fitnah* followed Muʿāwiyah's caliphate (661–680), which itself was not free from strife, and coincided with the caliphates of Muʿāwiyah's son Yazīd I (ruled 680–683), whom he designated as successor, and Yazīd's three successors. This *fitnah* was a second-generation reprise of the first; some of the personnel of the former were descendants or relatives of the leaders of the latter. Once again, different regions supported different claimants, as new tribal divisions emerged in the garrison towns; and once again, representatives of the Syrian Umayyads prevailed. In 680, at Karbalāʾ in Iraq, Yazīd's army murdered al-Ḥusayn, a son of ʿAlī and grandson of Muḥammad, along with a small group of supporters, accusing them of rebellion; and even though the Umayyads subdued Iraq, rebellions in the name of this or that relative of ʿAlī continued, attracting more and more non-Arab support and introducing new dimensions to his cause. In the Hejaz, the Marwānid branch of the Umayyads, descendants of Marwān I who claimed the caliphate in 685, fought against ʿAbd Allāh ibn az-Zubayr for years; by the time they defeated him, they had lost most of Arabia to Kharijite rebels.

Resistance to Muslim authority

During the period of the first two *fitnah*s, resistance to Muslim rule was an added source of conflict. Some of this resistance took the form of syncretic or anti-Islāmic religious movements. For example, during the second *fitnah,* in Iraq a Jew named Abū ʿIsā al-Iṣfahānī led a syncretic movement (that is, a movement combining different forms of belief or practice) on the basis of his claim to be a prophet (an option not generally open to Muslim rebels) and forerunner of the messiah. He viewed Muḥammad, as well as Jesus, as messengers sent not to all humanity but only to their own communities; so he urged each community to continue in its own tradition as he helped prepare for the coming of the messiah. In other areas, such as the newly conquered Maghrib, resistance took the form of large-scale military hostility. In the 660s the Umayyads had expanded their conflict with the Byzantine Empire by competing for bases in coastal North Africa; it soon became clear, however, that only a full-fledged occupation would serve their purposes. That occupation was begun by ʿUqbah ibn Nāfiʿ, the founder of al-Qayrawān (Kairouan, in modern Tunisia) and, as Sidi (Saint) ʿUqbah, the first of many Maghribi Muslim saints. It eventually resulted in the incorporation of large numbers of pagan or Christianized Berber tribes, the first large-scale forcible incorporation of tribal peoples since the secession of tribes under Abū Bakr. But first the Arab armies met fierce resistance from two individuals—one a man, Kusaylah, and one a woman, al-Kāhinah—who became Berber heroes. Berber resistance was not controlled until the end of the 7th century, after which the Berbers participated in the further conquest of the Maghrib and the Iberian peninsula.

During the caliphate of ʿAbd al-Malik ibn Marwān (ruled 685–705), which followed the end of the second *fitnah,* and under his successors during the next four decades, the problematic consequences of the conquests became much more visible. Like their Byzantine and late Sāsānian predecessors, the Marwānid caliphs nominally ruled the various religious communities but allowed the communities' own appointed or elected officials to administer most internal affairs. Yet now the right of religious communities to live in this fashion was justified by the Qurʾān and *sunnah;* as peoples with revealed books (*ahl al-kitāb*), they deserved protection (*dhimmah*) in return for a payment. The Arabs also formed a single religious community whose right to rule over the non-Arab protected communities the Marwānids sought to maintain.

The Dome of the Rock

To signify this supremacy, as well as his co-optation of previous legitimacy, ʿAbd al-Malik ordered the construction of the Dome of the Rock, a monumental mosque, in Jerusalem, a major centre of non-Muslim population. The site chosen was sacred to Jews and Christians because of its associations with biblical history; it held added meaning for Muslims, who believed it to be the starting point for Muḥammad's *miʿrāj* (midnight journey to heaven). Although this and other early mosques resembled contemporary Christian churches, gradually an Islāmic aesthetic emerged: a dome on a geometrical base, accompanied by a minaret from which to deliver the call to prayer; and an emphasis on surface decoration that combined arabesque and geometrical design with calligraphic representations of God's Word. ʿAbd al-Malik took other steps to mark the distinctiveness of Islāmic rule: for example, he encouraged the use of Arabic as the language of government and had Islāmized coins minted to replace the Byzantine and Sāsānian-style coinage that had continued to be used since the conquests. During the Marwānid period, the Muslim community was further consolidated by the regularization of the public cult and the crystallization of a set of five minimal duties (sometimes called pillars).

Yet the Marwānids also depended heavily on the help of non-Arab administrative personnel (*kuttāb;* singular, *kātib*) and on administrative practices (*e.g.,* a set of government bureaus) inherited from Byzantine and, in particular, late Sāsānian practice. Pre-Islāmic writings on governance translated into Arabic, especially from Middle Persian, influenced caliphal style. The governing structure at Damascus and in the provinces began to resemble pre-Islāmic monarchy, and thus appealed to a majority of subjects, whose heritage extolled the absolute authority of a divinely sanctioned ruler. Much of the inspiration for this development came from ʿAbd al-Malik's administrator in the eastern territories, al-Ḥajjāj ibn Yūsuf ath-Thaqafī, who was himself an admirer of Sāsānian practice.

The Marwānid caliphs, as rulers of Muslims and non-Muslims alike, had thus been forced to respond to a variety of expectations. Ironically, it was their defense of the importance and distinctiveness of the Arabic language and the Islāmic community, not their responsiveness to non-Muslim preferences, that prepared the way for the gradual incorporation of most of the subject population into the *ummah.* As the conquests slowed and the isolation of the fighters (*muqātilah*) became less necessary, it became more and more difficult to keep Arabs garrisoned. The sedentation of Arabs that had begun in the Hejaz was being repeated and extended outside the peninsula. As the tribal links that had so dominated Umayyad politics began to break down, the meaningfulness of tying non-Arab converts to Arab tribes as clients was diluted; moreover, the number of non-Muslims who wished to join the *ummah* was already becoming too large for this process to work effectively.

Simultaneously, the growing prestige and elaboration of things Arabic and Islāmic made them more attractive, to non-Arab Muslims and to non-Muslims alike. The more the Muslim rulers succeeded, the more prestige their customs, norms, and habits acquired. Heirs to the considerable agricultural and commercial resources of the Nile-to-Oxus region, they increased its prosperity and widened its horizons by extending its control far to the east and west.

Arabic as a lingua franca

Arabic, which occasionally had been used for administrative purposes in earlier empires, now became a valuable lingua franca. As Muslims continued to adapt to rapidly changing circumstances, they needed Arabic to reflect upon and elaborate what they had inherited from the Hejaz. Because the Qurʾān, translation of which was prohibited, was written in a form of Arabic that quickly became archaic to Muslims living in the garrisons, and because it contained references to life in Arabia before and during Muḥammad's time, full understanding of the text required special effort. Scholars began to study the religion and poetry of the *jāhilīyah,* the times of ignorance before God's revelation to Muḥammad. Philologians soon emerged, in the Hejaz as well as in the garrisons. Many Muslims cultivated reports, which came to be known as *ḥadīth,* of what Muḥammad had said and done, in order to develop a clearer and fuller picture of his *sunnah.* These materials were sometimes gathered into accounts of his campaigns, called *maghāzī.* The emulation of Muḥammad's *sunnah* was a major factor in the development of recognizably "Muslim" styles of personal piety and public decision making. As differences in the garrisons needed to be settled according to "Islāmic" principles, the caliphs appointed arbitrating judges, *qāḍī*s, who were knowledgeable in Qurʾān and *sunnah.* The pursuit of legal knowledge, *fiqh,* was taken up in many locales and informed by local pre-Islāmic custom and Islāmic resources. These special forms of knowledge began to be known as *ʿulūm* (singular, *ʿilm*); the persons who pursued them, as *ʿulamāʾ* (singular, *ʿālim*), a role that provided new sources of prestige and influence, especially for recent converts or sons of converts.

Muslims outside Arabia were also affected by interacting with members of the religious communities over which they ruled. When protected non-Muslims converted, they brought new expectations and habits with them; Islāmic eschatology is one area that reflects such enrichment. Unconverted protected groups (*dhimmī*s) were equally influential. Expressions of Islāmic identity often had to take into account the critique of non-Muslims, just as the various non-Muslim traditions were affected by contact with Muslims.

Interaction with non-Muslims

This interaction had special consequences in the areas of prophethood and revelation, where major shifts and accommodations occurred among Jews, Christians, Mazdeans, and Muslims during the first two centuries of their coexistence. Muslims attempted to establish Muḥammad's legitimacy as an heir to Jewish and Christian prophethood, while non-Muslims tried to distinguish their prophets and scriptures from Muḥammad and the Qurʾān. Within the emergent Islāmicate civilization, the separate religious communities continued to go their own way; but the influence of Muslim rule and the intervention of the caliphs in their internal affairs could not help but affect them. The Babylonian Talmud, completed during these years, bears traces of early interaction among communities. In Iraq caliphal policy helped promote the Jewish gaons (local rabbinic authorities) over the exilarch (a central secular leader). Mazdeans turned to the Nestorian Church to avoid Islām, or reconceptualized Zoroaster as a prophet sent to a community with a Book. With the *dhimmī* system (the system of protecting non-Muslims for payment), Muslim rulers formalized and probably intensified pre-Islāmic tendencies toward religious communalization. Furthermore, the greater formality of the new system could protect the subject communities from each other as well as from the dominant minority. So "converting" to Islām, at least in the Nile-to-Oxus region, meant joining one recognizably distinct social entity and leaving another. One of the most significant aspects of many Muslim societies was the inseparability of "religious" affiliation and group membership, a phenomenon that has translated poorly into the social structures of modern Muslim nations. In the central caliphal lands of the early 8th century, membership in the Muslim community offered the best chance for social and physical mobility, regardless of a certain degree of discrimination against non-Arabs. Among many astounding examples of this mobility is the fact that several of the early governors and independent dynasts of Egypt and the Maghrib were grandsons of men born in Central Asia.

The Marwānid Maghrib illustrates a kind of conversion more like that of the peninsular Arabs. After the defeat of initial Berber resistance movements, the Arab conquerors of the Maghrib quickly incorporated the Berber tribes en masse into the Muslim community, turning them immediately to further conquests. In 710 an Arab–Berber army set out for the Iberian Peninsula under the leadership of Ṭāriq ibn Ziyād (the name Gibraltar is derived from Jabal Ṭāriq, or "Mountain of Ṭāriq"). They defeated King Roderick in 711; raided into and through the Iberian Peninsula, which they called al-Andalūs; and ruled in the name of the Umayyad caliph. The Andalusian Muslims never had serious goals across the Pyrenees. In 732 Charles Martel encountered not a Muslim army but a summer raiding party; despite his "victory" over that party, Muslims continued their seasonal raiding along the southern French coast for many years.

Muslim Andalusia

Muslim Andalusia is particularly interesting because there the pressure for large-scale conversion that was coming to plague the Umayyads in Syria, Iraq, and Iran never developed. Muslims may never have become a majority throughout their 700-year Andalusian presence. Non-Muslims entered into the Muslim realm as Mozarabs, Christians who had adopted the language and manners, rather than the faith, of the Arabs. Given essentially the same administrative arrangements, the Iberian Christian population was later restored to dominance, while the Syrian Christian population was drastically reduced; but the Iberian Jewish population all but disappeared while the Nile-to-Oxus Jewish population survived.

The Berbers who remained in the Maghrib illustrate the mobility of ideologies and institutions from the central lands to more recently conquered territories. No sooner had they given up anti-Muslim resistance and joined the Muslim community than they rebelled again; but this time an Islāmic identity, Kharijism, provided the justification. Kharijite ideas had been carried to the Maghrib by refugees from the numerous revolts against the Marwānids. Kharijite egalitarianism suited the economic and social grievances of the Berbers as non-Arab Muslims under Arab rule. The revolts outlasted the Marwānids; they resulted in the first independent Maghribi dynasty, the Rustamid, founded by Muslims of Persian descent. The direct influence of the revolts was felt as late as the 10th century and survives among small communities in Tunisia and Algeria.

The third fitnah. Meanwhile, in the central caliphal lands, growing discontent with the emerging order crystallized in a multifaceted movement of opposition to the Marwānids. It culminated in the third *fitnah* (744–750), which resulted in the establishment of a new and final dynasty of caliphs, the ʿAbbāsids. Ever since the second *fitnah,* a number of concerned and self-conscious Muslims had begun to raise serious questions about the proper Muslim life and the Marwānids' ability to exemplify it, and to answer them by reference to key events in the *ummah*'s history. Pious Muslims tried to define a good Muslim and to decide whether a bad Muslim should be excluded from the community, or a bad caliph from office. They also considered God's role in determining a person's sinfulness and final dispensation. The proper relationship between Arab and non-Arab Muslims, and between Muslims and *dhimmī*s, was another important and predictable focus of reflection. The willingness of non-Arabs to join the *ummah* was growing; but the Marwānids had not found a solution that was either ideologically acceptable or fiscally sound. Because protected non-Muslim groups paid special taxes, fiscal stability seemed to depend on continuing to discourage conversion. One Marwānid, ʿUmar II (ruled 717–720), experimented unsuccessfully with a just solution. In these very practical and often pressing debates lay the germs of Muslim theology, as various overlapping positions, not always coterminous with political groupings, were taken: rejecting the history of the community by demanding rule by Muḥammad's family; rejecting the history of the community by following any pious Muslim and excluding any sinner; or accepting the history of the community, its leaders, and most of its members.

In the course of these debates the Marwānid caliphs began to seem severely deficient to a significant number of Muslims of differing persuasions and aspirations. Direct and implied criticism began to surface. Al-Ḥasan al-Baṣrī, a pious ascetic and a model for the early *ṣūfī*s, called on the Marwānids to rule as good Muslims, and on good Muslims to be suspicious of worldly power. Ibn Isḥāq composed an account of Muḥammad's messengership that emphasized the importance of the *anṣār,* the Yathribi tribes that accepted Muḥammad, and by implication the non-Arab converts (from whom Ibn Isḥāq himself was descended). The Marwānids were accused of *bidʿah,* new actions for which there were no legitimate Islāmic precedents. Their continuation of pre-Islāmic institutions—the spy system, extortion of deposed officials by torture, and summary execution—were some of their most visible "offenses." To the pious, the ideal ruler, or *imām* (the word also for a Muslim who led the *ṣalāt*), should, like Muḥammad, possess special learning and knowledge. The first four caliphs, they argued, had been *imām*s in this sense; but under the Umayyads the caliphate had been reduced to a military and administrative office devoid of *imāmah,* of true legitimacy. This piety-minded opposition to the Umayyads, as it has been aptly dubbed, now began to talk about a new dispensation. Some of the most vocal members found special learning and knowledge only in Muḥammad's family. Some defined Muḥammad's family broadly to include any Hāshimite; others, more narrowly, to include only descendants of ʿAlī. As the number of Muḥammad's descendants through ʿAlī had grown, numerous rebellions had broken out in the name of one or the other, drawing on various combinations of constituencies and reflecting a wide spectrum of Islāmic and pre-Islāmic aspirations.

Criticism of the Marwānid caliphs

In the late Marwānid period, the piety-minded opposition found expression in a movement organized in Khorāsān (Khurasan) by Abū Muslim, a semisecret operative of one particularly ambitious Hāshimite family, the ʿAbbāsids. The ʿAbbāsids, who were kin but not descendants of Muḥammad, claimed also to have inherited, a generation earlier, the authority of one of ʿAlī's actual descendants, Abū Hāshim. Publicly Abū Muslim called for any qualified member of Muḥammad's family to become caliph; but privately he allowed the partisans (*shīʿah*) of ʿAlī to assume that he meant them. Abū Muslim ultimately succeeded because he managed to link the concerns of the piety-minded in Syria and Iraq with Khorāsānian discontent. He played upon the grievances of its Arab tribes against the tribes of Syria and their representatives in the Khorāsānian provincial government, and on the millennial expectations of non-Arab converts and non-Muslims disenchanted with the injustices of Marwānid rule.

When in 750 the army organized and led by Abū Muslim succeeded in defeating the last Marwānid ruler, his caliph-designate represented only one segment of this broad coalition. He was the head of the ʿAbbāsid family, Abū al-ʿAbbās as-Saffāḥ, who now subordinated the claims of the party of ʿAlī to those of his own family, and who promised to restore the unity of the *ummah,* or *jamāʿah.* The circumstances of his accession reconfigured the piety-minded opposition that had helped bring him to power. The party (*shīʿah*) of ʿAlī refused to accept the compromise the ʿAbbāsids offered. Their former fellow-opponents did accept membership in the reunified *jamāʿah,* isolating the People of the Shīʿah and causing them to define themselves in terms of more radical points of view. Those who accepted the early ʿAbbāsids came to be known as the People of the Sunnah and Jamāʿah. They accepted the cumulative historical reality of the *ummah*'s first century: all of the decisions of the community, and all of the caliphs it had accepted, had been legitimate, as would be any subsequent caliph who could unite the community. The concept of *fitnah* acquired a fully historicist meaning: if internal discord were a trial sent by God, then any unifying victor must be God's choice.

Sunnites and Shīʿites. The historicists came to be known as Sunnites, their main opponents, as Shīʿites. These labels are somewhat misleading, because they imply that only the Sunnites tried to follow the *sunnah* of Muḥammad. In fact, each group relied on the *sunnah,* but emphasized different elements. For the Sunnites, who should more properly be called the Jamāʿi-Sunnites, the principle of solidarity was essential to the *sunnah.* The Shīʿites argued that the fundamental element of the *sunnah,* and one willfully overlooked by the Jamāʿi-Sunnites, was Muḥammad's devotion to his family and his wish that they succeed him through ʿAlī. These new labels expressed and consolidated the social reorganization that had been under way since the beginning of the conquests. The vast majority of Muslims now became consensus-oriented, while a small minority became oppositional. The inherent inimitability of Muḥammad's role had made it impossible for any form of successorship to capture universal approval.

When the ʿAbbāsids denied the special claims of the family of ʿAlī, they prompted the Shīʿites to define themselves as a permanent opposition to the status quo. The crystallization of Shīʿism into a movement of protest received its greatest impetus during and just after the lifetime of one of the most influential Shīʿite leaders of the early ʿAbbāsid period, Jaʿfar ibn Muḥammad (also called Jaʿfar aṣ-Ṣādiq; 765). Jaʿfar's vision and leadership allowed the Shīʿites to understand their chaotic history as a meaningful series of efforts by truly pious and suffering Muslims to right the wrongs of the majority. The leaders of the minority had occupied the office of *imām,* the central Shīʿite institution, which had been passed on from the first *imām,* ʿAlī, by designation down to Jaʿfar, the sixth. To protect his followers from increasing Sunnite hostility to the views of radical Shīʿites, known as the *ghulāt* ("extremists"), who claimed prophethood for ʿAlī, Jaʿfar made a distinction that both protected the uniqueness of prophethood and established the superiority of the role of *imām.* Since prophethood had ended, its true intent would die without the *imām*s, whose protection from error allowed them to carry out their indispensable task.

Crystallization of Shīʿite movement

Although Jaʿfar did develop an ideology that invited Sunnite toleration, he did not unify all Shīʿites. Differences continued to be expressed through loyalty to various of his relatives. During Jaʿfar's lifetime, his uncle Zayd revolted in Kūfah (740), founding the branch of the Shīʿism known as the *zaydīyah* (Zaydis), or Fivers (for their allegiance to the fifth *imām*), who became particularly important in southern Arabia. Any pious follower of ʿAlī could become their *imām,* and any *imām* could be deposed if he behaved unacceptably. The Shīʿite majority followed Jaʿfar's son Mūsā al-Kāẓim and *imām*s in his line through the 12th, who disappeared in 873. Those loyal to the 12 *imām*s became known as the Imāmīs or Ithnā ʿAsharīyah (Twelvers). They adopted a quietistic stance toward the status quo government of the ʿAbbāsids and prepared to wait until the 12th *imām* should return as the messiah

Branches of Shīʿism

to avenge injustices against Shīʿites and to restore justice before the Last Judgment. Some of Jaʿfar's followers, however, remained loyal to Ismāʿīl, Jaʿfar's eldest son who predeceased his father after being designated. These became the Ismāʿīlīyah (Ismāʿīlis) or Sabʿīyah (Seveners), and they soon became a source of continuing revolution in the name of Ismāʿīl's son Muḥammad at-Tamm, who was believed to have disappeared. Challenges to the ʿAbbāsids were not long in coming; of particular significance was the establishment, in 789, of the first independent Shīʿite dynasty, in present-day Morocco, by Idrīs ibn ʿAbd Allāh ibn Ḥasan II, who had fled after participating in an unsuccessful uprising near Mecca. Furthermore, Kharijite rebellions continued to occur regularly.

Legitimacy was a scarce and fragile resource in all premodern societies; in the early ʿAbbāsid environment, competition to define and secure legitimacy was especially intense. The ʿAbbāsids came to power vulnerable; their early actions undermined the unitive potential of their office. Having alienated the Shīʿites, they liquidated the Umayyad family, one of whom, ʿAbd ar-Raḥmān I, escaped and founded his own state in Andalusia. Although the ʿAbbāsids were able to buttress their legitimacy by employing the force of their Khorāsānian army, by appealing to their piety-minded support, and by emphasizing their position as heirs to the pre-Islāmic traditions of rulership, their own circumstances and policies militated against them. Despite their continuing preference for Khorāsānian troops, the ʿAbbāsids' move to Iraq and their execution of Abū Muslim disappointed the Khorāsānian chauvinists who had helped them. The non-Muslim majority often rebelled, too. Bih'āfrīd ibn Farwardīn claimed to be a prophet capable of incorporating both Mazdaism and Islām into a new faith. Hāshim ibn Ḥākim, called al-Muqannaʿ (the "Veiled One"), around 759 declared himself a prophet and then a god, heir to all previous prophets, to numerous followers of ʿAlī, and to Abū Muslim himself.

The ʿAbbāsid court at Baghdad

The ʿAbbāsids symbolized their connection with their pre-Islāmic predecessors by founding a new capital, Baghdad, near the old Sāsānian capital. They also continued to elaborate the Sāsānian-like structure begun by the Marwānid governors in Iraq. Their court life became more and more elaborate, the bureaucracy fuller, the inner sanctum of the palace fuller than ever with slaves and concubines as well as the retinues of the caliph's four legal wives. By the time of Hārūn ar-Rashīd (ruled 786–809), Europe had nothing to compare with Baghdad, not even the court of his contemporary Charlemagne (742–814). But problems surfaced, too. Slaves' sons fathered by Muslims were not slaves and so could compete for the succession. Despite the ʿAbbāsids' defense of Islām, unconverted Jews and Christians could be influential at court. The head (vizier or *wazīr*) of the financial bureaucracy sometimes became the effective head of government by taking over the chancery as well. Like all absolute rulers, the ʿAbbāsid caliphs soon confronted the insoluble dilemma of absolutism: the monarch cannot be absolute unless he depends on helpers, but his dependence on helpers undermines his absolutism. Hārūn ar-Rashīd experienced this paradox in a particularly painful way: having drawn into his service prominent members of a family of Buddhist converts, the Barmakids, he found them such rivals that he liquidated them within a matter of years. It was also during Hārūn's reign that Ibrāhīm ibn al-Aghlab, a trusted governor in Tunis, founded a dynasty that gradually became independent, as did the Ṭāhirids, the ʿAbbāsid governors in Khorāsān, two decades later.

The ʿAbbāsids' ability to rival their pre-Islāmic predecessors was enhanced by their generous patronage of artists and artisans of all kinds. The great 7,000-mile Silk Road from Ch'ang-an (now Sian, China) to Baghdad (then the two largest cities in the world) helped provide the wealth. The ensuing literary florescence was promoted by the capture of a group of Chinese papermakers at the Battle of Talas in 751. The ʿAbbāsids encouraged translation from pre-Islāmic languages, particularly Middle Persian, Greek, and Syriac. This activity provided a channel through which older thought could enter and be reoriented by Islāmicate societies. In the field of mathematics, al-Khwārizmī, from whose name the word *algorithm* is derived, creatively combined Hellenistic and Sanskritic concepts. The word *algebra* derives from the title of his major work, *Kitab al-jabr wa al-muqābalah* ("The Book of Integration and Equation"). Movements such as *falsafah* (a combination of the positive sciences with logic and metaphysics) and *kalām* (systematic theological discourse) applied Hellenistic thought to new questions. The translation of Indo-Persian lore promoted the development of *adab,* a name for a sophisticated prose literature as well as the set of refined urbane manners that characterized its clientele. Soon a movement called *shuʿūbīyah* arose to champion the superiority of non-Arabic tastes over the alleged crudeness of the poetry so dear to Arabic litterateurs. However, the great writer of early ʿAbbāsid times, al-Jāḥiẓ, produced a type of *adab* that fused pre-Islāmic and Islāmic concerns in excellent Arabic style. Many of these extra-Islāmic resources conflicted with Islāmic expectations. Ibn al-Muqaffaʿ, an administrator under al-Manṣūr (ruled 754–775), urged his master to emulate pre-Islāmic models, lest the law that the religious specialists (the *ʿulamāʾ*) were developing undermine caliphal authority irrevocably.

The ʿAbbāsids never acted on such advice completely; they even contravened it by appealing for piety-minded support. Having encouraged conversion, they tried to "purify" the Muslim community of what they perceived to be socially dangerous and alien ideas. Al-Mahdī (ruled 775–785) actively persecuted the Manichaeans, whom he defined as heretics so as to deny them status as a protected community. He also tried to identify Manichaeans who had joined the Muslim community without abandoning their previous ideas and practices. ʿAbbāsid "purification of Islām" ironically coincided with some of the most significant absorption of pre-Islāmic monotheistic lore to date, as illustrated by the stories of the prophets written by Al-Kisa'i, grammarian and tutor to a royal prince. Even though, like the Marwānids, the ʿAbbāsids continued to maintain administrative courts, not accessible to the *qāḍī*s, they also promoted the study of *ʿilm* and the status of those who pursued it. In so doing they fostered what Ibn al-Muqaffa' had feared—the emergence of an independent body of law, Sharīʿah, which Muslims could use to evaluate and circumvent caliphal rule itself.

Sharīʿah

A key figure in the development of Sharīʿah was Abū ʿAbd Allāh ash-Shāfiʿī, who died in 820. By his time Islāmic law was extensive but uncoordinated, reflecting differing local needs and tastes. Schools had begun to form around various recognized masters, such as al-Awzāʿī in Syria, Abū Ḥanīfah in Iraq, and Mālik ibn Anas, all of whom used some combination of local custom, personal reasoning, Qur'ān, and Ḥadīth. Ash-Shāfiʿī was born in Mecca, studied with Mālik, participated in a Shīʿite revolt in the Yemen, and was sent to Baghdad as a prisoner of the caliph. After his release he emigrated to Egypt, where he produced his most famous work. Like most other *faqih*s (students of jurisprudence, or *fiqh*), ash-Shāfiʿī viewed Muḥammad's community as a social ideal and his first four successors as rightly guided. So that this exemplary time could provide the basis for Islāmic law, he constructed a hierarchy of legal sources: Qur'ān; Ḥadīth, clearly traceable to Muḥammad and in some cases to his companions; *ijmāʿ* (consensus); and *qiyās* (analogy to one of the first three).

The way in which Islāmic law had developed had allowed many pre-Islāmic customs, such as the veiling and seclusion of women, to receive a sanction not given to them in the Qur'ān or the Ḥadīth. Ash-Shāfiʿī did not change that entirely. Law continued to be pursued in different centres, and several major "ways" (*madhhab*s) began to coalesce among Sunnites and Shīʿites alike. Among Sunnites, four schools came to be preeminent, Shāfiʿīyah (Shafiites), Mālikīyah (Malikites), Ḥanafīyah (Hanafites), and Ḥanābilah (Hanbalites), and each individual Muslim was expected to restrict himself to only one. Furthermore, the notion that the gate of *ijtihād* (personal effort at reasoning) closed in the 9th century was not firmly established until the 12th century. However, ash-Shāfiʿī's system was widely influential in controlling divergence and in limiting undisciplined forms of personal reasoning. It also stimulated the collecting and testing of *ḥadīth* for their unbroken traceability to Muḥammad or a companion. The need to verify *ḥadīth* stimulated a characteristic form of premodern Muslim intellectual and literary activity, the collecting of biographical materials into compendiums (*ṭabaqāt*). By viewing the Qur'ān and documentable *sunnah* as preeminent, ash-Shāfiʿī also undermined those in ʿAbbāsid court circles who wanted a more flexible base from which the caliph could operate. The Sharīʿah came to be a supremely authoritative, comprehensive set of norms and rules covering every aspect of life, from worship to personal hygiene. It applied equally to all Muslims, including the ruler, whom Sharīʿah-minded Muslims came to view as its protector, not its administrator or developer. While the caliphs were toying with theocratic notions of themselves as the shadow of God on earth, the students of legal knowledge were defining their rule as "nomocratic," based only on the law they protected and enforced.

According to the Sharīʿah, a Muslim order was one in which the ruler was Muslim and the Sharīʿah was enshrined as a potential guide to all; Muslims were one confessional community among many, each of which would have its own laws that would apply except in disputes between members of different communities. The Sharīʿah regulated relations and inequities among different segments of society, freeborn Muslim, slave, and protected non-Muslim. The process that produced Sharīʿah resembled the evolution of Oral Torah and rabbinic law, which the Sharīʿah resembled in its comprehensiveness, egalitarianism, and consensualism, in its absorption of local custom, in its resistance to distinguishing the sublime from the mundane, and in its independence from government. Like many Jews, many ultra-pious Muslims came to view the law as a divine rather than human creation.

Debate on the nature of the Qur'ān

The fourth fitnah. During the reign of al-Ma'mūn (813–833) the implications of all this *ʿilm*-based activity for caliphal authority began to become clear. Al-Ma'mūn came to the caliphate as the result of the fourth *fitnah*, which reflected the persisting alienation of Khorāsān. Al-Ma'mūn's father, Hārūn ar-Rashīd, provided for the empire to be divided at his death between two sons. Al-Amīn would rule in the capital and all the western domains; al-Ma'mūn, from his provincial seat at Merv in Khorāsān, would rule the less significant east. When Hārūn died, his sons struggled to expand their control; al-Ma'mūn won. During his reign, which probably represents the high point of caliphal absolutism, the court intervened in an unprecedented manner in the intellectual life of its Muslim subjects, who for the next generation engaged in the first major intra-Muslim conflict that focused on belief as well as practice. The Muslims, who now constituted a much more sizable proportion of the population but whose faith lacked doctrinal clarity, began to engage in an argument reminiscent of 2nd-century Christian discussions of the *logos*. Among Christians, for whom the Word was Jesus, the argument had taken a Christological form. But for Muslims the argument had to centre on the Qur'ān and its created or uncreated nature. Al-Ma'mūn, as well as his brother and successor al-Muʿtaṣim, was attracted to the Muʿtazilah (Mutazilites), whose school had been influenced by Hellenistic ideas as well as by contact with non-Muslim theologians. If the Qur'ān were eternal along with God, his unity would, for the Muʿtazilah, be violated. They especially sought to avoid literal exegesis of the Qur'ān, which in their view discouraged free will and produced embarrassing inconsistencies and anthropomorphisms. By arguing that the Qur'ān was created in time, they could justify metaphorical and changing interpretation. By implication, Muḥammad's position as deliverer of revelation was undermined because *ḥadīth* was made less authoritative.

The opponents of the Muʿtazilah, and therefore of the official position, coalesced around the figure of Aḥmad ibn Ḥanbal. A leading master of *ḥadīth*, he had many followers, some of them recent converts, whom he was able to mobilize in large public demonstrations against the doctrine of the created Qur'ān. Because viewing the Qur'ān as created would invalidate its absolute authority, Ibn Ḥanbal argued for an eternal Qur'ān and emphasized the importance of Muḥammad's *sunnah* to the understanding of it. By his time, major literary works had established a coherent image of the indispensability of Muḥammad's prophethood; in fact, just before the Muʿtazilite controversy began, Ibn Hishām had produced his classic recension of the *Sīrah*, or life, of Muḥammad, composed half a century earlier by Ibn Isḥāq. As in the early Christian Church, these were not merely dogmatic issues. They were rooted in the way ordinary Muslims lived, just as affection for a divine Christ had become popular sentiment by the time Arius and Athanasius debated. Although Muslims lacked an equivalent of the Christian Church, they resolved these issues similarly; like Jesus for the Christians, the Qur'ān for the Muslims was somehow part of God; *ḥadīth*-mindedness and emulation of Muḥammad's *sunnah* had become such an essential part of the daily life of ordinary people that the Muʿtazilite position, as intellectually consistent and attractive as it was, was unmarketable. In a series of forcible inquiries called *miḥnah*, al-Ma'mūn and al-Muʿtaṣim actively persecuted those who, like Ibn Ḥanbal, would not conform; but popular sentiment triumphed and after al-Muʿtaṣim's death the caliph al-Mutawakkil was forced to reverse the stand of his predecessors.

This caliphal failure to achieve doctrinal unity coincided with other crises. By al-Muʿtaṣim's reign the tribal troops were becoming unreliable and the Ṭāhirid governors of Khorāsān more independent. Al-Muʿtaṣim expanded his use of military slaves, finding them more loyal but more unruly, too. Soon he had to house them at Sāmarrā', a new capital north of Baghdad, where the caliphate remained until 892. For most of this period, the caliphs were actually under the control of their slave soldiery; and even though they periodically reasserted their authority, rebellions continued. Many were anti-Muslim, like that of the Iranian Bābak (whose 20-year-long revolt was crushed in 837); but increasingly they were intra-Muslim, like the Kharijite-led revolt of black agricultural slaves (Zanj) in southern Iraq (868–883). By 870, then, the Baghdad–Sāmarrā' caliphate had become one polity among many; its real rulers had no ideological legitimacy. At Córdoba the Umayyads had declared their independence; and the Maghrib was divided among several dynasties of differing persuasions, the Shīʿite Idrīsids, the Kharijite Rustamids, and the Jamāʿi-Sunnite Aghlabids. The former governors of the ʿAbbāsids, the Ṭūlūnids, ruled Egypt and parts of Arabia; Iran was divided between the Ṣaffārids, governors of the ʿAbbāsids in the south, and the Persian Sāmānids in the north.

Unifying forces in Islāmdom

The centrifugal forces represented by these administrative divisions should not obscure, however, the existence of numerous centripetal forces that continued to give Islāmdom, from Andalusia to Central Asia, other types of unity. The ideal of the caliphate continued to be a source of unity after the reality waned; among all the new states, no alternative to the caliphate could replace it. Furthermore, now that Muslims constituted a majority almost everywhere in Islāmdom, conflict began to be expressed almost exclusively in Islāmic rather than anti-Islāmic forms. In spite of continuing intra-Muslim conflict, Muslim worship and belief remained remarkably uniform. The annual pilgrimage to Mecca helped reinforce this underlying unity by bringing disparate Muslims together in a common rite. The pilgrimage, as well as the rise of prosperous regional urban centres, enhanced the trade that traversed Islāmdom regardless of political conflicts; along the trade routes that crisscrossed Eurasia, Islāmdom at its centre, moved not only techniques and goods but ideas as well. A network of credit and banking, caravansaries, and intercity mercantile alliances, tied far-flung regions together. Central was the caravan, then the world's most effective form of transport. The peripatetic nature of education promoted cross-fertilization. Already the *faqīr* (fakir), a wandering mendicant Ṣūfī dervish, was a familiar traveler. Across Islāmdom, similar mosque–market complexes sprang up in most towns; because municipal institutions were rare, political stability so unpredictable, and government intervention kept to a minimum (sometimes by design, more often by necessity), the Sharīʿah and the learned men who carried it became a mainstay of everyday life and social intercourse. The Sharīʿah, along with the widespread affection for the *sunnah* of Muḥammad, regulated, at least among pious Muslims, personal habits of the most specific sort, from the use of scent to the cut of a beard. Comprehensive and practical, the *sunnah* could amuse as well. When asked whether to trust in God or tie one's camel, so a popular *ḥadīth* goes, the Prophet replied, "Trust in God, then tie your camel."

The significance of *ḥadīth* and *sunnah* is represented by the ending date of the period of conversion and crystallization. No one can say exactly when the majority of Islāmdom's population became Muslim. Older scholarship looks to the end of the first quarter of the 9th century; newer scholarship to the beginning of the third quarter. In 870 a man died whose life's work symbolized the consolidation of Islām in everyday life: al-Bukhārī, who produced one of the six collections of *ḥadīth* recognized as authoritative by Jamāʿi-Sunnite Muslims. His fellow collector of *ḥadīth,* Muslim ibn al-Ḥajjāj, died about four years later. About the same time, classical thinkers in other areas of Islāmicate civilization died, among them the great author of *adab,* al-Jāḥiẓ (868/869), the great early ecstatic Ṣūfīs Abu'l Fayḍ Dhu'n-Nūn al-Miṣrī (861) and Abū Yazīd Bisṭāmī (874), the philosopher Yaʿqūb ibn Ishaq aṣ-Ṣabāḥ al-Kindī (870), and the historian of the conquests al-Balādhurī (*c.* 892). Men of different religious and ethnic heritages, they signified, by the last quarter of the 9th century, the full and varied range of intellectual activities of a civilization that had come of age.

Fragmentation and florescence (870–1041)

THE RISE OF COMPETITIVE REGIONS

The unifying forces operative at the end of the period of conversion and crystallization persisted during the period of fragmentation and florescence; but the caliphal lands in Iraq became less central. Even though Baghdad remained preeminent in cultural prestige, important initiatives were being taken from surrounding "regions": Andalusia; the Maghrib and sub-Saharan Africa; Egypt, Syria, and the holy cities (Mecca and Medina); Iraq; and Iran, Afghanistan, Transoxania, and, toward the end of the period, northern India. Regional courts could compete with the ʿAbbāsids and with each other as patrons of culture. Interregional and intra-regional conflicts were often couched in terms of loyalties formed in the period of conversion and crystallization, but local history provided supplemental identities. Although the ʿAbbāsid caliphate was still a focus of concern and debate, other forms of leadership became important. Just as being Muslim no longer meant being Arab, being cultured no longer meant speaking and writing exclusively in Arabic. Certain Muslims began to cultivate a second language of high culture, New Persian. As in pre-Islāmic times, written as well as spoken bilingualism became important. Ethnic differences were blurred by the effects of peripatetic education and shared languages. Physical mobility was so common that many individuals lived and died far from their places of birth. Cultural creativity was so noticeable that this period is often called the Renaissance of Islām.

New Persian language

Economic changes also promoted regional strengths. Although Baghdad continued to profit from its central location, caliphal neglect of Iraq's irrigation system and southerly shifts in the trans-Asian trade promoted the fortunes of Egypt; the opening of the Sahara to Maghribi Muslims provided a new source of slaves, salt, and minerals; and Egyptian expansion into the Mediterranean opened a major channel for Islāmicate influence on medieval Europe. Islāmdom continued to expand, sometimes as the result of aggression on the part of frontier warriors (*ghāzīs*), but more often as the result of trade. The best symbol of this expansiveness is Ibn Faḍlān, who left a provocative account of his mission in 921, on behalf of the Baghdad caliph, to the Volga Bulgars, among whom he met Swedes coming down the river to trade.

By the beginning of the period of fragmentation and florescence the subject populations of most Muslim rulers were predominantly Muslim, and nonsedentary peoples had ceased to play a major role. The period gave way to a much longer period (dated 1041–1405) in which migratory tribal peoples were once again critically important. In 1041 the reign of the Ghaznavid sultan Masʿūd I ended; by then the Ghaznavid state had lost control over the Seljuq Turks in their eastern Iranian domains and thus inaugurated Islāmdom's second era of tribal expansion. Because localism and cosmopolitanism coexisted in the period of fragmentation and florescence, the period is best approached through a region-by-region survey that underscores phenomena of interregional significance.

ANDALUSIA, THE MAGHRIB, AND SUB-SAHARAN AFRICA

Andalusia, far from the centre of Islāmdom, illustrated the extent of ʿAbbāsid prestige and the assertion of local creativity. In the beginning of the period, Islāmicate rule was represented by the Umayyads at Córdoba; established in 756 by a refugee from the ʿAbbāsid victory over the Syrian Umayyads, the Umayyad dynasty in Córdoba had replaced a string of virtually independent deputies of the Umayyad governors in the Maghrib. At first the Cordoban Umayyads had styled themselves *amīr*s, the title also used by caliphal governors and other local rulers; though refugees from ʿAbbāsid hostility, they continued to mention the ʿAbbāsids in the Friday worship session until 773. Their independence was not made official, however, until their best known member, ʿAbd ar-Raḥmān III (ruled 912–961), adopted the title of caliph in 929 and began having the Friday prayer recited in the name of his own house.

The fact that ʿAbd ar-Raḥmān declared his independence from the ʿAbbāsids while he modeled his court after theirs illustrates the period's cultural complexities. Like the ʿAbbāsids' and the Marwānids', ʿAbd ar-Raḥmān's absolute authority was limited by the nature of his army (Berber tribesmen and Slav slaves) and by his dependence on numerous assistants. His internal problems were compounded by external threats, from the Christian kingdoms in the north and the Fāṭimids in the Maghrib (see below). The Umayyad state continued to be the major Muslim presence in the peninsula until 1010, after which time it became, until 1031, but one of many independent city-states. Nowhere is the connection between fragmentation and florescence more evident than in the courts of these *mulūk al-ṭawāʾif,* or "party kings"; for it was they who patronized some of Andalusia's most brilliant Islāmicate culture. This florescence also demonstrated the permeability of the Muslim–Christian frontier. For example, the poet and theologian Ibn Ḥazm (994–1064) composed love poetry, such as *Ṭawk al-ḥamāmah* (*The Ring of the Dove*), which may have contributed to ideas of chivalric love among the Provençal troubadours.

The flowering of Islāmicate culture in Andalusia

In 870 the Maghrib was divided among several dynasties, all but one of foreign origin, and only one of which, the Aghlabids, nominally represented the ʿAbbāsids. The Muslim Arabs had been very different rulers than any of their predecessors—Phoenicians, Romans, Vandals, or Byzantines—who had occupied but not settled. Their interests in North Africa had been secondary to their objectives in the Mediterranean, so they had restricted themselves to coastal settlements, which they used as staging points for trade with the western Mediterranean or as sources of food for their "metropolitan" population. They had separated themselves from the Berbers with a fortified frontier. The Arabs, however, forced away from the coast in order to compete more effectively with the Byzantines, had quickly tried to incorporate the Berbers, who were also pastoralists. One branch of the Berbers, the Ṣanhājah, extended far into the Sahara, across which they had established a caravan trade with blacks in the Sudanic belt. At some time in the 10th century the Ṣanhājah nominally converted to Islām, and their towns in the Sahara began to assume Muslim characteristics. Around 990 a black kingdom in the Sudan, Ghana, extended itself as far as Audaghost, the Ṣanhājah centre in the Sahara. Thus was black Africa first brought into contact with the Muslim Mediterranean, and thus were the conditions set for dramatic developments in the Maghrib during the 12th and 13th centuries (see below, *Migration and renewal*).

In the late 9th century the Maghrib was unified and freed

Unification of the Maghrib

from outside control for the first time. Paradoxically, this independence was achieved by outsiders associated with an international movement of political activism and subversion. Driven underground by ʿAbbāsid intolerance and a maturing ideology of covert revolutionism, the Ismāʿīlī Shīʿites had developed mechanisms to maintain solidarity and undertake political action. These mechanisms can be subsumed under the term *daʿwah,* the same word that had been used for the movement that brought the ʿAbbāsids to power. The *daʿwah*'s ability to communicate rapidly over a large area rested on its traveling operatives as well as on a network of local cells. In the late 9th century an Ismāʿīlī movement, nicknamed the Qarāmiṭah (Qarmatians), had seriously but unsuccessfully threatened the ʿAbbāsids in Syria, Iraq, and Bahrain. Seeking other outlets, a Yemeni operative known as Abū ʿAbd Allāh ash-Shīʿī made contact, on the occasion of the *ḥajj,* with representatives of a Berber tribe that had a history of Kharijite hostility to caliphal control. The *ḥajj* had already become a major vehicle for tying Islāmdom's regions together, and Abū ʿAbd Allāh's movement was only one of many in the Maghrib that would be inaugurated thereby.

In 901 Abū ʿAbd Allāh arrived in the Petite Kabylie (in present-day Algeria); for eight years he prepared for an *imām,* preaching of a millennial restoration of justice after an era of foreign oppression. After conquering the Aghlabid capital al-Qayrawān (in present-day Tunisia), he helped free from a Sijilmassa prison his *imām,* ʿUbayd Allāh, who declared himself the *mahdī,* using a multivalent word that could have quite different meanings for different constituencies. Some Muslims applied *mahdī* to any justice-restoring divinely guided figure; others, including many Jamāʿi-Sunnites, to the apocalyptic figure expected to usher in the millennium before the Last Judgment; and still others, including most Shīʿites, to a returned or restored *imām.* Abū ʿAbd Allāh's followers may have differed in their expectations, but the *mahdī* himself was unequivocal: he was a descendant of ʿAlī and Fāṭimah through Ismāʿīl's disappeared son and therefore was a continuation of the line of the true *imām.* He symbolized his victory by founding a new capital named, after himself, al-Mahdīyah (in present-day Tunisia). During the next half century the "Fāṭimids" tried with limited success to expand westward into the Maghrib and north into the Mediterranean, where they made Sicily a naval base (912–913); but their major goal was Egypt, nominally under ʿAbbāsid control. From Egypt they would challenge the ʿAbbāsid caliphate itself. In 969 the Fāṭimid army conquered the Nile Valley and advanced into Palestine and southern Syria as well.

EGYPT, SYRIA, AND THE HOLY CITIES

The Fāṭimids in Cairo

The Fāṭimids established a new and glorious city, al-Qahirah ("The Victorious"; Cairo), to rival ʿAbbāsid Baghdad. They then adopted the title of caliph, laying claim to be the legitimate rulers of all Muslims as well as head of all Ismāʿīlīs. Now three caliphs reigned in Islāmdom, where there was supposed to be only one. In Cairo the Fāṭimids founded a great mosque–school complex, al-Azhar. They fostered local handicraft production and revitalized the Red Sea route from India to the Mediterranean. They built up a navy to trade as well as to challenge the Byzantines and underscore the ʿAbbāsid caliph's failure to defend and extend the frontiers. Fāṭimid occupation of the holy cities of Mecca and Medina, complete by the end of the 10th century, had economic as well as spiritual significance: it reinforced the caliph's claim to leadership of all Muslims; provided wealth; and helped him keep watch on the West Arabian coast, from the Hejaz to the Yemen, where a sympathetic Zaydī Shīʿite dynasty had ruled since 897. Fāṭimid presence in the Indian Ocean was even strong enough to establish an Ismāʿīlī missionary in Sind. The Fāṭimids patronized the arts; Fāṭimid glass and ceramics were some of Islāmdom's most brilliant. As in other regions, imported styles and tastes were transformed by or supplemented with local artistic impulses, especially in architecture, the most characteristic form of Islāmicate art.

The reign of one of the most unusual Fāṭimid caliphs, al-Ḥākim, from 996 to 1021, again demonstrated the interregional character of the Ismāʿīlī movement. Historians describe al-Ḥākim's personal habits as eccentric, mercurial, and unpredictable to the point of cruelty; his religious values, as inconsistent with official Ismāʿīlī teachings, tending toward some kind of accommodation with the Jamāʿi-Sunnite majority. After he vanished under mysterious circumstances, his religious revisionism was not pursued by his successors or by the Ismāʿīlī establishment in Egypt; but in Syria it inspired a peasant revolt that produced the Druze, who still await al-Ḥākim's return.

When the Fāṭimids expanded into southern Syria, another Shīʿite dynasty, the Ḥamdānid, of Bedouin origin, had been ruling northern Syria from Mosul since 905. In 944 a branch of the family had taken Aleppo; under the leadership of their most famous member, Sayf ad-Dawlah (ruled 945–967), the Ḥamdānids responded aggressively to renewed Byzantine expansionism in eastern Anatolia. They ruled from Aleppo until they were absorbed by the Fāṭimids after 1004; at their court some of Islāmdom's most lastingly illustrious writers found patronage. Two notable examples are the poet al-Mutanabbī (915–965), who illustrated the importance of the poet as a premodern press agent of the court, and al-Fārābī, who tried to reconcile reason and revelation.

Intellectual life under the Ḥamdānids

Al-Fārābī contributed to the ongoing Islāmization of Hellenistic thought. *Falsafah,* the Arabic cognate for the Greek *philosophia,* included metaphysics and logic, as well as the positive sciences, such as mathematics, music, astronomy, and anatomy. *Faylasūfs* often earned their living as physicians, astrologers, or musicians. The *faylasūf*'s whole way of life, like that of the *adīb,* reflected his studies. It was often competitive with that of more self-consciously observant Muslims because the *faylasūf* often questioned the relationship of revelation to real truth. The *faylasūfs* felt free to explore inner truths not exposed to the view of ordinary people; they practiced prudent concealment (*taqīyah*) of their deeper awareness wherever making it public might endanger the social order. The *faylasūfs* shared the principle of concealment with the Shīʿites; both believed, for rather different reasons, that inner truth was accessible to only a very few. This esotericism had counterparts in all premodern societies, where learning and literacy were severely restricted.

IRAQ

Cultural flowering in Iraq. By the late 9th and early 10th centuries the last remnant of the caliphal state was Iraq, under control of the Turkic soldiery. Political decline and instability did not preclude cultural creativity and productivity, however. In fact, Iraq's "generation of 870," loosely construed, contained some of the most striking and lastingly important figures in all of early Islāmicate civilization. Three of them illustrate well the range of culture in late 9th- and early 10th-century Iraq: the historian and Qurʾānic exegete aṭ-Ṭabarī (*c.* 839–923), the theologian Abū al-Ḥasan al-Ashʿarī (*c.* 873–*c.* 935), and the ecstatic mystic al-Ḥallāj (*c.* 858–922).

Aṭ-Ṭabarī

Abū Jaʿfar Muḥammad ibn Jarīr was born in Ṭabaristān, south of the Caspian Sea, and as a young man he traveled to Baghdad. Rarely could a man earn his living from religious learning; unless he found patronage, he would probably engage in trade or a craft. All the more astounding was the productivity of scholars like aṭ-Ṭabarī, who said that he produced 40 leaves a day for 40 years. The size of his extant works, which include a commentary on the Qurʾān and a universal history, testifies to the accuracy of his claim. His history is unique in sheer size and detail and especially in its long-term impact. His method involved the careful selection, organization, and juxtaposition of separate and often contradictory accounts cast in the form of *ḥadīth.* This technique celebrated the *ummah*'s collective memory and established a range of acceptable disagreement.

Al-Ashʿarī

Al-Ashʿarī, from Basra, made his contribution to systematic theological discourse (*kalām*). He had been attracted early to a leading Muʿtazilite teacher, but he broke away at the age of 40. He went on to use Muʿtazilite methods of reasoning to defend popular ideas such as the eternality and literal truth of the Qurʾān, and the centrality of Muḥammad's *sunnah* as conveyed by the *ḥadīth.*

Where his approach yielded objectionable results, such as an anthropomorphic rendering of God or a potentially polytheistic understanding of his attributes, al-Ashʿarī resorted to the principle of *bilā kayfah* ("without regard to the how"), whereby a person of faith accepts that certain fundamentals are true without regard to how they are true and that divine intention is not always accessible to human intelligence. Al-Ashʿarī's harmonization also produced a simple creed, which expressed faith in God, his angels, and his books, and affirmed belief in Muḥammad as God's last messenger and in the reality of death, physical resurrection, the Last Judgment, and heaven and hell. Taken together, aṭ-Ṭabarī's historiography and al-Ashʿarī's theology symbolize the consolidation of Jamāʿi-Sunnite, Sharīʿah-minded thought and piety.

Al-Ḥallāj

The most visible and powerful 10th-century exponent of Ṣūfism was al-Ḥallāj. By his day, Ṣūfism had grown far beyond its early forms, which were represented by al-Ḥasan al-Baṣrī (died 728), who practiced *zuhd,* or rejection of the world, and by Rābiʿah al-ʿAdawīyah (died 801), who formulated the Ṣūfī ideal of a disinterested love of God. The mystics Abū Yazīd Bisṭāmī (died 874) and al-Junayd (died 910) had begun to pursue the experience of unity with God, first by being "drunk" with his love and with love of him, and then by acquiring life-transforming self-possession and control. Masters (called *shaykh*s or *pīr*s) were beginning to attract disciples (*murīd*s) to their way. Like other Muslims who tried to go "beyond" the Sharīʿah to inner truth, the Ṣūfīs practiced concealment of inner awareness (*taqīyah*). Al-Ḥallāj, one of al-Junayd's disciples, began to travel and preach publicly, however. His success was disturbing enough for the authorities in Baghdad to have him arrested and condemned to death; he was tortured and beheaded, and finally his body was burned. Yet his career had shown the power of Ṣūfism, which would by the 12th century become an institutionalized form of Islāmic piety.

The Būyid dynasty. Long before, however, a major political change occurred at Baghdad. In 945 control over the caliphs passed from their Turkish soldiery to a dynasty known as the Būyids or Buwayhids. The Būyids came from Daylam, near the southern coast of the Caspian Sea. Living beyond the reach of the caliphs in Baghdad, its residents had identified with Imāmī Shīʿism. By about 930, three sons of a fisherman named Būyeh had emerged as leaders in Daylam. One of them conquered Baghdad, not replacing the caliph but ruling in his name. The fact that they were Shīʿite, as were the Idrīsids, Fāṭimids, and Ḥamdānids, led scholars to refer to the period from the mid-10th to mid-11th century as the Shīʿite century.

Būyid patronage of culture

Like other contemporary rulers, the Būyids were patrons of culture, especially of speculative thought (Shīʿism, Muʿtazilism, *kalām,* and *falsafah*). Jamāʿi-Sunnite learning continued to be patronized by the caliphs and their families. The Būyids favoured no one party over another. However, their openness paradoxically invited a hardening in Jamāʿi-Sunnite thought. Būyid attempts to maintain the cultural brilliance of the court at Baghdad were limited by a decline in revenue occasioned partly by a shift in trade routes to Fāṭimid Egypt, and partly by long-term neglect of Iraq's irrigation works. The caliphs had occasionally made land assignments (*iqṭāʿ*s) to soldiers in lieu of paying salaries; now the Būyids extended the practice to other individuals and thus removed an important source of revenue from central control. After 983, Būyid territories were split among various members of the family, and pressure was applied to their borders from both the west (by Ḥamdānids and Fāṭimids) and the east (by Sāmānids, Ghaznavids, and Seljuqs; see below).

The economic difficulties of Būyid Iraq promoted urban unrest, accounts of which provide a rare glimpse into the lives of ordinary Muslim town dwellers. Numerous movements served as outlets for socioeconomic grievances, directed most often toward the wealthy or the military. The concentration of wealth in the cities had produced a bipolar stratification system conveyed in the sources by a pair of words, *khāṣṣ* (special) and *ʿāmm* (ordinary). In the environment of 10th- and 11th-century Iraq, an instance of rising food prices or official maltreatment could easily spark riots of varying size, duration, and intensity. Strategies for protest included raiding, looting, and assault. Some movements were more coherently ideological than others, and various forms of piety could reflect socioeconomic distinctions. Some movements were particularly attractive to artisans, servants, and soldiers, as was the case with the proponents of Ḥadīth, whose mentor, Aḥmad ibn Ḥanbal (died 855), was viewed as a martyr because of his suffering at the hands of the Caliph. Other forms of piety, such as Shīʿism, could be associated with wealthier elements among the landowning and merchant classes.

Social action associations: *ʿayyār* and *futūwah*

Beneath the more organized forms of social action lay a more fluid kind of association, most often described by the labels *ʿayyār* and *futūwah.* These terms refer to individuals acting in concert, as needed, on the basis of certain rough-hewn concepts of proper male public behaviour. Such associations had counterparts in the late Hellenistic world, just as they have parallels in the voluntary protective associations formed in the 19th and 20th centuries whenever official institutions of protection have been either chronically or temporarily deficient. For some of the Islāmicate "gangs" or "clubs," thuggery may have been the norm; for others, the figure of the fourth caliph and first *imām,* ʿAlī, seems to have provided an exemplar. Even though Shīʿites had become a separate group with a distinctive interpretation of ʿAlī's significance, a more generalized affection for the family of the Prophet, and especially for ʿAlī, was widespread among Jamāʿi-Sunnites. ʿAlī had come to be recognized as the archetypal young male (*fatā*); a related word, *futūwah,* signified groups of young men who pursued such virtures as courage, aiding the weak, generosity, endurance of suffering, love of truth, and hospitality.

Premodern Islāmicate societies were characterized by a high degree of fluidity, occasionalism, and voluntarism in the structuring of associations, organizations, loyalties, and occupations. Although all societies must develop ways to maintain social boundaries, ease interaction among groups, and buffer friction, the ways in which Muslim societies have fulfilled these needs seem unusually difficult to delineate. For example, in Muslim cities of the period under discussion, the only official officeholders were appointees of the central government, such as the governor; the *muḥtasib,* a transformed Byzantine *agoranomos* who was monitor of public morality as well as of fair-market practice; or the *sahib ash-shurtah,* head of the police. In the absence of an organized church or ordained clergy, those whose influence derived from piety or learning were influential because they were recognized as such, not because they were appointed; and men of very different degrees of learning might earn the designation of *ʿālim.* Although the ruler was expected to contribute to the maintenance of public services, neither he nor anyone else was obligated to do so. Though the ruler might maintain prisons for those whose behaviour he disapproved, the local *qāḍī*s had need of none, relying generally on persuasion or negotiation and borrowing the caliphal police on the relatively rare occasion on which someone needed to be brought before them by force. There was no formalized mode of succession for any of the dynasties of the time. Competition, sometimes armed, was relied upon to produce the most qualified candidate.

Patronage was an important basis of social organization. The family served as a premodern welfare agency; where it was absent, minimal public institutions, such as hospitals, provided. One of the most important funding mechanisms for public services was a private one, the *waqf.* The *waqf* provided a legal way to circumvent the Sharīʿah's requirement that an individual's estate be divided among many heirs. Through a *waqf,* an individual could endow an institution or group with all or part of his estate, in perpetuity, before his death. A *waqf* might provide books for a school, candles or mats for a mosque, salaries of religious functionaries, or land for a hospital or caravansary. *Waqf* money or lands were indivisible, although they might contribute to the welfare of a potential heir who happened to be involved in the *waqf*-supported activity. The *waqf,* like other forms of patronage, provided needed social services without official intervention. On other occasions, wealthy individuals, especially those connected with the ruling family, might simply patronize favourite activities. In addition to patronage, many other overlapping ties bound individual Muslims together: loyalties to an occupation—soldier, merchant, learned man, artisan, government worker; loyalties to a town or neighbourhood, or to a form of piety, or to persons to whom one made an oath for a specific purpose; and ties to patron or to family, especially foster-parentage (*istināʿ*), the counterpart of which was significant in medieval Christendom.

The relationship of individual and group action

The Qurʾān and Sharīʿah discouraged corporate responsibility in favour of individual action; even the legal scope of partnership was limited. Yet the unstable political realities that had militated against the emergence of broad-based institutions sometimes called for corporate action, as when a city came to terms with a new ruler or invader. In those cases, a vaguely defined group of notables, known usually as *aʿyān,* might come together to represent their city in negotiations, only to cease corporate action when the more functional small-group loyalties could safely be resumed. Within this shifting frame of individuals and groups, the ruler was expected to maintain a workable, if not equitable, balance. More often than not the real ruler was a local *amīr* of some sort. For this reason, the de facto system of rule that emerged during this period, despite the persistence of the central caliphate in Baghdad, has sometimes been referred to as the *aʿyān–amīr* system.

The city's physical and social organization reflected this complex relationship between public and private, and between individual and group: physically separated quarters; multiple markets and mosques; mazelike patterns of narrow streets and alleys with dwellings oriented toward an inner courtyard; an absence of public meeting places other than bath, market, and mosque; and the concentration of social life in private residences. The *qāḍī* and *adīb* at-Tanūkhī provides a lively and humorous picture of 10th-century Baghdad, of a society of individuals with overlapping affiliations and shifting statuses: saints and scoundrels, heroes and rogues, rich men and poor. This mobility is illustrated by at-Tanūkhī's boast to a rival, "My line begins with me while yours ends with you." The prose genre of *maqāmah,* said to have been invented by al-Hamadhānī (died 1007), recounted the exploits of a clever, articulate scoundrel dependent on his own wits for his survival and success.

IRAN, AFGHANISTAN, AND INDIA

In the middle of the "Shīʿite century" a major Sunnite revival occurred in eastern Islāmdom in connection with the emergence of the second major language of Islāmicate high culture, New Persian. This double revival was accomplished by two Iranian dynasties, the Sāmānids and the Ghaznavids; Ghaznavid zeal even spilled over into India.

The Sāmānids. The Sāmānid dynasty (819–999) stemmed from a local family appointed by the ʿAbbāsids to govern at Bukhara and Samarkand. Gradually the Sāmānids had absorbed the domains of the rebellious Ṭāhirids and Ṣaffārids in northeastern Iran and reduced the Ṣaffārids to a small state in Sīstān. The Sāmānids, relying on Turkic slave troops, also managed to contain the migratory pastoralist Turkic tribes who continually pressed on Iran from across the Oxus River. In the 950s they even managed to convert some of these Turkic tribes to Islām.

The Sāmānid court at Bukhara attracted leading scholars, such as the philosophers Abū Bakr ar-Rāzī (died 925) and Avicenna (Ibn Sīnā; 980–1037), who later worked for the Būyids; and the poet Ferdowsī (died *c.* 1020). Though not Shīʿites, the Sāmānids expressed an interest in Shīʿite thought, especially in its Ismāʿīlī form, which was then the locus of so much intellectual vitality. The Sāmānids also fostered the development of a second Islāmicate language of high culture, New Persian. It combined the grammatical structure and vocabulary of spoken Persian with vocabulary from Arabic, the existing language of high culture in Iran. A landmark of this "Persianizing" of Iran was Ferdowsī's epic poem, the *Shāh-nāmeh* ("Book of Kings"), written entirely in New Persian in a long-couplet form (*masnavi*) derived from Arabic. Covering several thousand years of detailed mythic Iranian history, Ferdowsī brought Iran's ancient heroic lore, and its hero Rustam, into Islāmicate literature and into the identity of self-consciously Iranian Muslims. He began to compose the poem under the Sāmānids; but he dedicated the finished work to a dynasty that had meanwhile replaced them, the Ghaznavids.

Sebüktigin

The Ghaznavids. The Ghaźnavid dynasty was born in a way that had become routine for Islāmicate polities. Sebüktigin (ruled 977–997), a Sāmānid Turkic slave governor in Ghazna (now Ghaznī), in the Afghan mountains, made himself independent of his masters as their central power declined. His eldest son, Maḥmūd, expanded into Būyid territory in western Iran, identifying himself staunchly with Sunnite Islām. Presenting himself as a frontier warrior against the pagans, Maḥmūd invaded and plundered northwestern India, establishing a permanent rule in the Punjab; but it was through ruling Iran, which gave a Muslim ruler true prestige, that Maḥmūd sought to establish himself. He declared his loyalty to the ʿAbbāsid caliph, whose "investiture" he sought, and expressed his intention to defend Sunnite Islām against the Shīʿite Būyids. Although he and his regime were proud of their Turkic descent, Maḥmūd encouraged the use of New Persian, with its echoes of pre-Islāmic Iranian glory, for administration and for prose as well as poetry. This combination of Turkic identity and Persian language would characterize and empower many other Muslim rulers. To Ghazna Maḥmūd brought, sometimes by force, writers and artisans who could adorn his court. Among these was al-Bīrūnī (973–*c.* 1050), whose scholarly achievements no contemporary could rival. Before being brought to Ghazna, al-Bīrūnī had served the Sāmānids and the Khwārazm-Shāhs, a local dynasty just west of the Oxus River. His works included studies of astronomy (he even suggested a heliocentric universe), gems, drugs, mathematics, and physics; but his most famous book, inspired by accompanying Maḥmūd on his Indian campaigns, was a survey of Indian life, language, religion, and culture.

Like most other rulers of the day, Maḥmūd styled himself *amīr* and emphasized his loyalty to the caliph in Baghdad; but he and later Ghaznavid rulers also called themselves by the Arabic word, *sulṭān* (sultan). Over the next five centuries the office of sultan would become an alternative to caliph. The Ghaznavid state presaged other changes as well, especially by stressing the cleavage between ruler and ruled and by drawing into the ruling class not only the military but also the bureaucracy and the learned establishment. So tied was the ruling establishment to the ruler that it even moved with him on campaign. Ghaznavid "political theory" shared with other states the concept of the circle of justice or circle of power; *i.e.,* that justice is best preserved by an absolute monarch completely outside society; that such a ruler needs an absolutely loyal army; and that maintaining such an army requires prosperity, which in turn depends on the good management of an absolute ruler.

Bu'l-Fażl-i Bayhaqī (995–1077) worked in the Ghaznavid chancery and wrote a remarkable history of the Ghaznavids, the first major prose work in New Persian. He exhibited the broad learning of even a relatively minor figure at court; in his history he combined the effective writing skills of the chancery employee, the special knowledge of Qurʾān and *ḥadīth,* and the sophisticated and entertaining literature—history, poetry, and folklore—that characterized the *adīb.* He provided a vivid picture of life at court, graphically portraying the pitfalls of military absolutism—the dependence of the monarch on a fractious military and a large circle of assistants and advisors, who could mislead him and affect his decision making through internecine maneuvering and competition. In the reign of Maḥmūd's son, Masʿūd I, the weaknesses in the system had already become glaringly apparent. At the Battle of Dandānqān (1040), Masʿūd lost control of Khorāsān, his main holding in Iran, to the pastoralist Seljuq Turks; he then decided to withdraw to Lahore in his Indian domains, from which his successors ruled until overtaken by the Ghūrids in 1186.

THE DECLINE OF THE CALIPHATE AND RISE OF EMIRATES

By the end of Masʿūd's reign, government in Islāmdom had become government by *amīr.* Caliphal centralization had lasted 200 years; and even after the caliphal empire became too large and complex to be ruled from a single centre, the separate emirates that replaced it all defined their legitimacy in relation to it, for or against. In fact, the caliphate's first systematic description and justification was undertaken just when its impracticality was being demonstrated. As the Ghaznavids were ruling in Iran as "appointed" defenders of the caliph, a Baghdadi legal scholar named al-Māwardī retrospectively delineated the minimal requirements of the caliphate and tried to explain why it had become necessary for caliphal powers to be "delegated" in order for the *ummah*'s security to be maintained. Whereas earlier legists had tied the caliph's legitimacy to his defense of the borders, al-Māwardī separated the two, maintaining the caliph as the ultimate source of legitimacy and the guardian of pan-Islāmic concerns, and relegating day-to-day government to his "appointees." Al-Māwardī may have hoped that the Ghaznavids would expand far enough to be "invited" by the caliph to replace the uninvited Shīʿite Būyids. This replacement did occur, three years before al-Māwardī's death; however, it was not the Ghaznavids who appeared in Baghdad but rather the migratory pastoralist Turks who had meanwhile replaced them. The Seljuqs joined many other migrating groups to produce the next phase of Islāmicate history.

Migration and renewal (1041–1405)

Relationship of sedentary and migratory peoples

During this period, migrating peoples once again played a major role, perhaps greater than that of the Arabs during the 7th and 8th centuries. No other civilization in premodern history experienced so much in-migration, especially of alien and disruptive peoples, or showed a greater ability to assimilate as well as to learn from outsiders. Nowhere has the capacity of a culture to redefine and incorporate the strange and the foreign been more evident. In this period, which ends with the death in 1405 of Timur (Tamerlane), the last great tribal conqueror, the tense yet creative relationship between sedentary and migratory peoples emerged as one of the great themes of Islāmicate history, played out as it was in the centre of the great arid zone of Eurasia. Because this period can be seen as the history of peoples as well as of regions, and because the mobility of those peoples brought them to more than one cultural region, this period should be treated group by group rather than region by region.

As a general term "migrating" peoples is preferable because it does not imply aimlessness, as "nomadic" does; or herding, as "pastoralist" does; or kin-related, as "tribal" does. "Migrating" focuses simply on movement from one home to another. Although the Franks, as the crusaders are called in Muslim sources, differed from other migrating peoples, most of whom were pastoralists related by kinship, they too were migrating warriors organized to invade and occupy peoples to whom they were hostile and alien. Though not literally tribal, they appeared to behave like a tribe with a distinctive way of life and a solidarity based on common values, language, and objectives. Viewing them as alien immigrants comparable to, say, the Mongols, helps to explain their reception: how they came to be assimilated into the local culture and drawn into the intra-Muslim factional competition and fighting that was under way in Syria when they arrived.

TURKS

For almost 400 years a succession of Turkic peoples entered eastern Islāmdom from Central Asia. These nearly continuous migrations can be divided into three phases: Seljuqs (1055–92), Mongols (1256–1411), and neo-Mongols (1369–1405). Their long-term impact, more constructive than destructive on balance, can still be felt through the lingering heritage of the great Muslim empires they inspired. The addition of tribally organized warrior Turks to the already widely used Turkic slave soldiery gave a single ethnic group an extensive role in widening the gap between rulers and ruled.

Seljuq Turks. The Seljuqs were a family among the Oğuz Turks, a label applied to the migratory pastoralists of the Syrdarya–Oxus basin. Their name has come to stand for the group of Oğuz families led into Ghaznavid Khorāsān after they had been converted to Sunnite Islām, probably by Ṣūfī missionaries after the beginning of the 11th century. In 1040 the Seljuqs' defeat of the Ghaznavid sultan allowed them to proclaim themselves rulers of Khorāsān. Having expanded into western Iran as well, Toghrïl Beg, also using the title "sultan," was able to occupy Baghdad (1055) after "petitioning" the ʿAbbāsid caliph for permission. The Seljuqs quickly took the remaining Būyid territory and began to occupy Syria, whereupon they encountered Byzantine resistance in the Armenian highlands. In 1071 a Seljuq army under Alp-Arslan defeated the Byzantines at Manzikert north of Lake Van; while the main Seljuq army replaced the Fāṭimids in Syria, large independent tribal bands occupied Anatolia, coming closer to the Byzantine capital than had any other Muslim force.

Policies of Niẓām al-Mulk. The Seljuqs derived their legitimacy from investiture by the caliph, and from "helping" him reunite the *ummah;* yet their governing style prefigured the emergence of true alternatives to the caliphate. Some of their Iranian advisers urged them to restore centralized absolutism as it had existed in pre-Islāmic times and in the period of Marwānid–ʿAbbāsid strength. The best known proponent was Niẓām al-Mulk, chief minister to the second and third Seljuq sultans, Alp-Arslan and Malik-Shāh. Niẓām al-Mulk explained his plans in his *Seyāsat-nāmeh,* one of the best known manuals of Islāmicate political theory and administration. He was unable, however, to persuade the Seljuq sultans to assert enough power over other tribal leaders. Eventually the Seljuq sultans, like so many rulers before them, alienated their tribal supporters and resorted to the costly alternative of a Turkic slave core, whose leading members were appointed to tutor and train young princes of the Seljuq family to compete for rule on the death of the reigning sultan. The tutors were known as *atabeg*s; more often than not, they became the actual rulers of the domains assigned to their young charges, cooperating with urban notables (*aʿyān*) in day-to-day administration.

Founding of *madrasah*s

Although Niẓām al-Mulk was not immediately successful, he did contribute to long-term change. He encouraged the establishment of state-supported schools (*madrasahs*); those he personally patronized were called Niẓāmīyahs. The most important Niẓāmīyah was founded in Baghdad in 1067; here Niẓām al-Mulk gave government stipends to teachers and students whom he hoped he could subsequently not only appoint to the position of *qāḍī* but also recruit for the bureaucracy. Systematic and broad instruction in Jamāʿi-Sunnite learning would counteract the disruptive influences of non-Sunnite or anti-Sunnite thought and activity, particularly the continuing agitation of Ismāʿīlī Muslims. In 1090 a group of Ismāʿīlīs established themselves in a mountain fortress at Alamūt in the mountains of Daylam. From there they began to coordinate revolts all over Seljuq domains. Nominally loyal to the Fāṭimid caliph in Cairo, the eastern Ismāʿīlīs confirmed their growing independence and radicalism by supporting a failed contender for the Fāṭimid caliphate, Nizār. For that act they were known as the Nizārī Ismāʿīlīs. They were led by Ḥasan-e Ṣabbāḥ and were dubbed by their detractors the *hashīshīyah* (assassins) because they practiced political murder while they were allegedly under the influence of hashish.

Niẓām al-Mulk's *madrasah* system enhanced the prestige and solidarity of the Jamāʿi-Sunnite *ʿulamāʾ* without actually drawing them into the bureaucracy or combating anti-Sunnite agitation, but it also undermined their autonomy. It established the connection between state-supported education and office holding, and it subordinated the spiritual power and prestige of the *ʿulamāʾ* to the indispensable physical force of the military *amīr*s. Niẓām al-Mulk unintentionally encouraged the independence of these *amīr*s by extending the *iqṭāʿ* system beyond Būyid practice; he regularly assigned land revenues to individual military officers, assuming that he could keep them under bureaucratic control. When that failed, his system increased the *amīr*s' independence and drained the central treasury.

The *madrasah* system had other unpredictable results

Al-Ghazālī's teaching

that can be illustrated by al-Ghazālī, who was born in 1058 at Ṭūs and in 1091 was made head of the Baghdad Niẓāmīyah. For four years, to great admiration, he taught both *fiqh* and *kalam* and delivered critiques of *falsafah* and Ismāʿīlī thought. According to his autobiographical work *Al-Munqidh min aḍ-ḍalāl* (*The Deliverer from Error*), the more he taught, the more he doubted, until his will and voice became paralyzed. In 1095 he retreated from public life, attempting to arrive at a more satisfying faith. He undertook a radically skeptical reexamination of all of the paths available to the pious Muslim, culminating in an incorporation of the active, immediate, and inspired experience of the Ṣūfīs into the Sharīʿah-ordered piety of the public cult. For his accomplishments, al-Ghazālī was viewed as a renewer (*mujaddid*), a role expected by many Muslims to be filled by at least one figure at the turn of every Muslim century.

Ṭarīqah fellowships. In the 12th century Muslims began to group themselves into *ṭarīqah*s, fellowships organized around and named for the *ṭarīqah* ("way" or "path") of given masters. Al-Ghazālī may have had such a following himself. One of the first large-scale orders, the Qādirīyah, formed around the teachings of ʿAbd al-Qādir al-Jīlānī of Baghdad. Though rarely monastic in the European sense, the activities of a *ṭarīqah* often centred around assembly halls (called *khānqāh, zāwīyah,* or *tekke*) that could serve as places of retreat or accommodate special spiritual exercises. The *dhikr,* for example, is a ceremony in which devotees meditated on the name of God to the accompaniment of breathing exercises, music, or movement, so as to attain a state of consciousness productive of a sense of union with God. Although shortcuts and excesses have often made Ṣūfism vulnerable to criticism, its most serious practitioners have conceived of it as a disciplined extension of Sharīʿah-minded piety, not an escape. In fact, many Ṣūfīs have begun their path through supererogatory fulfillment of standard ritual requirements.

Thousands of *ṭarīqah*s sprang up over the centuries, some associated with particular occupations, locales, or classes. It is possible that by the 18th century most adult Muslim males had some connection with one or more *ṭarīqahs*. The structure of the *ṭarīqah* ensued from the charismatic authority of the master, who, though not a prophet, replicated the direct intimacy that the prophets had shared with God. This quality he passed on to his disciples through a hierarchically ordered network that could extend over thousands of miles. The *ṭarīqah*s thus became powerful centripetal forces among societies in which formal organizations were rare; but the role of the master became controversial because followers often made saints or intercessors of especially powerful Ṣūfī leaders and made shrines or pilgrimage sites of their tombs or birthplaces. Long before these developments could combine to produce stable alternatives to the caliphal system, Seljuq power had begun to decline, only to be replaced for a century and a half with a plethora of small military states. When the Frankish crusaders arrived in the Holy Land in 1099, no one could prevent them from quickly establishing themselves along the eastern Mediterranean coast.

FRANKS

The call for the Crusades. At the Council of Clermont in 1095 Pope Urban II responded to an appeal from the Byzantine emperor for help against the Seljuq Turks, who had expanded into western Anatolia just as the Kipchak Turks in the Ukraine had cut off newly Christian Russia from Byzantium. The First Crusade, begun the next year, brought about the conquest of Jerusalem in 1099. The Christian Reconquista (reconquest) of Spain was already under way, having scored its first great victory at Toledo in 1085. Ironically, modern historiography has concentrated on the crusades that failed and virtually ignored the ones that succeeded. In the four centuries between the fall of Toledo and the fall of Granada (1492), Spanish Christians replaced Muslim rulers throughout the Iberian Peninsula, although Muslims remained as a minority under Christian rule until the early 17th century. In the 200 years from the fall of Jerusalem to the end of the Eighth Crusade (1291), western European crusaders failed to halt the Turkish advance or to establish a permanent presence in the Holy Land. By 1187 local Muslims had managed to retake Jerusalem and thereby contain Christian ambitions permanently. By the time of the Fourth Crusade (1202–04) the crusading movement had been turned inward against Christian heretics such as the Byzantines.

Successful crusades in the Iberian Peninsula

Effect of the Crusades in Syria. The direct impact of the Crusades on Islāmdom was limited largely to Syria. For the century during which western European Christians were a serious presence there, they were confined to their massive coastal fortifications. The crusaders had arrived in Syria at one of its most factionalized periods prior to the 20th century. Seljuq control, never strong, was then insignificant; local Muslim rule was anarchic; the Seljuq regime in Baghdad was competing with the Fāṭimid regime in Egypt; and all parties in Syria were the target of the Nizārī Ismāʿīlī movement at Alamūt. The crusaders soon found it difficult to operate as more than just another faction. Yet the significance of the crusaders as a force against which to be rallied should not be underestimated any more than should the significance of Islāmdom as a force against which Christendom could unite.

The crusaders' situation encouraged interaction with the local population and even assimilation. They needed the food, supplies, and services available in the Muslim towns. Like their Christian counterparts in Spain, they took advantage of the enemy's superior skills, in medicine and hygiene, for example. Because warfare was seasonal and occasional, they spent much of their time in peaceful interaction with their non-Christian counterparts. Some early-generation crusaders intermarried with Arab Muslims or Arab Christians and adopted their personal habits and tastes, much to the dismay of Christian latecomers. An intriguing account of life in Syria during the Crusades can be found in the *Kitāb al-Iʿtibār* ("Book of Reflection"), the memoirs of Usamah ibn Munqidh (1095–1188). Born in Syria, he was a small boy when the first generation of Franks controlled Jerusalem. As an adult he fought with Saladin (see below) and lived to see him unite Egypt with Syria and restore Jerusalem to Muslim control. In this fine example of Islāmicate autobiographical writing, Usamah draws a picture of the Crusades not easily found in European sources: Christians and Muslims observing, and sometimes admiring, each others' skills and habits, from the battlefield to the bathhouse. Although the Franks in Syria were clearly influenced by the Muslims, the Crusades seem to have contributed relatively little to the overall impact of Islāmicate culture on Europe, even though they constituted the most prolonged direct contact.

Although the crusaders never formed a united front against the Muslims, Syrian Muslims did eventually form a united front against them, largely through the efforts of the family of the *amīr* Zangī, a Turkic slave officer appointed Seljuq representative in Mosul in 1127. After Zangī had extended his control through northern Syria, one of his sons and successors, Nureddin (Nūr ad-Dīn), based at Aleppo, was able to tie Zangī's movement to the frontier warrior (*ghāzī*) spirit. This he used to draw together urban and military support for a *jihād* against the Christians. After taking Damascus, he established a second base in Egypt. He offered help to the failing Fāṭimid regime in return for being allowed to place one of his own lieutenants, Saladin (Ṣalāḥ ad-Dīn Yūsuf ibn Ayyūb), as chief minister to the Fāṭimid caliph, thus warding off a crusader alliance with the Fāṭimids. This action gave Nureddin two fronts from which to counteract the superior seaborne and naval support the crusaders were receiving from western Europe and the Italian city-states. Three years before Nureddin's death in 1174, Saladin substituted himself for the Fāṭimid caliph he theoretically served, thus ending more than 200 years of Fāṭimid rule in Egypt. When Nureddin died, Saladin succeeded him as head of the whole movement. When Saladin died in 1193, he had recaptured Jerusalem (1187) and begun the reunification of Egypt and Syria; his successors were known, after his patronymic, as the Ayyūbids. The efforts of a contemporary ʿAbbāsid caliph, an-Nāṣir, to revive the caliphate seem pale by comparison.

The Ayyūbids ruled in Egypt and Syria until 1250, when they were replaced first in Egypt and later in Syria by

Final expulsion of crusaders by Ayyūbids

the leaders of their own slave-soldier corps, the Mamlūks. It was they who expelled the remaining crusaders from Syria, subdued the remaining Nizārī Ismāʿīlīs there, and consolidated Ayyūbid holdings into a centralized state. That state became strong enough in its first decade to do what no other Muslim power could: in 1260 at ʿAyn Jālūt, south of Damascus, the Mamlūk army defeated the recently arrived Mongols and expelled them from Syria.

MONGOLS

The Mongols were pagan, horse-riding tribes of the northeastern steppes of Central Asia. In the early 13th century, under the leadership of Genghis Khan, they formed, led, and gave their name to a confederation of Turkic tribes that they channeled into a movement of global expansion, spreading east into China, north into Russia, and west into Islāmdom. Like other migratory peoples before them, Arabs, Berbers, and Turks, they had come to be involved in citied life through their role in the caravan trade. Unlike others, however, they did not convert to Islām before their arrival. Furthermore, they brought a greater hostility to sedentary civilization, a more ferocious military force, a more cumbersome material culture, a more complicated and hierarchical social structure, and a more coherent sense of tribal law. Their initial impact was physically more destructive than that of previous invaders, and their long-term impact perhaps more socially and politically creative.

First Mongol incursions. The first Mongol incursions into Islāmdom in 1220 were a response to a challenge from the Khwārezm-Shāh ʿAlāʾ ad-Dīn Muḥammad, the aggressive reigning leader of a dynasty formed in the Oxus Delta by a local governor who had rebelled against the Seljuq regime in Khorāsān. Under Genghis Khan's leadership, Mongol forces destroyed numerous cities in Transoxania and Khorāsān in an unprecedented display of terror and annihilation. By the time of Genghis Khan's death in 1227, his empire stretched from the Caspian Sea to the Sea of Japan. A later successor, Möngke, decided to extend the empire in two new directions. From the Mongol capital of Karakorum, he simultaneously dispatched Kublai Khan to southern China (where Islām subsequently began to expand inland) and Hülegü to Iran (1256). Hülegü had already received Sunnite ambassadors who encouraged him to destroy the Ismāʿīlī state at Alamūt; this he did and more, reaching Baghdad in 1258, where he terminated and replaced the caliphate. The ʿAbbāsid line continued, however, until 1517; the Mamlūk sultan Baybars I, shortly after his defeat of the Mongols, invited a member of the ʿAbbāsid house to "invest" him and to live in Cairo as spiritual head of all Muslims.

The Mongol regimes in Islāmdom quickly became rivals. The Il-Khans controlled the Tigris–Euphrates valley and Iran; the Chagatai dominated the Syrdarya and Oxus basins, the Kābul mountains, and eventually the Punjab; and the Golden Horde was concentrated in the Volga basin. The Il-Khans ruled in the territories where Islām was most firmly established. They patronized learning of all types and scholars from all parts of the vast Mongol empire, especially China. Evincing a special interest in nature, they built a major observatory in Azerbaijan. Just as enthusiastically as they had destroyed citied life, they now rebuilt it, relying as had all previous invaders of Iran on the administrative skills of indigenous Persian-speaking bureaucrats. The writings of one of these men, ʿAṭā Malek Joveynī, who was appointed Mongol governor in Baghdad in 1259, described the type of rule the Mongols sought to impose. It has been called the military patronage state because it involved a reciprocal relationship between the foreign tribal military conquerors and their subjects. The entire state was defined as a single mobile military force connected to the household of the monarch; with no fixed capital, it moved with the monarch. All non-Turkic state workers, bureaucratic or religious, even though not military specialists, were defined as part of the army (*asker*); the rest of the subject population, as the herds (*raʿīyah*). The leading tribal families could dispose of the wealth of the conquered populations as they wished, except that their natural superiority obligated them to reciprocate by patronizing whatever of excellence the cities could produce. What the Ghaznavids and Seljuqs had begun, the Mongols now accomplished. The self-confidence and superiority of the leading families were bolstered by a fairly elaborate set of tribal laws, inherited from Genghis Khan and known as the Yasa, which served to regulate personal status and criminal liability among the Mongol elite, as did the Sharīʿah among Muslims. In Il-Khanid hands, this dynastic law merely coexisted but did not compete with Sharīʿah; but in later Turkic regimes a reconciliation was achieved that extended the power of the rulers beyond the limitations of an autonomous Sharīʿah.

Military patronage system of Mongol rule

Conversion of Mongols to Islām. For a time the Il-Khans tolerated and patronized all religious persuasions, Sunnite, Shīʿite, Buddhist, Nestorian Christian, Jewish, and pagan. But in 1295 a Buddhist named Maḥmūd Ghāzān became Khan and declared himself Muslim, compelling other Mongol notables to follow suit. His patronage of Islāmicate learning fostered such brilliant writers as Rashīd ad-Dīn, the physician and scholar who authored one of the most famous Persian universal histories of all time. The Mongols, like other Islāmicate dynasties swept into power by a tribal confederation, were able to unify their domains for only a few generations. By the 1330s their rule had begun to be fragmented among myriad local leaders. Meanwhile, on both Mongol flanks, other Turkic Muslim powers were increasing in strength.

To the east the Delhi Sultanate of Turkic slave soldiers withstood Mongol pressure, benefited from the presence of scholars and administrators fleeing Mongol destruction, and gradually began to extend Muslim control south into India, a feat that was virtually accomplished under Muḥammad ibn Tughluq. Muslim Delhi was a culturally lively place that attracted a variety of unusual persons. Muḥammad ibn Tughluq himself was, like many later Indian Muslim rulers, well read in philosophy, science, and religion. Not possessing the kind of dynastic legitimacy the pastoralist Mongols had asserted, he tied his legitimacy to his support for the Sharīʿah, and he even sought to have himself invested by the ʿAbbāsid "caliph" whom the Mamlūks had taken to Cairo. His concern with the Sharīʿah coincided with the growing popularity of Ṣūfism, especially as represented by the massive Chishti *ṭarīqah*. Its most famous leader, Niẓām ad-Dīn Awliyāʾ, had been a spiritual adviser to many figures at court before Muḥammad ibn Tughluq came to the throne, as well as to individual Hindus and Muslims alike. In India, Ṣūfism, which inherently undermined communalism, was bringing members of different religious communities together in ways very rare in the more westerly parts of Islāmdom.

To the west, the similarly constituted Mamlūk state continued to resist Mongol expansion. Its sultans were chosen, on a nonhereditary basis, from among a group of freed slaves who acted as the leaders of the various slave corps. At the death of one sultan the various military corps would compete to see whose leader would become the next sultan. The leaders of the various slave corps formed an oligarchy that exercised control over the sultan. Although political instability was the frequent and natural result of such a system, cultural florescence did occur. The sultans actively encouraged trade and building, and Mamlūk Cairo became a place of splendour, filled with numerous architectural monuments. While the Persian language was becoming the language of administration and high culture over much of Islāmdom, Arabic alone continued to be cultivated in Mamlūk domains, to the benefit of a diversified intellectual life. Ibn an-Nafīs (died 1288), a physician, wrote about pulmonary circulation 300 years before it was "discovered" in Europe. For Mamlūk administrative personnel, al-Qalqashandī composed an encyclopaedia in which he surveyed not only local practice but also all the information that a cultivated administrator should know. Ibn Khallikān composed one of the most important Islāmicate biographical works, a dictionary of eminent men. Sharīʿah-minded studies were elaborated: the *ʿulamāʾ* worked out a political theory that tried to make sense of the sultanate, and they also explored the possibility of enlarging on the Sharīʿah by reference to *falsafah* and Ṣūfism.

Cairo under Mamlūk rule

However, in much the same way as ash-Shāfiʿī had responded in the 9th century to what he viewed as dangerous legal diversity, another great legal and religious reformer, Ibn Taymīyah, living in Mamlūk Damascus in the late 13th and early 14th century, cautioned against such extralegal practices and pursuits. He insisted that the Sharīʿah was complete in and of itself and could be adapted to every age by any *faqih* who could analogize according to the principle of human advantage (*maṣlaḥah*). A Ḥanbalī himself, Ibn Taymīyah became as popular as his school's founder, Aḥmad ibn Ḥanbal. Like him, Ibn Taymīyah attacked all practices that undermined what he felt to be the fundamentals of Islām, including all forms of Shīʿite thought as well as aspects of Jamāʿi-Sunnite piety (often influenced by the Ṣūfīs) that stressed knowledge of God over service to him. Most visible among such practices was the revering of saints' tombs, which was condoned by the Mamlūk authorities. Ibn Taymīyah's program and popularity so threatened the Mamlūk authorities that they put him in prison, where he died. His movement did not survive, but when his ideas surfaced, in the revolutionary movement of the Wahhābīyah in the late 18th century, their lingering power became dramatically evident.

Further west, the Rūm Seljuqs at Konya submitted to the Mongols in 1243 but survived intact. They continued to cultivate the Islāmicate arts, architecture in particular. The most famous Muslim ever to live at Konya, Jalāl ad-Dīn ar-Rūmī, had emigrated from eastern Iran with his father before the arrival of the Mongols. In Konya, Jalāl ad-Dīn, attracted to Ṣūfī activities, attached himself to the master Shams ad-Dīn. The poetry inspired by Jalāl ad-Dīn's association with Shams ad-Dīn is unparalleled in Persian literature. Its recitation, along with music and movement, was a key element in the devotional activities of Jalāl ad-Dīn's followers, who came to be organized into a Ṣūfī *ṭarīqah* named the Mevleviyah (Mawlawīyah) after their title of respect for him, Mevlana ("Our Master"). In his poetry Jalāl ad-Dīn explored all varieties of metaphors, including intoxication, to describe the ineffable ecstasy of union with God.

Poetry of Jalāl ad-Dīn ar-Rūmī

Ascent of the Ottoman Turks. It was not from the Rūm Seljuqs, however, that lasting Muslim power in Anatolia was to come, but rather from one of the warrior states on the Byzantine frontier. The successive waves of Turkic migrations had driven unrelated individuals and groups across central Islāmdom into Anatolia. Avoiding the Konya state, they gravitated toward an open frontier to the west, where they began to constitute themselves, often through fictitious kinship relationships, into quasi-tribal states that depended on raiding each other and Byzantine territory and shipping. One of these, the Osmanlıs, or Ottomans, named for their founder, Osman I (ruled 1281–1324), was located not on the coast, where raiding had its limits, but in Bithynia just facing Constantinople. In 1326 they won the town of Bursa and made it their first capital. From Anatolia they crossed over into Thrace in the service of rival factions at Constantinople, then began to occupy Byzantine territory, establishing their second capital at Edirne on the European side. Their sense of legitimacy was complex. They were militantly Muslim, bound by the *ghāzī* spirit, spurred on in their intolerance of local Christians by Greek converts and traveling Ṣūfīs who gravitated to their domains. At the same time, *ʿulamāʾ* from more settled Islāmic lands to the east encouraged them to abide by the Sharīʿah and tolerate the Christians as protected non-Muslims. The Ottomans also cast themselves as deputies of the Rūm Seljuqs, who were themselves originally "deputized" by the ʿAbbāsid caliph. Finally they claimed descent from the leading Oğuz Turk families, who were natural rulers over sedentary populations. Under Murad I (ruled *c.* 1360–89) the state began to downplay its warrior fervour in favour of more conventional Islāmicate administration. Instead of relying on volunteer warriors, Murad established a regular cavalry, which he supported with land assignments, as well as a specially trained infantry force called the "new troops," Janissaries, drawn from converted captives. Expanding first through western Anatolia and Thrace, the Ottomans under Bayezid I (ruled 1389–1403) turned their eyes toward eastern and southern Anatolia; just as they had incorporated the whole, they encountered a neo-Mongol conqueror expanding into Anatolia from the east who utterly defeated their entire army in a single campaign (1402).

Timur's efforts to restore Mongol power. Timur (Tamerlane) was a Turk, not a Mongol; but he aimed to restore Mongol power. He was born a Muslim in the Syrdarya valley and served local pagan Mongol warriors and finally the Chagatai heir-apparent; but he rebelled and made himself ruler in Khwārezm in 1380. He planned to restore Mongol supremacy under a thoroughly Islāmic program. He surpassed the Mongols in terror, constructing towers out of the heads of his victims. Having established himself in Iran, he moved first on India and then on Ottoman Anatolia and Mamlūk Syria; but before he could consolidate his realm, he died. His impact was twofold: his defeat of the Ottomans inspired a comeback that would produce one of the greatest Islāmicate empires of all time, and one of the Central Asian heirs to his tradition of conquest would found another great Islāmicate empire in India. These later empires managed to find the combination of Turkic and Islāmic legitimacy that could produce the stable centralized absolutism that had eluded all previous Turkic conquerors.

ARABS

When the Fāṭimids conquered Egypt in 969, they left a governor named Zīrī in the Maghrib. By 1041 the dynasty founded by Zīrī declared its independence from the Fāṭimids, but it too was challenged by breakaways such as the Zanātah in Morocco and the Ḥammādids in Algeria. Gradually the Zīrids were restricted to the eastern Maghrib. There they were invaded from Egypt by two Bedouin Arab tribes, the Banū Halīl and the Banū Sulaym, at the instigation (1052) of the Fāṭimid ruler in Cairo. This mass migration of warriors as well as wives and children is known as the Hilālian invasion. Though initially disruptive, the Hilālian invasion had an important cultural impact; it resulted in a much greater spread of the Arabic language than had occurred in the 7th century and inaugurated the real Arabization of the Maghrib.

Hilālian invasion of the eastern Maghrib

BERBERS

When the Arab conquerors arrived in the Maghrib in the 7th century, the indigenous peoples they met were the Berbers, a group of predominantly but not entirely migratory tribes who spoke a recognizably common Hamito-Semitic language with significant dialectal variations. Berber tribes could be found from present-day Morocco to present-day Algeria, and from the Mediterranean to the Sahara. As among the Arabs, small tribal groupings of Berbers occasionally formed short-lived confederations or became involved in caravan trade. No previous conqueror had tried to assimilate the Berbers, but the Arabs quickly converted them and enlisted their aid in further conquests. Without their help, for example, Andalusia could never have been incorporated into the Islāmicate state. At first only Berbers nearer the coast were involved, but by the 11th century Muslim affiliation had begun to spread far into the Sahara.

The Ṣanhājah confederation. One particular western Saharan Berber confederation, the Ṣanhājah, was responsible for the first Berber-directed effort to control the Maghrib. The Ṣanhājah were camel herders who traded mined salt for gold with the black kingdoms of the south. By the 11th century their power in the western Sahara was being threatened by expansion both from other Berber tribes, centred at Sijilmassa, and from the Soninke state at Ghana to the south, which had actually captured their capital of Audaghost in 990. The subsequent revival of their fortunes parallels Muḥammad's revitalization of the Arabs 500 years earlier, in that Muslim ideology reinforced their efforts to unify several smaller groups. The Ṣanhājah had been in contact with Islām since the 9th century, but their distance from major centres of Muslim life had kept their knowledge of the faith minimal. In 1035, however, Yaḥyā ibn Ibrāhīm, a chief from one of their tribes, the Gudālah, went on *ḥajj*. For the Maghribi pilgrim, the cultural impact of the *ḥajj* was experienced not only in Mecca and

Yaḥyā ibn Ibrāhīm's conversion to Islām

Medina but also on the many stops along the 3,000-mile overland route. When Yaḥyā returned, he was accompanied by a teacher from Nafis (in present-day Libya), ʿAbd Allāh ibn Yasīn, who would instruct the Berbers in Islām as teachers under ʿUmar I had instructed the Arab fighters in the first Muslim garrisons. Having met with little initial success, the two are said to have retired to a *ribāṭ*, a fortified place of seclusion, perhaps as far south as an island in the Sénégal River, to pursue a purer religious life. The followers they attracted to that *ribāṭ* were known, by derivation, as *al-murābiṭūn* ("the people of the retreat"); the dynasty they founded came to be known by the same name, or Almoravids in its Anglicized form. In 1042 Ibn Yasīn declared a *jihād* against the Ṣanhājah tribes, including his own, as people who had embraced Islām but then failed to practice it properly. By his death in 1059, the Ṣanhājah confederation had been restored under an Islāmic ideology; and the conquest of Morocco, which lacked strong leadership, was under way.

The Almoravid dynasty. Ibn Yasīn's spiritual role was taken by a consultative body of *ʿulamāʾ*. His successor as military commander was Abū Bakr ibn ʿUmar. While pursuing the campaign against Morocco, Abū Bakr had to go south, leaving his cousin Yūsuf ibn Tāshufīn as his deputy. When Abū Bakr tried to return, Ibn Tāshufīn turned him back to the south, where he remained until his death in 1087. Under Ibn Tāshufīn's leadership, by 1082, Almoravid control extended as far as Algiers. In 1086 Ibn Tāshufīn responded to a request for help from the Andalusian party kings, unable to defend themselves against the Christian kingdoms in the north, such as Castile. By 1110 all Muslim states in Andalusia had come under Almoravid control.

Like most other Jamāʿi-Sunnite rulers of his time, Ibn Tāshufīn had himself "appointed" deputy by the caliph in Baghdad. He also based his authority on the claim to bring correct Islām to peoples who had strayed from it. For him "correct" Islām meant the Sharīʿah as developed by the Mālikī *faqih*s, who played a key role in the Almoravid state by working out the application of the Sharīʿah to everyday problems. Like their contemporaries elsewhere, they received stipends from the government, sat in the ruler's council, went on campaign with him, and gave him recommendations (*fatwa*s) on important decisions. This was an approach to Islām far more current than the one it had replaced, but still out of touch with the liveliest intellectual developments. During the next phase of Berber activism, newer trends from the east reached the Maghrib.

Ibn Tūmart's revolt

A second major Berber movement originated in a revolt begun against Almoravid rule in 1125 by Ibn Tūmart, a settled Maṣmūdah Berber from the Atlas Mountains. Like Ibn Yasīn, Ibn Tūmart had been inspired by the *ḥajj*, which he used as an opportunity to study in Baghdad, Cairo, and Jerusalem, acquainting himself with all current schools of Islāmic thought and becoming a disciple of the ideas of the recently deceased al-Ghazālī. Emulating his social activism, Ibn Tūmart was inspired to act on the familiar Muslim dictum, "Command the good and forbid the reprehensible." His early attempts took two forms, disputations with the scholars of the Almoravid court and public chastisement of Muslims who in his view contradicted the rules of Islām; he went so far as to throw the Almoravid ruler's sister off her horse because she was unveiled in public. His activities aroused hostility and he fled to the safety of his own people. There, like Muḥammad, he grew from teacher of a personal following to leader of a social movement.

Like many subsequent reformers, especially in Africa and other outlying Muslim lands, Ibn Tūmart used Muḥammad's career as a model. He interpreted the Prophet's rejection and retreat as an emigration (*hijrah*) that enabled him to build a community, and he divided his followers into *muhājirūn* ("fellow emigrants") and *anṣār* ("helpers"). He preached the idea of surrender to God to a people who had strayed from it. Thus could Muḥammad's ability to bring about radical change through renewal be invoked without actually claiming the prophethood that he had sealed forever. Ibn Tūmart further based his legitimacy on his claim to be a *sharīf* (descendant of Muḥammad) and the *mahdī*, not in the Shīʿite sense but in the more general sense of a human sent to restore pure faith. In his view Almoravid students of legal knowledge were so concerned with pursuing the technicalities of the law that they had lost the purifying fervour of their own founder, Ibn Yasīn. They even failed to maintain proper Muslim behaviour, be it the veiling of women in public or the condemning of the use of wine, musical instruments, and other unacceptable, if not strictly illegal, forms of pleasure. Like many Muslim revitalizers before and since, Ibn Tūmart decried the way in which the law had taken on a life of its own, and he called upon Muslims to rely on the original and only reliable sources, the Qurʾān and *ḥadīth.* Although he opposed irresponsible rationalism in the law, in matters of theological discourse he leaned toward the limited rationalism of the Ashʿarite school, which was becoming so popular in the eastern Muslim lands. Like the Ashʿarites, he viewed the unity of God as one of Islām's fundamentals and denounced any reading of the Qurʾān that led to anthropomorphism. Because he focused on attesting the unity of God (*tawḥīd*), he called his followers al-Muwaḥḥidūn (Almohads), "those who attest the unity of God." Ibn Tūmart's movement signified the degree to which Maghribis could participate in the intellectual life of Islāmdom as a whole; but his need to use Berber for his many followers who did not know Arabic also illustrates the limits of interregional discourse.

The Almohad dynasty. By 1147, 17 years after Ibn Tūmart's death, Almohads had replaced Almoravids in all their Maghribi and Andalusian territories. In Andalusia their arrival slowed the progress of the Christian Reconquista. There, as in the Maghrib, arts and letters were encouraged: an example is an important movement of *falsafah* that included Ibn Ṭufayl, Ibn al-ʿArabī, and Ibn Rushd (Latin Averroës), the Andalusian *qāḍī* and physician whose interpretations of Aristotle became so important for medieval European Christianity. During the late Almohad period in Andalusia the intercommunal nature of Islāmicate civilization became especially noticeable in the work of non-Muslim thinkers, such as Moses Maimonides, who participated in trends outside their own communities even at the expense of criticism from within. By the early 13th century, Almohad power began to decline; a defeat in 1212 at Las Navas de Tolosa by the Christian kings of the north forced a retreat to the Maghrib. But the impact of Almohad cultural patronage on Andalusia long outlasted Almohad political power; successor dynasties in surviving Muslim states were responsible for some of the highest achievements of Andalusian Muslims, among them the Alhambra palace in Granada. Furthermore, the 400-year southward movement of the Christian–Muslim frontier resulted, ironically, in some of the most intense Christian–Muslim interaction in Andalusian history. The Cid could fight for both sides; Muslims, as Mudejars, could live under Christian rule and contribute to its culture; Jews could translate Arabic and Hebrew texts into Castilian. Almohads were replaced in the Maghrib as well, through a revolt by their own governors: the Ḥafṣids in Tunis and the Marīnid Berber dynasty in Fès. There too, however, Almohad influence outlasted their political presence: both towns became centres, in distinctively Maghribi form, of Islāmicate culture and Islāmic piety.

Continued spread of Islāmic influence. As the Maghrib became firmly and distinctively Muslim, Islām moved south. The spread of Muslim identity into the Sahara and the involvement of Muslim peoples, especially the Tuareg, in trans-Saharan trade provided several natural channels of influence. By the time of the Marīnids, Ḥafṣids, and Mamlūks, several major trade routes had established crisscrossing lines of communication: from Cairo to Timbuktu, from Tripoli to Bornu and Lake Chad, from Tunis to Timbuktu at the bend of the Niger River, and from Fès and Tafilalt through major Saharan entrepôts into Ghana and Mali. The rise at Timbuktu of Mali, the first great western Sudanic empire with a Muslim ruler, attested the growing incorporation of sub-Saharan Africa into the North African orbit. The reign of Mansa Mūsā, who even went on pilgrimage, demonstrated the influence of Islām on at least the upper echelons of African society.

Travel writings of Ibn Baṭṭūṭah

The best picture of Islāmdom in the 14th century appears in the work of a remarkable Maghribi *qāḍī* and traveler, Ibn Baṭṭūṭah (1304–1368/77). In 1325, the year that Mansa Mūsā went on pilgrimage, Ibn Baṭṭūṭah also left for Mecca, from his hometown of Tangiers. He was away for almost 30 years, visiting most of Islāmdom, including Andalusia, all of the Maghrib, Mali, Syria, Arabia, Iran, India, the Maldive Islands, and, he claimed, China. He described the unity within diversity that was one of Islāmdom's most prominent features. Although local customs often seemed at variance with his notion of pure Islāmic practice, he felt at home everywhere. Despite the divisions that had occurred during Islām's 700-year history, a Muslim could attend the Friday worship session in any Muslim town in the world and feel comfortable, a claim that is difficult if not impossible to make for any other major religious tradition at any time in its history. By the time of Ibn Baṭṭūṭah's death, Islāmdom comprised the most far-flung yet interconnected set of societies in the world. As one author has pointed out, Thomas Aquinas (*c.* 1224–74) might have been read from Spain to Hungary and from Sicily to Norway; but Ibn al-ʿArabī (1165–1240) was read from Spain to Sumatra and from the Swahili coast to Kazan on the Volga River. By the end of the period of migration and renewal, Islām had begun to spread not only into sub-Saharan Africa but also into the southern seas with the establishment of a Muslim presence in the Straits of Malacca. Conversion to Islām across its newer frontiers was at first limited to a small elite, who supplemented local religious practices with Muslim ones. Islām could offer not only a unifying religious system but also social techniques, including alphabetic literacy, a legal system applicable to daily life, a set of administrative institutions, and a body of science and technology—all capable of enhancing the power of ruling elements and of tying them into a vast and lucrative trading network.

The period of migration and renewal exposed both the potentiality and the limitations of government by tribal peoples. This great problem of Islāmicate history received its most sophisticated analysis from a Maghribi Muslim named Ibn Khaldūn (1332–1406), a contemporary of Petrarch. His family had migrated from Andalusia to the Maghrib, and he himself was born in Ḥafṣid territory. He was both a *faylasūf* and a *qāḍī*, a combination more common in Andalusia and the Maghrib than anywhere else in Islāmdom. His *falsafah* was activist; he strove to use his political wisdom to the benefit of one of the actual rulers of the day. To this end he moved from one court to another before becoming disillusioned and retiring to Mamlūk Cairo as a *qāḍī*. His life thus demonstrated the importance and the constraints of royal patronage as a stimulant to intellectual creativity. In his *Muqaddimah* (the introduction to his multivolume world history) he used his training in *falsafah* to discern patterns in history. Transcending the critiques of historical method made by historians of the Būyid period, such as al-Masʿūdī, Miskawayh, and as-Suli, Ibn Khaldūn established careful standards of evidence. Whereas Muslim historians conventionally subscribed to the view that God passed sovereignty and hegemony (*dawlah*) from one dynasty to another through his divine wisdom, Ibn Khaldūn explained it in terms of a cycle of natural and inevitable stages. By his day it had become apparent that tribally organized migratory peoples, so favoured by much of the ecology of the Maghrib and the Nile-to-Oxus region, could easily acquire military superiority over settled peoples if they could capitalize on the inherently stronger group feeling (*ʿaṣabīyyah*) that kinship provides. Once in power, according to Ibn Khaldūn, conquering groups pass through a phase in which a small number of "builders" among them bring renewed vitality to their conquered lands. As the family disperses itself among sedentary peoples and ceases to live the hard life of migration, it becomes soft from the prosperity it has brought and begins to degenerate. Then internal rivalries and jealousies force one member of the family to become a king who must rely on mercenary troops and undermine his own prosperity by paying for them. In the end, the ruling dynasty falls prey to a new tribal group with fresh group feeling. Thus did Ibn Khaldūn call attention to the unavoidable instability of all premodern Muslim dynasties, caused by their lack of the regularized patterns of succession that were beginning to develop in European dynasties.

Consolidation and expansion (1405–1683)

Shifting power of Muslims and Europeans

After the death of Timur in 1405, power began to shift from migrating peoples to sedentary populations living in large centralized empires. After about 1683, when the last Ottoman campaign against Vienna failed, the great empires for which this period is so famous began to shrink and weaken, just as western Europeans first began to show their potential for worldwide expansion and domination. When the period began, Muslim lands had begun to recover from the devastating effects of the Black Death (1346–48), and many were prospering. Muslims had the best opportunity in history to unite the settled world, but by the end of the period, they had been replaced by Europeans as the leading contenders for this role. Muslims were now forced into direct and repeated contact with Europeans, through armed hostilities as well as through commercial interactions; and often the Europeans competed well. Yet Muslim power was so extensive, and the western Europeans such an unexpected source of competition, that Muslims were able to realize that their situation had changed only after they no longer had the strength to resist. Furthermore, the existence of several strong competitive Muslim states militated against a united response to the Europeans and could even encourage some Muslims to align themselves with the European enemies of others.

In this period, long after Islāmdom was once thought to have peaked, centralized absolutism reached its height, aided in part by the exploitation of gunpowder warfare and in part by new ways to fuse spiritual and military authority. Never before had Islāmicate ideals and institutions better demonstrated their ability to encourage political centralization, or to support a Muslim style of life where there was no organized state, be it in areas where Islām had been long established, or in areas where it was newly arrived. The major states of this period impressed contemporary Europeans; in them some of the greatest Islāmicate artistic achievements were made. In this period Muslims formed the cultural patterns that they brought into modern times, and adherence to Islām expanded to approximately its current distribution. As adherence to Islām expanded, far-flung cultural regions began to take on a life of their own. The unity of several of these regions was expressed through empire—the Ottomans in southeastern Europe, Anatolia, the eastern Maghrib, Egypt, and Syria; the Ṣafavids in Iran and Iraq; the Indo-Timurids (Mughals) in India. In these empires, Sunnite and Shīʿite became identities on a much larger scale than ever before, expressing competition between large populations; simultaneously Shīʿism acquired a permanent base from which to generate international opposition. Elsewhere, less formal and often commercial ties bound Muslims from distant locales; growing commercial and political links between Morocco and the western Sudan produced a trans-Saharan Maghribi Islām; Egyptian Islām influenced the central and eastern Sudan; and steady contacts between East Africa, South Arabia, southern Iran, southwest India, and the southern seas promoted a recognizable Indian Ocean Islām, with Persian as its lingua franca. In fact, Persian became the closest yet to an international language; but the expansion and naturalization of Islām also fostered a number of local languages into vehicles for Islāmicate administration and high culture—Ottoman, Chagatai, Swahili, Urdu, and Malay. Everywhere Muslims were confronting adherents of other religions, and new converts often practiced Islām without abandoning their previous practices. The various ways in which Muslims responded to religious syncretism and plurality continue to be elaborated to the present day.

Expansion of Muslim presence

This was a period of major realignments and expansion. The extent of Muslim presence in the Eastern Hemisphere in the early 15th century was easily discernible, but only with difficulty could one have imagined that it could soon produce three of the greatest empires in world history. From the Atlantic to the Pacific, from the Balkans to Sumatra, Muslim rulers presided over relatively small kingdoms; but nowhere could the emergence of a world-class dynasty be predicted. In Andalusia only one Muslim state, Granada, remained to resist Christian domination of the Iberian Peninsula. The Maghrib, isolated between an almost all-Christian Iberia and an eastward-looking Mamlūk Egypt and Syria, was divided between the Marīnids and Ḥafṣids. Where the Sahara shades off into the Sudanic belt, the empire of Mali at Gao was ruled by a Muslim and included several Saharan "port" cities, such as Timbuktu, that were centres of Muslim learning. On the Swahili coast, oriented as always more toward the Indian Ocean than toward its own hinterland, several small Muslim polities centred on key ports such as Kilwa. In western Anatolia and the Balkan Peninsula the Ottoman state under Sultan Meḥmed I was recovering from its defeat by Timur. Iraq and western Iran were the domains of Turkic tribal dynasties known as the Black Sheep (Kara Koyunlu) and the White Sheep (Ak Koyunlu); they shared a border in Iran with myriad princelings of the Timurid line; and the neo-Mongol, neo-Timurid Uzbek state ruled in Transoxania. North of the Caspian, several Muslim khanates ruled as far north as Moscow and Kazan. In India, even though Muslims constituted a minority, they were beginning to assert their power everywhere except the south, which was ruled by Vijayanagar. In Islāmdom's far southeast, the Muslim state of Samudra held sway in Sumatra, and the rulers of the Moluccas had recently converted to Islām and begun to expand into the southern Malay Peninsula. Even where no organized state existed, as in the outer reaches of Central Asia and into southern China, scattered small Muslim communities persisted, often centred on oases. By the end of this period, Islāmdom's borders had retreated only in Russia and Iberia; but these losses were more than compensated by continuing expansion in Europe, Africa, Central Asia, and South and Southeast Asia. Almost everywhere this plethora of states had undergone realignment and consolidation, based on experimentation with forms of legitimation and structure.

OTTOMANS

The conquest of Constantinople (1453)

Continuation of Ottoman rule. After the Ottoman state's devastating defeat by Timur, its leaders had to retain the vitality of the warrior spirit (without its unruliness and intolerance) and the validation of the Sharīʿah (without its confining independence). In 1453, Meḥmed II, the Conqueror, fulfilled the warrior ideal by conquering Constantinople (soon to be known as Istanbul), putting an end to the Byzantine Empire, and subjugating the local Christian and Jewish populations. Even by then, however, a new form of legitimation was taking shape. The Ottomans continued to wage war against Christians on the frontier and to levy and convert (through the *devşirme*) young male Christians to serve in the sultan's household and army; but warriors were being pensioned off with land grants and replaced by troops more beholden to the sultan. Except for those forcibly converted, the rest of the non-Muslim population was protected for payment according to the Sharīʿah and the preference of the *ulema* (the Turkish spelling of *ʿulamāʾ*), and organized into self-governing communities known as *millet*s. Furthermore, the sultans began to claim the caliphate because they met two of its traditional qualifications: they ruled justly, in principle according to the Sharīʿah, and they defended and extended the frontiers, as in their conquest of Mamlūk Egypt, Syria, and the holy cities in 1516–17. Meanwhile they began to undercut the traditional oppositional stance of the *ulema* by building on Seljuq and Mongol practice in three ways: they promoted state-supported training of *ulema*; they defined and paid holders of religious offices as part of the military; and they aggressively asserted the validity of dynastic law alongside Sharīʿah. Simultaneously, they emphasized their inheritance of Byzantine legitimacy by transforming Byzantine symbols, such as Hagia Sophia (Church of the Divine Wisdom), into symbols for Islām; and by favouring their empire's European part, called, significantly, Rūm.

Four-part administrative structure

Reign of Süleyman I. The classical Ottoman system crystallized during the reign of Süleyman I, the Lawgiver (ruled 1520–66). He also pushed the empire's borders almost to their furthest limits, to the walls of Vienna in the northwest, throughout the Maghrib up to Morocco in the southwest, into Iraq to the east, and to the Yemen in the southeast. During Süleyman's reign the Ottomans even sent an expedition into the southern seas to help Aceh against the Portuguese colonizers. In theory, Süleyman presided over a balanced four-part structure: the palace household, which contained all of the sultan's wives, concubines, children, and servants; the bureaucracy (chancery and treasury); the armed forces; and the religious establishment. Important positions in the army and bureaucracy went to the cream of the *devşirme,* Christian youths converted to Islām and put through special training at the capital to be the sultan's personal "slaves." *Ulema* who acquired government posts had undergone systematic training at the major *medrese*s (*madrasah*s) and so in the Ottoman state were more integrated than were their counterparts in other states; yet they were free-born Muslims, not brought into the system as slaves of the sultan. The ruling class communicated in a language developed for their use only, Ottoman, which combined Turkic syntax with largely Arabic and Persian vocabulary. It was in this new language that so many important figures demonstrated the range and sophistication of Ottoman interests, such as the historian Mustafa Naima, the encyclopaedist Kâtip Çelebi, and the traveler Evliya Çelebi. The splendour of the Ottoman capital owed not a little to Süleyman's chief architect, the Greek *devşirme* recruit Sinan, who transformed the city's skyline with magnificent mosques and *medreses.*

The extent of Ottoman administration. Even in North Africa and the Fertile Crescent, where Ottoman rule was indirect, the effect of its administration, especially its land surveys and *millet* and tax systems, could be felt; remnants of the Ottoman system continue to play a role in the political life of modern states such as Israel and Lebanon, despite the fact that Ottoman control had already begun to relax by the first quarter of the 17th century. By then control of the state treasury was passing, through land grants, into the hands of local *aʿyān,* and they gradually became the real rulers, serving local rather than imperial interests. Meanwhile discontinuance of the *devşirme* and the rise of hereditary succession to imperial offices shut off new sources of vitality. Monarchs, confined to the palace during their youth, became weaker and participated less in military affairs and government councils. As early as 1630, Sultan Murad IV was presented by one of his advisers with a memorandum explaining the causes of the perceived decline and urging a restoration of the system as it had existed under Süleyman. Murad IV tried to restore Ottoman efficiency and central control, and his efforts were continued by subsequent sultans aided by a talented family of ministers known as the Köprülüs. However, during a war with Austria and Poland from 1682 to 1699, in which a major attack on Vienna failed (1683), the Ottomans suffered their first serious losses to an enemy and exposed the weakness of their system to their European neighbours. They signed two treaties, at Carlowitz in 1699 and at Passarowitz in 1718, that confirmed their losses in southeastern Europe, signified their inferiority to the Habsburg coalition, and established the defensive posture they would maintain into the 20th century.

ṢAFAVIDS

The Ṣafavid state began not from a band of *ghāzī* warriors but from a local Ṣūfī *ṭarīqah* of Ardabīl in Azerbaijan. The *ṭarīqah* was named after its founder, Shaykh Ṣafī od-Dīn (1252/53–1334), a local holy man. As for many *ṭarīqah*s and other voluntary associations, Sunnite and Shīʿite alike, affection for the family of ʿAlī was a channel for popular support. During the 15th century Shaykh Ṣafī's successors transformed their local *ṭarīqah* into an interregional movement by translating ʿAlid loyalism into full-fledged Imāmi Shīʿism. By asserting that they were the Ṣūfī "perfect men" of their time as well as descendants and representatives of the last *imām,* they strengthened the support of their Turkic tribal disciples (known as the Kizilbash, or "Red

Heads," because of their symbolic 12-fold red headgear). They also attracted support outside Iran, especially in eastern Anatolia (where the anti-Ottoman Imāmi Bekṭāshī *ṭarīqah* was strong), in Syria, the Caucasus, and Transoxania. The ability of the Iranian Shīʿite state to serve as a source of widespread local opposition outside of Iran was again to become dramatically apparent many years later, with the rise of the ayatollah Ruhollah Khomeini's Islāmic Republic in the late 1970s.

Proclamation of Ismāʿīl I as shah

Expansion in Iran and beyond. By 1501 the Ṣafavids were able to defeat the Ak Koyunlu rulers of northern Iran, whereupon their teenage leader Ismāʿīl I (ruled 1501–24) had himself proclaimed shah, using that pre-Islāmic title for the first time in almost 900 years and thereby invoking the glory of ancient Iran. The Ṣafavids thus asserted a multivalent legitimacy that flew in the face of Ottoman claims to have restored caliphal authority for all Muslims. Eventually, irritant became threat: by 1510, when Ismāʿīl had conquered all of Iran (to approximately its present frontiers) as well as the Fertile Crescent, he began pushing against the Uzbeks in the east and the Ottomans in the west, both of whom already suffered from significant Shīʿite opposition that could easily be aroused by Ṣafavid successes. Having to fight on two fronts was the most difficult military problem any Muslim empire could face. According to the persisting Mongol pattern, the army was a single force attached to the household of the ruler and moving with him at all times; so the size of an area under effective central control was limited to the farthest points that could be reached in a single campaign season. After dealing with his eastern front, Ismāʿīl turned west. At Chāldirān (1514) in northwestern Iraq, having refused to use gunpowder weapons, Ismāʿīl suffered the kind of defeat at Ottoman hands that the Ottomans had suffered from Timur. Yet through the war of words waged in a body of correspondence between Shah Ismāʿīl and the Ottoman sultan Selim I, and through the many invasions from both fronts that occurred during the next 60 years, the Ṣafavid state survived and prospered. Still living off its position at the crossroads of the trans-Asian trade that had supported all previous empires in Iraq and Iran, it was not yet undermined by the gradual emergence of more significant sea routes to the south.

The first requirement for the survival of the Ṣafavid state was the conversion of its predominantly Jamāʿī-Sunnite population to Imāmi Shīʿism. This was accomplished by a government-run effort supervised by the state-appointed leader of the religious community, the *ṣadr.* Gradually forms of piety emerged that were specific to Ṣafavid Shīʿism; they centred on pilgrimage to key sites connected with the *imām*s, as well as on the annual remembering and reenacting of the key event in Shīʿite history, the caliph Yazīd I's destruction of Imām al-Ḥusayn at Karbalāʾ on the 10th of Muḥarram, AH 61 (680 CE). The 10th of Muḥarram, or ʿĀshūrāʾ, already marked throughout Islāmdom with fasting, became for Iranian Shīʿites the centre of the religious calendar. The first 10 days of Muḥarram became a period of communal mourning, during which the pious imposed suffering on themselves to identify with their martyrs of old, listened to sermons, and recited appropriate elegiac poetry. In later Ṣafavid times the name for this mourning, *taʿzīyeh,* also came to be applied to passion plays performed to reenact events surrounding al-Ḥusayn's martyrdom. Through the depths of their empathetic suffering, Shīʿites could help to overturn the injustice of al-Ḥusayn's martyrdom at the end of time, when all wrongs would be righted, all wrongdoers punished, and all true followers of the *imām*s rewarded.

Shah ʿAbbās I. The state also survived because Ismāʿīl's successors moved, like the Ottomans, toward a type of legitimation different from the one that had brought them to power. This development began in the reign of Ṭahmāsp (1524–76) and culminated in the reign of the greatest Ṣafavid shah, ʿAbbās I (ruled 1588–1629). Since Ismāʿīl's time, the tribes had begun to lose faith in the Ṣafavid monarch as spiritual leader; now ʿAbbās appealed for support more as absolute monarch and less as the charismatic Ṣufi master or incarnated *imām.* At the same time he freed himself from his unruly tribal *amīr*s by depending more and more on a paid army of converted Circassian, Georgian, and Armenian Christian captives. Meanwhile he continued to rely on a large bureaucracy headed by a chief minister with limited responsibilities; but, unlike his Ottoman contemporaries, he distanced members of the religious community from state involvement while allowing them an independent source of support in their administration of the *waqf* system. Because the Shīʿite *ʿulamāʾ* had a tradition of independence that made them resist incorporation into the military "household" of the shah, ʿAbbās' policies were probably not unpopular; but they eventually undermined his state's legitimacy. By the end of the period under discussion, it was the religious leaders, the *mujtahid*s, who would claim to be the spokesmen for the hidden *imām.* Having shared the ideals of the military patronage state, the Ottoman state became more firmly militarized and religious, as the Ṣafavid became more civilianized and secular. The long-term consequences of this breach between government and the religious institution were extensive, culminating in the establishment of the Islāmic Republic of Iran in 1978.

Ṣafavid capital at Isfahan

ʿAbbās expressed his new role by moving his capital in about 1597–98 to Isfahan in Fārs, the central province of the ancient pre-Islāmic Iranian empires and symbolically more Persian than Turkic. Isfahan, favoured by a high and scenic setting, became one of the most beautiful cities in the world, leading its boosters to say, "Isfahan is half the world." It came to contain, often thanks to royal patronage, myriad palaces, gardens, parks, mosques, *medrese*s, caravansaries, workshops, and public baths. Many of these still stand, including the famed Masjed-e Shah, a mosque that shares the great central mall with an enormous covered bazaar and many other structures. It was here that ʿAbbās received diplomatic and commercial visits from Europeans, including a Carmelite mission from Pope Clement XIII (1604) and the adventuring Sherley brothers from Elizabethan England. Just as his visitors hoped to use him to their own advantage, ʿAbbās hoped to use them to his, as sources of firearms and military technology, or as pawns in his economic warfare against the Ottomans, in which he was willing to seek help from apparently anyone, including the Russians, Portuguese, and Habsburgs.

Under Ṣafavid rule, Iran in the 16th and 17th centuries became the centre of a major cultural flowering expressed through the Persian language and through the visual arts. This flowering extended to Ṣafavid neighbour states as well—Ottomans, Uzbeks, and Indo-Timurids. Like other Shīʿite dynasties before them, the Ṣafavids encouraged the development of *falsafah* as a companion to Shīʿite esotericism and cosmology. Two major thinkers, Mīr Dāmād and his disciple Mullā Ṣadrā, members of the Ishrāqī, or illuminationist, school, explored the realm of images or symbolic imagination as a way to understand issues of human meaningfulness. The Ṣafavid period was also important for the development of Shīʿite Sharīʿah-minded studies, and it produced a major historian, Iskandar Beg Munshī, chronicler of ʿAbbās' reign.

Decline of central authority. None of ʿAbbās' successors was his equal, though his state, ever weaker, survived for a century. The last effective shah, Ḥusayn I (1694–1722), could defend himself neither from tribal raiding in the capital nor from interfering *mujtahid*s led by Mohammad Bāqir Majlisī (whose writings later would be important in the Islāmic Republic of Iran). In 1722, when Maḥmūd of Qandahār led an Afghan tribal raid into Iran from the east, he easily took Isfahan and destroyed what was left of central authority.

INDO-TIMURIDS (MUGHALS)

Foundation by Bābur. Although the Mongol-Timurid legacy influenced the Ottoman and Ṣafavid states, it had its most direct impact on Bābur (1483–1530), the adventurer's adventurer and founder of the third major empire of the period. Bābur's father, ʿUmar Shaykh Mīrzā (died 1494) of Fergana, was one among many Timurid "princes" who continued to rule small pieces of the lands their great ancestor had conquered. After his father's death the 11-year-old Bābur, who claimed descent not only from

Timur but also from Genghis Khan (on his mother's side), quickly faced one of the harshest realities of his time and place—too many princes for too few kingdoms. In his youth he dreamed of capturing Samarkand as a base for reconstructing Timur's empire. For a year after the Ṣafavid defeat of the Uzbek Muḥammad Shaybānī Khān, Bābur and his Chagatai followers did hold Samarkand, as Ṣafavid vassals; but when the Ṣafavids were in turn defeated, Bābur lost not only Samarkand but his native Fergana as well. He was forced to retreat to Kābul, which he had occupied in 1504. From there he never restored Timur's empire; rather, barred from moving north or west, he took the Timurid legacy south, to a land on which Timur had made only the slightest impression.

Victory in India

When Bābur turned toward northern India, it was ruled from Delhi by the Lodī sultans, one of many local Turkic dynasties scattered through the subcontinent. In 1526 at Pānīpat, Bābur met and defeated the much larger Lodī army. In his victory he was aided, like the Ottomans at Chāldirān, by his artillery. By his death just four years later, he had laid the foundation for a remarkable empire, known most commonly as the Mughal (*i.e.,* Mongol). It is more properly called Indo-Timurid because the Chagatai Turks were distinct from the surviving Mongols of the time and because Bābur and his successors acknowledge Timur as the founder of their power.

Bābur is also remembered for his memoirs, the *Bābur-nāmeh.* Written in Chagatai, then an emerging Islāmicate literary language, his work gives a lively and compelling account of the wide range of interests, tastes, and sensibilities that made him so much a counterpart of his contemporary, the Italian Niccolò Machiavelli (1469–1527).

Reign of Akbar. Süleyman's and ʿAbbās' counterpart in the Indo-Timurid dynasty was their contemporary, Akbar (ruled 1556–1605), the grandson of Bābur. At the time of his death, he ruled all of present-day India north of the Deccan Plateau and Gondwana, and more: one diagonal of his empire extended from the Hindu Kush to the Bay of Bengal; the other, from the Himalayas to the Arabian Sea. Like its contemporaries to the west, particularly the Ottomans, this state endured because of a regularized and equitable tax system that provided the central treasury with funds to support the ruler's extensive building projects as well as his *manṣabdār*s, the military and bureaucratic officers of the imperial service. For these key servants, Akbar, again like his counterparts to the west, relied largely on foreigners who were trained especially for his service. Like the Janissaries, the *manṣabdār*s were not supposed to inherit their offices, and, although they were assigned lands to supervise, they themselves were paid through the central treasury to assure their loyalty to the interests of the ruler.

Akbar's centralized state

Although Akbar's empire was, like Süleyman's and ʿAbbās', a variation on the theme of the military patronage state, his situation, and consequently many of his problems, differed from theirs in important ways. Islām was much more recently established in most of his empire than in either of the other two, and Muslims were not in the majority. Although the other two states were not religiously or ethnically homogeneous, the extent of their internal diversity could not compare with Akbar's, where Muslims and non-Muslims of every stripe alternately coexisted and came into conflict—Jacobites (members of the Monophysite Syrian church), Ṣūfīs, Ismāʿīlī Shīʿites, Zoroastrians, Jains, Jesuits, Jews, and Hindus. Consequently, Akbar was forced even more than the Ottomans to confront and address the issue of religious plurality. The option of aggressive conversion was virtually impossible in such a vast area, as was any version of the Ottoman *millet* system in a setting in which hundreds if not thousands of *millet*s could be defined.

In some ways Akbar faced, in exaggerated form, the situation that the Arab Muslims faced when they were a minority in the Nile-to-Oxus region in the 7th–9th century. Granting protected status to non-Muslims, even those who were not really "peoples of the book" in the original sense, with an organized religion of their own, was legally and administratively justifiable; but unless they could be kept from interacting too much with the Muslim population, Islām itself could be affected. The power of Ṣūfī *ṭarīqah*s like the influential Chishtis, and of the Hindu mystical movement of Gurū Nānak, were already promoting intercommunal interaction and cross-fertilization. Akbar's response was different from that of the ʿAbbāsid caliph al-Mahdi. Instead of institutionalizing intolerance of non-Muslim influences, and instead of hardening communal lines, Akbar banned intolerance and even the special tax on non-Muslims. To keep the *ʿulamāʾ* from objecting, he tried, for different reasons than had the Ottomans and Ṣafavids, to tie them to the state financially. His personal curiosity about other religions was exemplary; with the help of Abu'l-Faẓl, his Ṣūfī adviser and biographer, he established a kind of salon for religious discussion. A very small circle of personal disciples seems to have emulated Akbar's own brand of *tawḥīd-i ilāhī* ("divine oneness"). This appears to have been a general monotheism akin to what the *ḥanīf*s of Mecca, and Muḥammad himself, had once practiced, as well as to the boundary-breaking pantheistic awareness of great Ṣūfīs like ar-Rūmī and Ibn al-ʿArabī, who was very popular in South and Southeast Asia. Akbar combined toleration for all religions with condemnation of practices that seemed to him humanly objectionable, such as enslavement and the immolation of widows.

Continuation of the empire. For half a century, Akbar's first two successors, Jahāngīr and Shāh Jahān, continued his policies. A rebuilt capital at Delhi was added to the old capitals of Fatehpur Sīkri and Āgra, site of Shāh Jahān's most famous building, the Tāj Mahal. The mingling of Hindu and Muslim traditions was expressed in all the arts, especially in naturalistic and sensuous painting; extremely refined and sophisticated design in ceramics, inlay-work, and textiles; and in delicate yet monumental architecture. Shāh Jahān's son, Dārā Shikōh (1615–59), was a Ṣūfī thinker and writer who tried to establish a common ground for Muslims and Hindus. In response to such attempts, a Sharīʿah-minded movement of strict communalism arose, connected with a leader of the Naqshbandī *ṭarīqah* named Shaykh Aḥmad Sirhindī. With the accession of Aurangzeb (ruled 1659–1707) the tradition of ardent ecumenicism, which would reemerge several centuries later in a non-Muslim named Mohandas K. (Mahatma) Gandhi, was replaced with a stricter communalism that imposed penalties on protected non-Muslims and stressed the shah's role as leader of the Muslim community, by virtue of his enforcing the Sharīʿah. Unlike the Ottoman and Ṣafavid domains, the Indo-Timurid empire was still expanding right up to the beginning of the 18th century; but the empire began to disintegrate shortly after the end of Aurangzeb's reign, when Ṣafavid and Ottoman power were also declining rapidly.

Rise of communalism

Between the 15th and 18th century the use of coffee, tea, and tobacco, despite the objections of the *ʿulamāʾ*, became common in all three empires. Teahouses became important new centres for male socializing, in addition to the home, the mosque, the marketplace, and the public bath. (Female socializing was restricted largely to the home and the bath.) In the teahouses men could practice the already well-developed art of storytelling and take delight in the clever use of language. *The Thousand and One Nights* (*Alf laylah wa laylah*), the earliest extant manuscripts of which date from this period, and the stories of the Arabian hero ʿAntar must have been popular, as were the tales of a wise fool known as Mullah Nasroddin in Persian (Nasreddin), Hoca in Turkish, and Juḥā in Arabic. The exploits of Nasroddin, sometimes in the guise of Ṣūfī dervish or royal adviser, often humorously portray centralized absolutism and mysticism: "Nasroddin was sent by the King to investigate the lore of various kinds of Eastern mystical teachers. They all recounted to him tales of the miracles and the sayings of the founders and great teachers, all long dead, of their schools. When he returned home, he submitted his report, which contained the single word 'Carrots.' He was called upon to explain himself. Nasroddin told the King: 'The best part is buried; few know—except the farmer—by the green that there is orange underground; if you don't work for it, it will deteriorate; there are a great many donkeys associated with it.' "

TRANS-SAHARAN ISLĀM

When the Ottomans expanded through the southern Mediterranean coast in the early 16th century, they were unable to incorporate Morocco, where a new state had been formed in reaction to the appearance of the Portuguese. The Portuguese were riding the momentum generated by their own seaborne expansion as well as by the fulfillment of the Reconquista and the establishment of an aggressively intolerant Christian regime in the centre of the Iberian Peninsula. In Morocco, it was neither the fervour of warriors nor Shīʿite solidarity nor Timurid restoration that motivated the formation of a state; rather it was a very old form of legitimacy that had proved to be especially powerful in Africa, that of the *sharīf*s, descendants of Muḥammad. It had last been relied on with the Idrīsids; now the *sharīf*s were often associated with Ṣūfī holy men, known as marabouts. It was one such Ṣūfī, Sīdī Barakāt, who legitimated the Saʿdī family of *sharīf*s as leaders of a *jihad* that expelled the Portuguese and established an independent state (1511–1603) strong enough to expand far to the south. Meanwhile the greatest Muslim kingdom of the Sudan, Songhai, was expanding northward; and its growing control of major trade routes into Morocco provoked Moroccan interference. Invaded in 1591, Songhai was ruled as a Moroccan vassal for 40 years, during which time Morocco itself was experiencing political confusion and instability. Morocco was reunited in 1668 by the ʿAlawite *sharīf*s. A holy family of Sijilmassa, they were brought to power by Arab tribal support, which they eventually had to replace with a costly army of black slaves. Like the Saʿdīs, they were legitimated in two ways: by the recognition of leading Ṣūfīs and by the special spiritual quality (*barakah*) presumed to have passed to them by virtue of their descent from the Prophet through ʿAlī. Although they were not Shīʿites, they cultivated charismatic leadership that undermined the power of the *ʿulamāʾ* to use the Sharīʿah against them. They also recognized the limits of their authority as absolute monarchs, dividing their realm into the area of authority and the area of no authority (where many of the Berber tribes lived). Thus the Moroccan *sharīf*s solved the universal problems of legitimacy, loyalty, and control in a way tailored to their own situation.

Islāmic states in the Sudanic region

While the Saʿdī dynasty was ruling in Morocco, but long before its incursions into the Sahara, a number of small Islāmic states were strung from one end of the Sudanic region to the other: Senegambia, Songhai, Aïr, Mossi, Nupe, Hausa, Kanem-Bornu, Darfur, and Funj. Islām had come to these areas along trade and pilgrimage routes, especially through the efforts of a number of learned teaching-trading families such as the Kunta. Ordinarily the ruling elites became Muslim first, employing the skills of Arab immigrants, traders, or travelers, and taking political and commercial advantage of the Arabic language and the Sharīʿah without displacing indigenous religious practices or legitimating principles. By the 16th century the Muslim states of the Sudanic belt were in contact not only with the major Muslim centres of the Maghrib and Egypt, but also with each other through an emerging trans-Sudanic pilgrimage route. Furthermore, Islām had by then become well enough established to provoke efforts at purification comparable with the Almoravid movement of the 11th century. Sometimes these efforts were gradualist and primarily educational, as was the case with the enormously influential Egyptian scholar as-Suyūṭī (1445–1505). His works, read by many West African Muslims for centuries after his death, dealt with numerous subjects, including the coming of the *mahdī* to restore justice and strengthen Islām. He also wrote letters to Muslim scholars and rulers in West Africa more than 2,000 miles away, explaining the Sharīʿah and encouraging its careful observance.

Other efforts to improve the observance of Islām were more militant. Rulers might forcibly insist on an end to certain non-Muslim practices, as did Muḥammad Rumfa (ruled 1463–99) in the Hausa city-state of Kano, or Muḥammad I Askia, the greatest ruler of Songhai (ruled 1493–1528). Often, as in the case of both of these rulers, militance was encouraged by an aggressive reformist scholar like al-Maghīlī (flourished 1492), whose writings detailed the conditions that would justify a *jihād* against Muslims who practiced their faith inadequately. Like many reformers, al-Maghīlī identified himself as a *mujaddid*, a figure expected to appear around the turn of each Muslim century. (The 10th century of the *hijrah* era began in 1494.) To the east in Ethiopia, an actual *jihād* was carried out by Aḥmad Grāñ (*c.* 1506–43), in the name of opposition to the Christian regime and purification of "compromised" Islām. Further to the east, a conquest of Christian Nubia by Arab tribes of Upper Egypt resulted in the conversion of the pagan Funj to Islām and the creation of a major Muslim kingdom there. Although most indigenous West African scholars looked to foreigners for inspiration, a few began to chart their own course. In Timbuktu, where a rich array of Muslim learning was available, one local scholar and member of a Tukulor learned family, Aḥmad Bābā, was writing works that were of interest to North African Muslims. Local histories written in Arabic also survive, such as the *Taʾrīkh al-fattāsh* (written by several generations of the Kāti family, from 1519 to 1665), a chronological history of Songhai, or as-Saʿdī's *Taʾrīkh al-Sūdān* (completed in 1655). By the end of the period of consolidation and expansion, Muslims in the Sudanic belt were being steadily influenced by North African Islām but were also developing distinctive traditions of their own.

INDIAN OCEAN ISLĀM

A similar relationship was simultaneously developing across another "sea," the Indian Ocean, which tied South and Southeast Asian Muslims to East African and south Arabian Muslims the way the Sahara linked North African and Sudanic Muslims. Several similarities are clear: the alternation of advance and retreat, the movement of outside influences along trade routes, and the emergence of significant local scholarship. There were differences, too: Indian Ocean Muslims had to cope with the Portuguese threat and to face Hindus and Buddhists more than pagans, so that Islām had to struggle against sophisticated and refined religious traditions that possessed written literature and considerable political power.

The Aceh state

The first major Muslim state in Southeast Asia, Aceh, was established around 1524 in northern and western Sumatra in response to more than a decade of Portuguese advance. Under Sultan Iskandar Muda (ruled 1608–37), Aceh reached the height of its prosperity and importance in the Indian Ocean trade, encouraging Muslim learning and expanding Muslim adherence. By the end of the 17th century, Aceh's Muslims were in touch with major intellectual centres to the west, particularly in India and Arabia, just as West African Muslims were tied to centres across the Sahara. Because they could draw on many sources, often filtered through India, Sumatran Muslims may have been exposed to a wider corpus of Muslim learning than Muslims in many parts of the heartland. Aceh's scholarly disputes over Ibn al-ʿArabī were even significant enough to attract the attention of a leading Medinan, Ibrahim al-Kurani, who in 1640 wrote a response. The same kind of naturalization and indigenization of Islām that was taking place in Africa was also taking place here; for example, ʿAbd ar-Raʾūf of Singkel, after studying in Arabia from about 1640 to 1661, returned home, where he made the first "translation" of the Qurʾān into Malay, a language that was much enriched during this period by Arabic script and vocabulary. This phenomenon extended even to China. Liu Chih, a scholar born around 1650 in Nanking, created serious Islāmicate literature in Chinese, including works of philosophy and law.

In the early 17th century another Muslim commercial power emerged when its ruler, the prince of Tallo, converted; Macassar (now Ujung Pandang) became an active centre for Muslim competition with the Dutch into the third quarter of the 17th century, when its greatest monarch, Ḥasan ad-Dīn (ruled 1631–70), was forced to cede his independence. Meanwhile, however, a serious Islāmic presence was developing in Java, inland as well as on the coasts; by the early 17th century the first inland Muslim state in Southeast Asia, Mataram, was established. There Ṣūfī holy men performed a missionary function similar to that being performed in Africa. Unlike the more seriously Islāmized states in Sumatra, Mataram suffered, as did its counterparts in West Africa, from its inability to suppress indigenous beliefs to the satisfaction of the more conservative *'ulamā'*. Javanese Muslims, unlike those in Sumatra, would have to struggle for centuries to negotiate the confrontation between Hindu and Muslim cultures. Their situation underscores a major theme of Islāmicate history through the period of consolidation and expansion; that is, the repeatedly demonstrated absorptive capacity of Muslim societies, a capacity that was soon to be challenged in unprecedented ways.

Reform, dependency, and recovery (1683 to the present)

The history of the Muslims in modern times has often been explained in terms of the impact of "the West." From this perspective, the 18th century was a period of degeneration and a prelude to European domination, symbolized by Napoleon's conquest of Egypt in 1798. Given the events of the 1980s, however, it is possible to argue that the period of Western domination was an interlude in the ongoing development of indigenous styles of modernization. In order to examine that hypothesis, it is necessary to begin the "modern" period with the 18th century, when activism and revival were present throughout Islāmdom. The three major Muslim empires did experience a decline during the 18th century, as compared to their own earlier power and to the rising powers in Europe; but most Muslims were not yet aware that Europe was partly to blame. Similar decline had occurred many times before, a product of the inevitable weaknesses of the military conquest state turned into centralized absolutism, overdependence on continuous expansion, weakening of training for rule, the difficulty of maintaining efficiency and loyalty in a large, complex royal household and army, and the difficulty of maintaining sufficient revenues for an increasingly lavish court life. Furthermore, population increased, as it did almost everywhere in the 18th-century world, just as inflation and expensive reform reduced income to central governments. Given the insights of Ibn Khaldūn, however, one might have expected a new group with a fresh sense of cohesiveness to restore political strength.

Economic and social strengths in western Europe

Had Muslims remained on a par with all other societies, they might have revived. But by the 18th century one particular set of societies in western Europe had developed an economic and social system capable of transcending the 5,000-year-old limitations of the agrarian-based settled world as defined by the Greeks (who called it Oikoumene). Unlike most of the lands of Islāmdom, those societies were rich in natural resources (especially the fossil fuels that could supplement human and animal power) and poor in space for expansion. Cut off by Muslims from controlling land routes from the East, European explorers had built on and surpassed Muslim seafaring technology to compete in the southern seas and discover new sea routes—and, accidentally, a new source of wealth in the Americas. In Europe, centralized absolutism, though an ideal, had not been the success it was in Islāmdom. Emerging from the landed classes rather than from the cities, it had benefited from and been constrained by independent urban commercial classes. In Islāmdom, the power of merchants had been inhibited by imperial overtaxation of local private enterprise, appropriation of the benefits of trade, and the privileging of foreign traders through agreements known as the Capitulations.

In Europe independent financial and social resources promoted an unusual freedom for technological experimentation and, consequently, the technicalization of other areas of society as well. Unlike previous innovations in the Oikoumene, Europe's technology could not easily be diffused to societies that had not undergone the prerequisite fundamental social and economic changes. Outside of Europe, gradual assimilation of the "new," which had characterized change and cultural diffusion for 5,000 years, had to be replaced by hurried imitation, which proved enormously disorienting. This combination of innovation and imitation produced an unprecedented and persisting imbalance among various parts of the Oikoumene. Muslims' responses paralleled those of other "non-Western" peoples but were often filtered through and expressed in peculiarly Islāmic or Islāmicate symbols and motifs. The power of Islām as a source of public values had already waxed and waned many times; it intensified in the 18th and 19th centuries, receded in the early 20th century, and surged again after the mid-20th century. Thus European colonizers appeared in the midst of an ongoing process that they greatly affected but did not completely transform.

PRE-COLONIAL REFORM AND EXPERIMENTATION (1683–1818)

From the mid-17th century through the 18th and early 19th centuries certain Muslims expressed an awareness of internal weakness. In some areas, Muslims were largely unaware of the rise of Europe; in others, such as India, Sumatra, and Java, the 18th century actually brought European control. Responses to decline, sometimes official and sometimes unofficial, sometimes Islāmizing, sometimes Europeanizing, fell into two categories, as the following examples demonstrate.

In some areas, leaders attempted to revive existing political systems. In Iran, for example, attempts at restoration combined military and religious reform. Around 1730 a Turk from Khorāsān named Nāder Qolī Beg reorganized the Ṣafavid army in the name of the Ṣafavid shah, whom he replaced with himself in 1736. Nāder Shāh extended the borders of the Ṣafavid state further than ever; he even defeated the Ottomans and may have been aspiring to be the leader of all Muslims. To this end he made overtures to neighbouring rulers, seeking their recognition by trying to represent Iranian Shī'ism as a *madhhab* alongside the Sunnite *madhhabs*. After he was killed in 1747, however, his reforms did not survive and his house disintegrated. Karīm Khān Zand, a general from Shīrāz, ruled in the name of the Ṣafavids but did not restore real power to the shah. By the time the Qājārs (1779–1925) managed to resecure Iran's borders, reviving Ṣafavid legitimacy was impossible.

Restoration in the Ottoman Empire

In the Ottoman Empire, restoration involved selective imitation of things European. Its first phase, from 1718 to 1730, is known as the Tulip Period, because of the cultivation by the wealthy of a Perso-Turkish flower then popular in Europe. Experimentation with European manners and tastes was matched by experimentation with European military technology. Restoration depended on reinvigorating the military, the key to earlier Ottoman success, and Christian Europeans were hired for the task. After Nāder Shāh's defeat of the Ottoman army, this first phase of absolutist restoration ended, but the pursuit of European fashion had become a permanent element in Ottoman life. Meanwhile, central power continued to weaken, especially in the area of international commerce. The certificates of protection that had accompanied the Capitulations arrangements for foreign nationals were extended to non-Muslim Ottoman subjects, who gradually oriented themselves toward their foreign associates. The integration of such groups into the Ottoman state was further weakened by the recognition, in the disastrous Treaty of Küçük Kaynarca (1774), of the Russian tsar as protector of the Ottoman's Greek Orthodox *millet*. A second stage of absolutist restoration occurred under Selim III, who became sultan in the first year of the French Revolution and ruled until 1807. His military and political reforms, referred to as the New Order (Nizam-ı Cedid), went beyond the Tulip Period in making use of things European; for example, the enlightened monarch, as exemplified by Napoleon himself, became an Ottoman ideal. Here, as in Egypt under Muḥammad 'Alī (reigned 1805–48), the famed core of Janissaries that had been a source of Ottoman strength was destroyed and replaced with European-trained troops.

In other areas, leaders envisioned or created new social orders that were self-consciously Islāmic. The growing popularity of westernization and a decreasing reliance on Islām as a source of public values was counterbalanced in many parts of Islāmdom by all sorts of Islāmic activism, ranging from educational reform to *jihād*. "Islāmic" politics often were marked by an oppositional quality that drew on long-standing traditions of skepticism about government. Ṣūfism could play very different roles. In the form of renovated *ṭarīqah*s it could support reform and stimulate pan-Islāmic awareness. Ṣūfīs often encouraged the study of *ḥadīth* so as to establish the Prophet Muḥammad as a model for spiritual and moral reconstruction and to invalidate many unacceptable traditional or customary Islāmic practices. Ṣūfī *ṭarīqah*s provided interregional communication and contact and an indigenous form of social organization that could even lead to the founding of a dynasty, as in the case of the Libyan monarchy.

Ṣūfism could also be condemned as a source of degeneracy. The most famous and influential militant anti-Ṣūfī movement arose in the Arabian Peninsula and called itself al-Muwaḥḥidūn ("the Monotheists"); but it came to be known as Wahhābīyah, after its founder, Muḥammad ibn ʿAbd al-Wahhāb (1703–92). Inspired by Ibn Taymīyah (see above *Migration and renewal (1041–1405)*), Ibn al-Wahhāb argued that the Qurʾān and *sunnah* could provide the basis for a reconstruction of Islāmic society out of the degenerate form in which it had come to be practiced. Islām itself was not an inhibiting force; "traditional" Islām was. Far from advocating the traditional, the Wahhābīs argued that what had become traditional had strayed very far from the fundamental, which can always be found in the Qurʾān and *sunnah.* The traditional they associated with blind imitation (*taqlīd*); reform, with making the pious personal effort (*ijtihād*) necessary to understand the fundamentals. Within an Islāmic context, this type of movement was not conservative, because it sought not to conserve what had been passed down but to renew what had been abandoned. The Wahhābī movement attracted the support of a tribe in the Najd led by Muḥammad ibn Saʿūd. Although the first state produced by this alliance did not last, it laid the foundations for the existing Saudi state in Arabia and inspired similar activism elsewhere down to the present day.

Islāmic activism in West Africa

In West Africa a series of activist movements appeared from the 18th century into the 19th. There as in Arabia, Islāmic activism was directed less at non-Muslims than at Muslims who had gone astray. As in many of Islāmdom's outlying areas, emergent groups of indigenous educated, observant Muslims, such as the Tukulor, were finding the casual, syncretistic, opportunistic nature of official Islām to be increasingly intolerable. Such Muslims were inspired by reformist scholars from numerous times and places—al-Ghazālī, as-Suyūṭī, Maghili; by a theory of *jihād* comparable to that of the Wahhābīs; and by expectations of a *mujaddid* as the Islāmic century turned in AH 1200 (AD 1785). In what is now northern Nigeria, the discontent of the 1780s and '90s erupted in 1804, when Usman dan Fodio declared a *jihād* against the Hausa rulers. Others followed, among them Muhammad al-Jaylani in Aïr, Shehuh Ahmadu Lobbo in Macina, al-Ḥajj ʿUmar Tal (a member of the reformist Tijānī *ṭarīqah*) in Fouta Djallon, and Samory in the Malinke (Mandingo) states. *Jihād* activity continued for a century; it again became millennial near the turn of the next Muslim century in AH 1300 (AD 1882), as the need to resist against European occupation became more urgent. For example, Muḥammad Aḥmad declared himself to be the *mahdī* in the Sudan in 1881.

In the Indian Ocean area, Islāmic activism was more often intellectual and educational. Its best exemplar was Shāh Walī Allāh of Delhi (1702–62), the spiritual ancestor of many later Indian Muslim reform movements. During his lifetime the collapse of Muslim political power was painfully evident. He tried to unite the Muslims of India, not around Ṣūfism as Akbar had tried to do, but around the Sharīʿah. Like Ibn Taymīyah, he understood the Sharīʿah to be based on firm sources—Qurʾān and *sunnah*—that could with pious effort be applied to present circumstances. Once again, the study of *ḥadīth* provided a rich array of precedents and inspired a positive spirit of social reconstruction akin to that of the Prophet Muḥammad.

DEPENDENCY (1818–1962)

The many efforts to revive and resist were largely unsuccessful. By 1818, British hegemony over India was complete; and many other colonies and mandates followed between then and the aftermath of World War I. Not all Muslim territories were colonized, but nearly all experienced some kind of dependency, be it psychological, political, technological, cultural, or economic. Perhaps only the Saudi regime in the central parts of the Arabian Peninsula could be said to have escaped any kind of dependency; but even there oil exploration, begun in the 1930s, brought European interference. In the 19th century westernization and Islāmic activism coexisted and competed. By the turn of the 20th century secular ethnic nationalism had become the most common mode of protest in Islāmdom; but the spirit of Islāmic reconstruction was also kept alive, either in conjunction with secular nationalism or in opposition to it.

In the 19th-century Ottoman Empire, selective westernization coexisted with a reconsideration of Islām. The program of reform known as the Tanzimat, which was in effect from 1839 to 1876, aimed to emulate European law and administration by giving all Ottoman subjects, regardless of religious confession, equal legal standing and by limiting the powers of the monarch. In the 1860s a group known as the Young Ottomans tried to identify the basic principles of European liberalism and even love of nation with Islām itself. In Iran, the Qājār shahs brought in a special "Cossack Brigade," trained and led by Russians, while at the same time the Shīʿite *mujtahid*s viewed the decisions of their spiritual leader as binding on all Iranian Shīʿites and declared themselves to be independent of the shah. (One Shīʿite revolt, that of the Bāb [died 1850], led to a whole new religion, Bahāʾī.) Like the Young Ottomans, Shīʿite religious leaders came to identify with constitutionalism in opposition to the ruler.

Reaction against Europeans and westernization

Islāmic protest often took the form of *jihād* against the Europeans: by Southeast Asians against the Dutch; by the Sanūsī *ṭarīqah* over Italian control in Libya; by the Mahdist movement in the Sudan; or by the Ṣaliḥī *ṭarīqah* in Somalia, led by Sayyid Muḥammad ibn ʿAbd Allāh Ḥasan, who was tellingly nicknamed the Mad Mullah by Europeans. Sometimes religious leaders, like those of the Shīʿites in Iran, took part in constitutional revolutions (1905–11). Underlying much of this activity was a pan-Islāmic sentiment that drew on very old conceptions of the *ummah* as the ultimate solidarity group for Muslims. Three of the most prominent Islāmic reconstructionists were Jamāl ad-Dīn al-Afghānī, his Egyptian disciple Muḥammad ʿAbduh, and the Indian poet Sir Muḥammad Iqbāl. All warned against blind pursuit of Westernization, arguing that the blame for the weaknesses of Muslims lay not with Islām, but rather with Muslims themselves, because they had lost touch with the progressive spirit of social, moral, and intellectual reconstruction that had made early Islāmicate civilization one of the greatest in human history. Although al-Afghānī, who taught and preached in many parts of Islāmdom, acknowledged that organization by nationality might be necessary, he viewed it as inferior to Muslim identity. He further argued that Western technology could advance Muslims only if they retained and cultivated their own spiritual and cultural heritage. He pointed out that at one time Muslims had been intellectual and scientific leaders in the world, identifying a Golden Age under the ʿAbbāsid caliphate and pointing to the many contributions Muslims had made to "the West." Like al-Afghānī, Iqbāl assumed that without Islām Muslims could never regain the strength they had possessed when they were a vital force in the world, united in a single international community and unaffected by differences of language or ethnos. This aggressive recovery of the past became a permanent theme of Islāmic reconstruction. In many regions of Islāmdom the movement known as Salafīyah also identified with an ideal time in history, that of the "pious ancestors" (*salaf*) in the early Muslim state of Muḥammad and his companions, and advocated past-oriented change to bring present-day Muslims up to the progressive standards of an earlier ideal. In addition to clearly Islāmic thinkers, there were others, such as the Egyptian Muṣṭafā Kāmil, whose nationalism was not simply secular. Kāmil saw Egypt as simultaneously European, Ottoman, and Muslim. The Young Turk Revolution of

1908 was followed by a period in which similarly complex views of national identity were discussed in the Ottoman Empire.

RECOVERY (1922 TO THE PRESENT)

Progress of secular nationalism. Despite the ideological appeal of such positions, the need to throw off European control promoted the fortunes of secular nationalism and other narrower forms of loyalty. Especially after Japan's defeat of Russia in 1905, nationalist fervour increased. Sometimes it was associated with related ideologies, such as pan-Arabism, Pan-Turkism, or Arab socialism. Many nationalists enthusiastically admired things European despite the fact that they were committed to resisting or removing European control. Often accepting European assessments of traditional religion as a barrier to modernization, many nationalists sought an identity in the pre-Islāmic past. Kemal Atatürk looked to the Turkic past in Central Asia and Anatolia to transform Ottomanism into a Turkish identity not dependent on Islām. "Islāmic" dress was discouraged. Muslim males, who prayed with covered heads, were now asked to replace the fez, which could be kept on during prayer, with the brimmed hat, which could not. Arabic script, too closely associated with Islām, was replaced with the Roman, after the Cyrillic (the alphabet of Central Asian Turks) had been considered and rejected. In Iran, Reza Shah Pahlavi argued that the Islāmic period was but an accidental interlude in the continuous history, since Achaemenid times, of Iran as a unified entity. The Egyptian Taha Hussein connected his country's national identity with Pharaonic times and with Mediterranean–European culture; and therefore it could easily partake of modern Western civilization. Christians were thus as much Egyptians as were Muslims; the accompanying development of a standard literary Arabic, *fuṣḥā,* emphasized the unity of all Arabs, regardless of confession. These approaches allowed, indeed required, all religious communities to partake of a single legal and societal system, at the price of denying the public relevance of a primary loyalty for the majority of the population.

Other nationalists made more of Islām. In Saudi Arabia and Pakistan, for example, Islām played a primary role in the formation of a national identity. In Pakistan it provided, according to the statesman Mohammed Ali Jinnah, an alternative for Muslims who would otherwise have to share in an identity defined by a Hindu majority. In many Arab countries, especially in the Maghrib, secular nationalism's downgrading of Islām was muted by a qualified acceptance of Islām as one, but not the only, important source of loyalty. At the same time there were Muslims who opposed nationalism altogether. In India, Mawlanā Abu'l-'Alā' Mawdūdī, who was the founder of the Jamā'at-i Islāmī, opposed both secular and religious nationalism and argued for the Islāmization of society and an Islāmic alternative to nationalism. In Egypt, Sayyid Quṭb and Ḥasan al-Bannā', who were the mentors of the Muslim Brotherhood, fought for the educational, moral, and social reform of an Islāmic Egypt and indeed of all Islāmdom.

Opposition to nationalism

Creating national identities. Only a few existing states where Muslims predominate, such as Turkey and Saudi Arabia, had no colonial interval; most became independent after World War II. An even larger number of countries have Muslim minorities. Like the citizens of many new nations, Muslims have not found the creation of national identities to be easy, especially considering the pace at which it has had to occur. More than two-thirds of the world's nations have come into existence since the end of World War II; foreign dependency is a living memory for many of their citizens, or at least for the parents and grandparents of their citizens. Many of them are not nation-states—that is, states established by a group of people who decided that they belonged together and therefore went about acquiring sovereignty over a territory—but rather are state-nations, composed of groups of people who acquired or were given sovereignty over a territory and then had to develop a sense of nationality. The most obvious state-nations are Syria, Iraq, Lebanon, and Jordan. All resulted from the interaction of intra-European rivalry and diplomacy with the aspirations of a prominent Ottoman-Hāshimite sharifian family in Mecca to create a single Arab state in the East. Instead of a single state, however, three monarchies emerged: the kingdom of Ḥusayn ibn 'Alī in the Hejaz (to be replaced by the Saudis), the kingdom of Fayṣal I in Iraq (because he had to be compensated for being ousted from Syria), and the kingdom of Abdullah in Transjordan. Lebanon was carved from French Syria with borders that would establish a bare Christian majority loyal to the French. In Ottoman Palestine, Jewish nationalists clashed with Arab nationalists, at a time when both groups felt betrayed by the British. In subsequent armed clashes, Zionist groups defended a set of boundaries as artificial as many others, creating a state that has remained a target for anti-imperialist sentiment. Eventually, Jewish nationalism spawned another nationalism, that of the Palestinians, inchoate before the founding of Israel but crystallized by the failure of any party to the conflict—Arab states, foreign powers, Palestinian leaders, or Israel itself—to make a place for most of the former Arab residents of Palestine.

Many Muslim countries were united by negative nationalism, aimed at ejecting a common enemy; but turning negative into positive has been difficult. Rarely have the groups that achieved independence survived. Often, as in Libya or Iraq or Egypt, further revolutions have occurred, in many cases led by the military, whose role as a vehicle for modernization cannot be underestimated. Subsequent governments have had to deal with the social and economic problems that plague all developing countries, as well as with regional rivalries and conflicts. Almost nowhere did the colonizers leave an infrastructure sufficient to support the growth of population that European medicine and hygiene had produced.

Relation of religion and nationality. Given the multicommunal structure of premodern Muslim societies, the relation between religion and nationality has been another major problem. Nationalism has frequently led to competition and rivalry among a new nation's religious communities. As they became independent, citizens of the nations of Islāmdom could draw on no direct equivalent of national identity. The broadest identity was provided by membership in a pan-territorial community like the *ummah* of all Muslims, or the Greek Orthodox Church, or the Turkic tribes; the narrowest, family or neighbourhood. In the middle of the spectrum was membership in a local confessional community, with all its implications of status, occupation, manners, and customs. Citizens of the new nations would theoretically have to find an identity that could subsume and supersede all others; and the rulers of new nations would have to take the unprecedented step of declaring all citizens subject to the same law, rather than members of quasi-autonomous, self-governing religious communities with their own legal systems. Yet the significance of being a member of a religious community could not easily be undone or replaced.

Rivalry among religious communities

Many countries inherited a relatively simple form of this problem: the people within their borders were primarily of one faith, Islām, and of one form of that faith, the Sunnite. That majority adherence could in some way be associated with or bolster the national identity, while discomfiting only a small number of people. Turkey, Iran, Jordan, Indonesia, the Yemens, and all the states of North Africa and the Arabian Peninsula fall into this category. Even so, religious minorities in these countries (such as the Armenians) suffered and shrank; for Jews communal lines were hardened by the emergence of the state of Israel, the hostility it evoked from most Arab states, and its aggressive efforts at ingathering. The self-consciously Islāmic government in Iran has also introduced a religious intolerance that, while it is discouraged by the Sharī'ah, is encouraged by local sentiment as well as by the staunch nationalism Iran shares with secular states. In reaction to the Pahlavi state they overthrew, in which trying to restore Zoroastrianism had not been unthinkable, the leaders of the Islāmic Republic of Iran have associated being Iranian with being Muslim.

Farther from the centre of Islāmdom, Islām plays various roles as a minority religion. Among Turks in the southern republics of the Soviet Union, for example, Islām is an important source of identity. Muslims living in western Europe and the Americas are generally able to form communities and practice their religion as they will: in Canada, for example, Ismāʿīlī Muslims, under the guidance of Aga Khan IV, form a cohesive group that promotes the economic and cultural development of its members. In the United States, tenets of Islām were embraced by the founders of the American Muslim Mission (originally called Nation of Islam) in the early 1930s. As the community has developed, its leaders have increasingly emphasized the Qurʾān and Muḥammad's example as sources of authority.

Pan-Islāmic organizations

Survival of Islāmic activism. Although Islāmic activism never disappeared during the years in which Muslim countries were becoming independent, other ideological orientations seemed more important between the end of World War II and the declaration of the Islāmic Republic of Iran in 1979. Many Westerners or westernized Muslims expected religion to recede as modernization progressed. Already in the 1950s, however, the Muslim Brotherhood in Egypt called for an exclusively Islāmic state in place of the secular multi-communal state that Gamal Abdel Nasser had founded. In the early 1960s new circumstances were beginning to foster increased self-consciously Islāmic activity, some popular, some supported by official institutions. In these years critics of Mohammad Reza Shah Pahlavi began to rally around the exiled ayatollah Ruhollah Khomeini; the writings of ʿAli Shariʿati began to influence Muslims inside and outside Iran; and two great pan-Islāmic organizations were formed, the Muslim World League (1962) and the Organization of the Islāmic Conference (1971). Although Westerners have become most familiar with activism's violent forms, its educational, cultural, pietistic, and political dimensions have been more extensive. All these developments occurred in the wake of the formation of the Organization of Petroleum Exporting Countries in 1961 and culminated in Egypt's success in its war with Israel in 1973. The resurgence of economic and military power was not the only factor that could foster those who had maintained an interest in Islām all along. In a few parts of the Muslim world, petroleum-based prosperity promoted increased international influence and pride; elsewhere modernization was producing widespread educational and economic cleavages and populations with very low median ages. As dissatisfaction with the material failures of secular modernization grew, so did disenchantment with the Western ideologies that had undergirded it. While these other ideologies were being tried and discredited, Islām had remained relatively peripheral to public policy, and thus unassailable. All the while, citizens of Muslim countries were echoing the anti-imperialist rhetoric that was increasing throughout the developing world.

Situation of Muslim women. For women, modernization is especially problematic. Urged on the one hand to be liberated from Islām and thereby become modern, they are told by others to be liberated from being Western through being self-consciously Muslim. There is little information on the situation of ordinary women in premodern Islāmdom, but evidence from the modern period underscores the enormous variety of settings in which Muslim women live and work, as well as the inability of the stereotype of meek, submissive, veiled passivity to reflect the quality of their lives. As always, Muslim women live in cities, towns, villages, and among migratory pastoral tribes; some work outside the home, some inside, some not at all; some wear concealing clothing in public, most do not; for some, movement outside the home is restricted, for most not; and, for many, public modesty is common, as it is for many Muslim men. For many, the private home and the public bath continue to be the centres of social interaction; for others, the world of employment and city life is an option. As always, few live in polygamous families. Strict adherence to the Sharīʿah's provision for women to hold their property in their own right has produced Muslim women of great wealth, in the past as well as today. Clearly, any simple description of the lives of Muslim women is misleading.

Modern Islām's unifying forces. Modern Islāmdom can appear so diverse as to defy description, yet it is also held together by stronger centripetal forces than almost any other pan-national solidarity group. The *ḥajj* attracts more than 1,000,000 Muslims annually; and, despite significant religious cleavages, Islām remains one of the least sectarian of world religions. Most Muslims live in societies in which the force of tradition is very strong and in which modernization has also penetrated to some extent. The majority of Muslims remain, as they have always been, agricultural. A very small minority are migratory pastoralists; a larger minority are village, town, and city dwellers. In all settings tradition, including religious tradition, is being drawn upon as a source of change and modernization, with the consequence that the Western equation of modernization and secularization has been severely tested and even undermined.

Yet the role of tradition varies. Some Islāmic activists rely on a kind of secularized "cultural" Islām, somewhat like cultural Judaism, that depends very little on personal piety or the observance of Islāmic law or the many customs that have come to be associated with being Muslim, while others cling to the customs associated with Islām with little awareness of Islām's more learned side. Labels such as Shīʿite, which always carried an oppositional quality, may be formerly nonessential attributes that have become salient in the wake of the success of the ayatollah Khomeini in Iran. When disadvantaged persons who happen to be Shīʿites find an opening for communal protest, or when those for whom Shīʿite theology means little find its vision of justice and radical revolution appropriate to their specific circumstances, an old label acquires a new valence.

Like any other explanatory system, Islām has always had to provide a way of talking about the world, of establishing identity in the world, and of managing the world's affairs. In performing these functions, Islām has from its inception been forced to compete with other explanatory systems for the "mental space" of its adherents and simultaneously to define its stance toward preexisting and ongoing extra-Islāmic influences. Islām continues to compete, aided unwittingly by the weaknesses of its competitors, spurred on by the freshness of its own demands for public attention, and fueled by the remarkable ability of many of its adherents to respond to the connection between the mundane and spiritual that has been the hallmark of all religious life.

BIBLIOGRAPHY

Surveys: The most visionary general work on Islāmic history is MARSHALL G.S. HODGSON, *The Venture of Islam: Conscience and History in a World Civilization,* 3 vol. (1974), which sets Islām into a world historical context. A similar but shorter work, sumptuously illustrated, is FRANCIS ROBINSON, *Atlas of the Islamic World Since 1500* (1982).

Regions of Islāmdom: PETER B. CLARKE, *West Africa and Islam: A Study of Religious Development from the 8th to the 20th Centuries* (1982); JAMIL M. ABUN-NASR, *A History of the Maghrib,* 2nd ed. (1975); CLIFFORD GEERTZ, *Islam Observed: Religious Development in Morocco and Indonesia* (1968, reissued 1971); S.M. IKRAM, *Muslim Rule in India and Pakistan, 711–1858 A.C.,* rev. ed. (1966); RAPHAEL ISRAELI, *Muslims in China: A Study in Cultural Confrontation* (1980); and NEHEMIA LEVTZION (ed.), *Conversion to Islam* (1979).

Periods and aspects of Islāmicate history: On premodern Islāmicate social structure, see ROY P. MOTTAHEDEH, *Loyalty and Leadership in an Early Islamic Society* (1980); IRA LAPIDUS, *Muslim Cities in the Later Middle Ages* (1967); and S.D. GOITEIN, *A Mediterranean Society: The Jewish Communities of the Arab World as Portrayed in the Documents of the Cairo Geniza,* 4 vol. (1967–83). HAMILTON A.R. GIBB, *Studies on the Civilization of Islam* (1962, reissued 1982), is a collection of interpretive articles on history, historiography, literature, and philology. RENÉ GROUSSET, *The Empire of the Steppes: A History of Central Asia* (1970; originally published in French, 1939); and JOHN J. SAUNDERS, *The History of the Mongol Conquests* (1971), deal with the Mongol conquests. JOHN J. SAUNDERS (ed.), *The Muslim World on the Eve of Europe's Expansion* (1966), combines primary sources on the last three great empires; and the most comprehensive account of modern Islām, with an especially fine treatment of the 18th century, is JOHN OBERT VOLL, *Islam, Continuity and Change in the Modern*

World (1982). On Muslim women, see, for example, LOIS BECK and NIKKI KEDDIE (eds.), *Women in the Muslim World* (1978); ELIZABETH WARNOCK FERNEA and BASIMA QATTAN BEZIRGAN (eds.), *Middle Eastern Muslim Women Speak* (1977, reprinted 1984); and JANE I. SMITH (ed.), *Women in Contemporary Muslim Societies* (1980).

Collections of primary sources in English translation: ERIC SCHROEDER, *Muhammad's People* (1955); ARTHUR JEFFERY (ed.), *A Reader of Islam* (1962, reprinted 1980); JOHN ALDEN WILLIAMS (ed.), *Islam* (1961, reissued 1967), and *Themes of Islamic Civilization* (1971, reprinted 1982); WILLIAM H. MCNEILL and MARILYN ROBINSON WALDMAN, *The Islâmic World* (1973, reprinted 1983); JAMES KRITZECK, *Anthology of Islamic Literature* (1964, reissued 1975); and BERNARD LEWIS (ed.), *Islam: From the Prophet Muhammad to the Capture of Constantinople,* 2 vol. (1974, reissued 1976).

Major reference works: The Encyclopaedia of Islam, 5 vol. (1913–36), and a new edition, of which 5 vol. appeared from 1960 to 1986; *The Shorter Encyclopaedia of Islam* (1953, reprinted 1974), with articles culled from the *Encyclopaedia of Islam; The Cambridge History of Islam,* 2 vol. (1970, reprinted in 4 vol., 1980); JEAN SAUVAGET, *Jean Sauvaget's Introduction to the History of the Muslim East: A Bibliographical Guide* (1965, reprinted 1982; originally published in French, 2nd ed., 1961), a dated but still useful annotated bibliographic guide; and CLIFFORD EDMUND BOSWORTH, *The Islamic Dynasties: A Chronological and Genealogical Handbook,* rev. ed. (1980). JEAN JACQUES WAARDENBURG, *L'Islam dans le miroir de l'Occident,* 3rd rev. ed. (1970); and EDWARD W. SAID, *Orientalism* (1978, reissued 1979), are critiques of Western approaches to Islām.

(M.R.W.)

Israel

Israel (in full State of Israel; Hebrew Medinat Yisra'el; Arabic Dawlat Isrā'īl) is a Middle Eastern republic situated at the eastern end of the Mediterranean Sea. It is bounded to the north by Lebanon, to the northeast by Syria, to the east and southeast by Jordan, to the southwest by Egypt, and to the west by the Mediterranean Sea. The total area is 7,992 square miles (20,700 square kilometres) excluding East Jerusalem and other territories occupied in the 1967 war. Jerusalem is the capital and the seat of government.

Following the United Nations partition of Palestine, Israel emerged as a sovereign state on May 15, 1948. It was the first Jewish state to be established in nearly 2,000 years. Its creation represented a fulfillment of the historic ideal of the Jewish people stemming from the traditional religious belief in God's promise of the land of Israel to the people of Israel. The ideal found practical expression in a desire to forge a nation without dependence on the goodwill of others. The establishment of Israel as a member of the family of nations signified a decisive step in modern Jewish history.

Among the population of Israel are hundreds of thousands of immigrants, many of them survivors of Nazi persecution in Europe or victims of anti-Semitism elsewhere. Israeli society has engaged in pioneering activities, including the rehabilitation of neglected agricultural lands. This has led to the creation of a Jewish rural population, which, though it makes up only about one-eighth of the total, also represents something almost unknown in the Diaspora (the historical scattering of the Jews in countries outside of Palestine). The revival of the Hebrew language has helped to make possible the cultural integration of the newcomers.

Hostile relations between Israel and its neighbouring Arab states have prevailed from the outset, with Israel obtaining victories over the Arabs after battles fought in 1948–49, 1956, 1967, and 1973. Territory occupied by Israeli forces after the 1967 and 1973 conflicts—including East Jerusalem, the West Bank, the Gaza Strip, and the Golan Heights region of Syria—is not treated in this article, although it is still held by Israel.

The article is divided into the following sections:

Physical and human geography

THE LAND

Relief. Israel may be divided into four natural regions. These are (1) the Mediterranean coastal plain, (2) the hill regions of northern and central Israel, (3) the Great Rift Valley, and (4) the Negev.

The coastal plain is a narrow strip about 115 miles (185 kilometres) long, widening to a breadth of about 20 miles in the south. In the north of the country, the mountains of Galilee constitute the highest part of Israel; their highest point is Mt. Meron, or (in Arabic) Jebel Jarmaq (3,963 feet [1,208 metres]). To the east these mountains terminate in an escarpment overlooking the Great Rift Valley. The mountains of Galilee are separated from the hills of Samaria and Judaea to the south by the Plain of Esdraelon ('Emeq Yizre'el), which, running approximately northwest to southeast, connects the coastal plain with the Great

Rift Valley. The Mount Carmel range, which culminates in a 1,791-foot peak, reaches northwest from the hills of Samaria and Judaea almost to the coast of Haifa.

The Jordan River

The Great Rift Valley, a long fissure in the Earth's crust, begins beyond the northern frontier of Israel and runs the length of the country to the Gulf of Aqaba. The Jordan River, which forms part of the frontier between Israel and Jordan, runs southward from Dan on Israel's northern frontier, where it is 500 feet above sea level, first into ‘Emeq H̱ula (H̱ula Basin), then into the freshwater Sea of Galilee, also known as Lake Tiberias or Yam Kinneret (689 feet below sea level), and finally into the highly saline Dead Sea, which is about 1,312 feet below sea level and which represents the lowest point of a natural landscape feature on the Earth's surface. The Negev, in the southern part of Israel, forms an arrow-shaped wedge of territory that comes to a point at the port of Elat (Eilat) on the Gulf of Aqaba.

Drainage. The principal drainage system is represented by the Jordan River. Other principal rivers in Israel are the Yarqon, which empties into the Mediterranean near Tel Aviv; the Qishon, which runs through the western part of the Plain of Esdraelon to drain into the Mediterranean at Haifa; and a small section of the Yarmuk, a tributary of the Jordan. The remaining streams, usually seasonal, flow through streambeds called wadis.

Soils. The coastal plain is covered mainly by alluvial soils. Because of its proximity to the coastal plain, parts of the arid northern Negev, where soil development would not be expected, have windblown loess soils. The soils of Galilee change from calcareous rock in the coastal plain to cenomanian and turonian limestone in Upper Galilee and to Eocene formations (from 38,000,000 to 54,000,000 years old) in the lower part of the region. Rock salt and gypsum are abundant in the Great Rift Valley.

Climate. Israel experiences great climatic contrasts. In the south, rainfall is light, amounting to about one inch (25 millimetres) a year in the ‘Arava Valley south of the Dead Sea, while in the north it is relatively heavy, amounting to 44 inches a year in the Upper Galilee region. Annual rainfall occurs on from 40 to 60 days spread over a season between October and April. Summers are dry and hot, but in the coastal areas sea breezes exert a moderating influence. Temperature depends largely on elevation and distance from the sea. The average daily maximum temperature in the coastal areas ranges between about 90° F (32° C) in August to about 65° F (18° C) in January. At Elat, in the south, daytime temperatures reach about 70° F (21° C) in January and may rise as high as 114° F (46° C) in August. Relative humidity is highest near the coast and is higher on summer than on winter nights. The Jordan Valley is hotter and drier than the coast. The hill regions have occasional snows in winter.

Plant and animal life. Vegetation is widely varied. The original evergreen forests have largely disappeared because of many centuries of cultivation and goat herding. The hills are mostly covered by maquis (wild shrub vegetation), and only desert scrub grows wild in the Negev and on the sand dunes of the coastal plain. North of Beersheba most of the country is under cultivation or is used for hill grazing; where irrigation is available, citrus groves and eucalyptus (introduced from Australia) and conifer plantations flourish. Millions of trees have been planted under a reafforestation program.

Nature reserves

Animal life is similarly varied. Mammals include wildcats, wild boars, gazelles, ibex, jackals, hyenas, hares, coneys, badgers, and tiger weasels. Among the reptiles, the agama and gecko lizards, the viper, and the carpet viper are found. Birds include the partridge, tropical cuckoo, bustard, sand grouse, and desert lark. There are many kinds of fish and insects. Invasions of desert locusts sometimes occur. Several regions have been set aside as nature reserves, notably parts of the ‘Arava in the south and Mt. Carmel, Mt. Meron, and the remains of the H̱ula Lake and marshes in the north.

Settlement patterns. The character of many regions has been altered by new patterns of Jewish settlement. The first modern-day Jewish settlers established themselves on the coastal plain in the 1880s. Later they also moved into the valleys of the interior and into parts of the hill districts, as well as into the Negev. The formerly Arab-populated areas of the coastal plain, the Judaean foothills, and the Jordan and ‘Arava valleys became almost exclusively Jewish. Although the majority of the Bedouin of the Negev left the region when it became Israeli territory, the Negev remained largely the domain of Arab nomads. The non-Jewish population is concentrated mainly in the north, where Arabs constitute a substantial part of the population of Galilee.

Rural settlement. The rural Jewish population amounts to about one-tenth of the total Jewish population. More than half the rural inhabitants are immigrants who arrived after 1948; two-thirds of the settlements were established after that date. The settlements are organized into kibbutzim, which are collective groups voluntarily practicing joint production and consumption; moshavim, which are smallholders' cooperatives practicing joint sales and purchases, making common use of machinery, ideally prohibiting hired labour, and leasing national land, usually for 49 years; and individually owned farms or villages in which private ownership is practiced. The kibbutzim and moshavim perform pioneer work in underdeveloped areas and security functions in border areas, and they contribute substantially to the national ability to absorb new immigrants.

Of the total rural population only about one-fifth, including the Bedouin, is non-Jewish. Before 1948 Jewish and Arab agricultural settlements existed side by side but were completely independent of each other. Since then, however, the demand for labour has resulted in thousands of Arab workers from the villages and the occupied territories finding employment in the citrus groves, in industry, or as construction labourers. This movement, together with increased agricultural mechanization, has led to a drop in the number of Jewish agricultural workers. In Arab villages fewer than half of the adult labourers, both men and women, are engaged in working the land. There has been a growing tendency to practice intensive cultivation, to diversify crops, and to extend farm areas. Most Arab farmers work their own land; some either lease land or work for Arab or Jewish landlords.

New towns

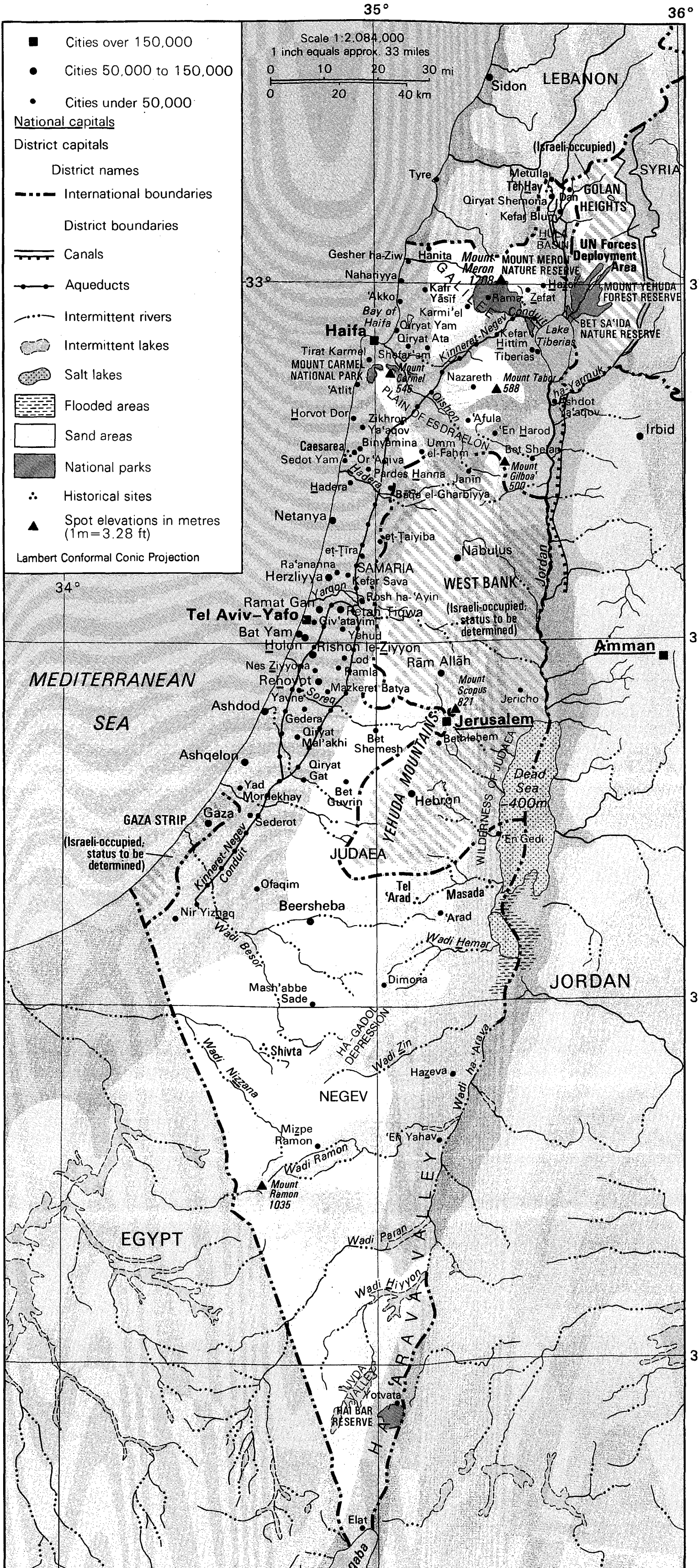
Urban settlement. The great majority of the Jewish population is urban. With the increasing mechanization and efficiency of Jewish agriculture, the proportion of people living on the land has been decreasing. As a result of industrial and service sector development, the two large conurbations of Tel Aviv–Yafo and Haifa and of the city of Jerusalem contain about one-fourth of the total population. Great efforts have been made by the authorities to prevent overconcentration of population in these areas. In both the north and south, new towns have been built whose populations consist largely of new immigrants. These towns serve as centres of regional settlement or else fulfill special economic tasks such as the manufacture of textiles, clothing, or machinery.

The major urban centres inhabited by Arabs include cities and towns with both Arab and Jewish populations, such as Jerusalem, Haifa, ‘Akko, Lod (Lydda), Ramla, and Yafo, and towns with entirely Arab populations, such as Nazareth. In towns with both Arab and Jewish populations, many of the former differences in the way of life of the two communities are diminishing, even though Arabs and Jews usually live in different quarters.

THE PEOPLE

Groups historically associated with the contemporary country. Jews constitute more than four-fifths of the total population, Muslims about one-eighth, Christians a small percentage, and Druzes and others the remainder.

Jews. In origin, as well as in physical features, the Jewish population lacks uniformity. Immigrants differed in racial origin and culture and brought with them languages and customs from a variety of countries. Consciousness of geographic origin and descent is, however, gradually being superseded by a national consciousness, especially among the young. Religious Jewish groups immigrating to Israel generally continue to pray in the synagogues of their respective communities. The two main religious



MAP INDEX

Political subdivisions

Central	32 05 N 34 55 E
Haifa	32 35 N 35 00 E
Jerusalem	31 45 N 35 00 E
Northern	32 50 N 35 20 E
Southern	30 40 N 34 50 E
Tel Aviv	32 05 N 34 48 E

Cities and towns

ʿAfula	32 36 N 35 17 E
ʿAkko	32 55 N 35 05 E
ʿArad	31 15 N 35 13 E
Ashdod	31 49 N 34 39 E
Ashdot Yaʿaqov	32 40 N 35 35 E
Ashqelon	31 40 N 34 35 E
ʿAtlit	32 41 N 34 56 E
Bāqa el-Gharbiyya	32 25 N 35 03 E
Bat Yam	32 01 N 34 45 E
Beersheba (Beʾer Shevaʿ)	31 14 N 34 47 E
Bet Guvrin	31 36 N 34 54 E
Bet Sheʾan	32 30 N 35 30 E
Bet Shemesh	31 45 N 35 00 E
Binyamina	32 31 N 34 57 E
Dan	33 14 N 35 39 E
Dimona	31 04 N 35 02 E
Elat	29 33 N 34 57 E
ʿEn Gedi	31 27 N 35 23 E
ʿEn H̱arod	32 33 N 35 23 E
ʿEn Yahav	30 38 N 35 11 E
Gedera	31 49 N 34 46 E
Gesher ha-Ziw	33 02 N 35 06 E
Givʿatayim	32 04 N 34 48 E
H̱adera	32 26 N 34 55 E
Haifa (H̱efa)	32 50 N 35 00 E
H̱anita	33 05 N 35 10 E
Haẕeva	30 48 N 35 15 E
H̱aẕor	32 59 N 35 33 E
H̱efa, see Haifa	
Herzliyya	32 10 N 34 51 E
H̱olon	32 01 N 34 46 E
H̱orvot Dor	32 37 N 34 55 E
Jerusalem (Yerushalayim)	31 46 N 35 14 E
Kafr Yāsīf	32 57 N 35 10 E
Karmiʾel	32 55 N 35 18 E
Kefar Blum	33 10 N 35 36 E
Kefar H̱ittim	32 48 N 35 30 E
Kefar Sava	32 10 N 34 54 E
Lod	31 58 N 34 54 E
Mashʾabbe Sade	31 00 N 34 47 E
Mazkeret Batya	31 51 N 34 50 E
Metulla	33 16 N 35 35 E
Miẕpe Ramon	30 36 N 34 48 E
Nahariyya	33 00 N 35 05 E
Nazareth (Naẕerat)	32 42 N 35 18 E
Nes Ẕiyyona	31 55 N 34 48 E
Netanya	32 20 N 34 51 E
Nir Yizẖaq	31 14 N 34 22 E
Ofaqim	31 17 N 34 37 E
Or ʿAqiva	32 30 N 34 55 E
Pardes H̱anna	32 28 N 34 58 E
Petaẖ Tiqwa	32 05 N 34 53 E
Qiryat Ata	32 48 N 35 06 E
Qiryat Gat	31 36 N 34 46 E
Qiryat Malʾakhi	31 44 N 34 44 E
Qiryat Shemona	33 13 N 35 34 E
Qiryat Yam	32 51 N 35 04 E
Raʿananna	32 11 N 34 53 E
Rama	32 56 N 35 22 E
Ramat Gan	32 05 N 34 49 E
Ramla	31 55 N 34 52 E
Reẖovot	31 54 N 34 49 E
Rishon le-Ẕiyyon	31 58 N 34 48 E
Rosh ha-ʿAyin	32 06 N 34 57 E
Safad, see Ẕefat	
Sederot	31 31 N 34 35 E
Sedot Yam	32 29 N 34 53 E
Shefarʿam	32 48 N 35 10 E
Ṭaiyiba, eṭ	32 16 N 35 01 E
Tel Aviv–Yafo	32 04 N 34 46 E
Tiberias (Teverya)	32 47 N 35 32 E
Ṭīra, eṭ-	32 14 N 34 57 E
Tirat Karmel	32 46 N 34 58 E
Umm el-Faẖm	32 31 N 35 09 E
Yad Mordekhay	31 35 N 34 33 E
Yavne	31 53 N 34 45 E
Yehud	32 02 N 34 53 E
Yerushalayim, see Jerusalem	
Yotvata	29 53 N 35 03 E
Ẕefat (Safad)	32 58 N 35 30 E
Zikhron Yaʿaqov	32 34 N 34 57 E

Physical features and points of interest

Aqaba, Gulf of	29 15 N 34 45 E
ʿArava, Wadi ha-	30 58 N 35 24 E
ʿArava Valley, ha-	30 10 N 35 10 E
Besor, Wadi	31 28 N 34 22 E
Bet Saʿida Nature Reserve	32 52 N 35 38 E
Caesarea (H̱orbat Qesari), *historical site*	32 30 N 34 53 E
Carmel, Mount (Har Karmel)	32 44 N 35 02 E
Dead Sea (Yam ha-Melaẖ)	31 30 N 35 30 E
Esdraelon, Plain of (ʿEmeq Yizreʿel)	32 36 N 35 14 E
Gadol Depression, ha-	30 56 N 34 59 E
Galilee (ha-Galil), *region*	32 54 N 35 20 E
Galilee, Sea of, see Tiberias, Lake	
Gilboaʿ, Mount	32 29 N 35 25 E
H̱adera, *river*	32 27 N 34 53 E
Hai Bar Reserve	29 50 N 34 58 E
Haifa, Bay of (Mifraẕ H̱efa)	32 53 N 35 03 E
H̱emar, Wadi	31 08 N 35 22 E
H̱iyyon, Wadi	30 12 N 35 07 E
H̱ula Basin	33 08 N 35 37 E
Jordan (ha-Yarden), *river*	31 46 N 35 33 E
Judaea (Yehuda), *region*	31 35 N 35 00 E
Judaea, Hills of, see Yehuda Mountains	
Judaea Mountains, see Yehuda Mountains	
Karmel, Har, see Carmel, Mount	
Kinneret, Yam, see Tiberias, Lake	
Kinneret–Negev Conduit	32 52 N 35 32 E
Masada (H̱orvot Meẕada), *historical site*	31 19 N 35 21 E
Mediterranean Sea	32 30 N 34 30 E
Melaẖ, Yam ha-, see Dead Sea	
Meron, Mount	33 00 N 35 25 E
Meẕada, H̱orvot, see Masada	
Midbar Yehuda, see Wilderness of Judaea	
Mount Carmel National Park	32 42 N 35 03 E
Mount Meron Nature Reserve	32 59 N 35 26 E
Mount Yehuda Forest Reserve	32 56 N 35 43 E
Negev, *region*	30 30 N 34 55 E
Niẕẕana, Wadi	30 57 N 34 23 E
Paran, Wadi	30 24 N 35 10 E
Qesari, H̱orbat, see Caesarea	
Qishon, *river*	32 49 N 35 02 E
Ramon, Mount	30 30 N 34 38 E
Ramon, Wadi	30 36 N 34 55 E
Samaria, *region*	32 15 N 35 10 E
Shivta (Subeita), *historical site*	30 53 N 34 38 E
Soreq, *river*	31 56 N 34 42 E
Subeita, see Shivta	
Tabor, Mount (Har Tavor)	32 41 N 35 23 E

Tel 'Arad, *historical site*	31 17 N 35 08 E
Tel Ḫay, *historical site*	33 14 N 35 35 E
Tiberias, Lake (Yam Kinneret)	32 48 N 35 35 E
'Uvda Valley	29 57 N 34 57 E
Wilderness of Judaea (Midbar Yehuda), *region*	31 30 N 35 18 E
Yarden, ha-, see Jordan	
Yarmuk, ha-, *river*	32 38 N 35 34 E
Yarqon, *river*	32 06 N 34 47 E
Yehuda, see Judaea	
Yehuda Mountains (Judaea Mountains)	31 40 N 35 10 E
Yizre'el, 'Emeq, see Esdraelon, Plain of	
Z̮in, Wadi	30 57 N 35 19 E

groupings are formed by those who follow the Ashkenazic rite (of Jews from central and eastern Europe and their descendants in other parts of the world) and those who follow the Sefardic and Oriental rite (of Jews from the Mediterranean region and from the Middle and Far East). Thus there are traditionally two chief rabbis in Israel, one Ashkenazi and one Sefardi. Religious Jewry in Israel constitutes a significant and articulate section of the population. Disputes often arise between this group and a strong movement that seeks to prevent religious bodies and authorities from dominating national life.

Muslims. The largest religious minority group is the Muslims, who constitute more than two-thirds of the Arab population. Practically all of the Muslims adhere to the Sunnī rite. A minority are peasants, and most Muslims now live in towns. Like all other religious communities, the Muslims enjoy considerable autonomy in dealing with matters of marriage and divorce and have separate religious courts. The state supervises their religious institutions. Among the Muslims are the Bedouin, most of whom live in the Negev with the rest living in Galilee. New economic opportunities, and the fact that the borders with neighbouring Arab countries are closed, have encouraged the Bedouin to adopt a more sedentary mode of life.

Christians. Most Christians are town dwellers, and the majority speak Arabic. Christian communities exercise autonomy in religious and communal affairs. The Greek Orthodox and Greek Catholic (Melchite) are the largest of these, most of which are headquartered in Jerusalem. Apart from the Greek Orthodox patriarchate in Jerusalem, the Christian churches are dependent to a degree on supreme hierarchs abroad. These communities include Roman Catholics and Uniates (Melchites, Maronites, Chaldean Catholics, Syrian Catholics, and Armenian Catholics).

Jerusalem is also the seat of two Russian Orthodox missions: one represents the Moscow patriarchate and the other, the Russian church in exile. The Evangelical, Episcopal, and Lutheran churches are small and primarily Arabic-speaking.

The Druze community

Druzes. The Druzes, who live in villages in Galilee and on Mt. Carmel and who maintain excellent relations with the Israeli majority, have since 1957 constituted a separate Arab community. Most of the Druzes are agriculturists who preserve their traditional way of life. They pay homage to Jethro (Moses' father-in-law), whose putative grave is near Kefar Ḫittim in Galilee. The Druzes serve in the army.

Bahā'īs. The Bahā'ī faith is the only religion other than Judaism whose world centre is in Israel. A shrine, an archives building, and an administrative centre are located on Mt. Carmel in Haifa. There are a few hundred adherents, most of whom are employed at the centre in Haifa.

Circassians. The Circassians, who are Sunnī Muslims, emigrated from the Caucasus in the 1870s. They number a few thousand and live in villages in Galilee, preserving their language and their traditions. Older Circassians speak Arabic, but the younger generation speaks Hebrew. The men serve in the Israeli Army.

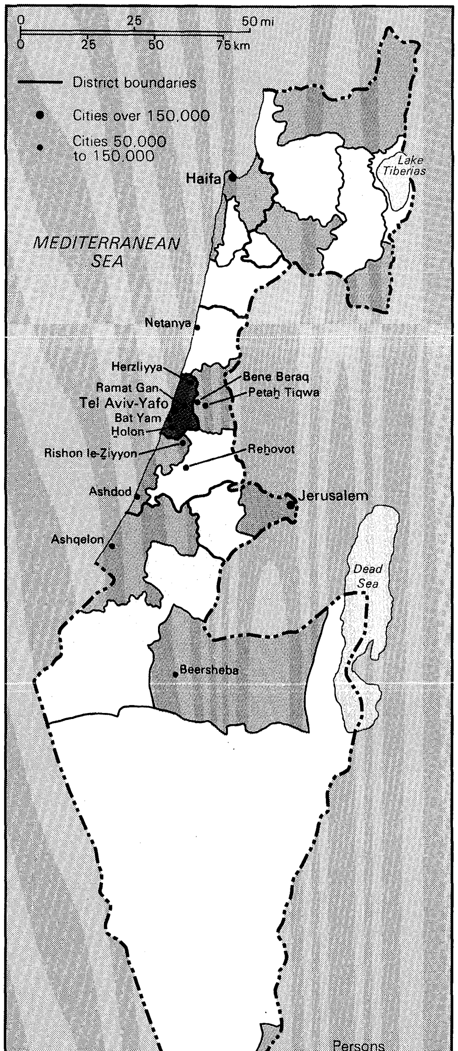
Samaritans. About half of the few hundred surviving members of the Samaritan community live near Tel Aviv in the town of Ḫolon. They preserve their separate religious and communal organization but participate in national life as part of the Jewish section of the population.

Demography. In 1948 the Jewish population in Israel numbered about 650,000. Between 1948 and 1970 about 1,300,000 Jewish immigrants entered the country, and about 200,000 Jews left it, although some later returned. Of the Jewish community the largest proportion was born in Israel, followed by those born in Europe and America, Africa, and elsewhere in Asia. In 1948 about 155,000 Arabs remained. Growth in the Arab population (not including those living in East Jerusalem) has been relatively large, including refugees who have returned. The average lifespan is one of the highest in the world.

THE ECONOMY

The increase in the Jewish population was the most distinctive cause of the rapid rise in the gross national product after 1948. Although most immigrants had to change occupations, a nucleus of highly skilled labour facilitated economic expansion. The establishment and rapid growth of institutions of higher learning and research helped increase the nation's potential. Large amounts of capital arrived in the form of money involving no financial obligation by the state. This included gifts from world Jewry, reparations from the Federal Republic of Germany for the persecution of Jews by Adolf Hitler, grants-in-aid from the U.S. government, and capital brought in by immigrants. It has been supplemented by loans and commercial credits and by foreign investment.

The goals of economic policy are continued economic growth, the reinforcement of a competitive capacity, and



Population density of Israel.

further integration of Israel's economy with the world economy. Progress toward these goals has been made under difficult conditions, which have included a rapid increase of population; a boycott and a blockade by the neighbouring Arab countries except, from 1979, Egypt; heavy expenditure on defense; a scarcity of natural resources, including water; a high standard of living; inflation; and a restricted home market that limits the economies of methods of mass production.

Resources. *Mineral resources.* Mineral resources include potash, bromine, and magnesium, the last two of which are obtained from the waters of the Dead Sea; copper ore, which is located in the ʿArava Valley; phosphates and small amounts of gypsum in the Negev; and some marble in Galilee. There are oil deposits in the northern Negev and south of Tel Aviv and deposits of natural gas also in the northern Negev and northeast of Beersheba. Limited exploitation of oil began in the 1950s.

Electric power resources. Electricity is principally generated from thermal stations. The electricity industry is nationalized, and the government has encouraged intensive rural electrification; electricity for agriculture and industry is provided at favourable rates.

Atomic energy. The Israel Atomic Energy Commission was established in 1952. It has undertaken a comprehensive survey of the country's natural resources and trains scientific and technical personnel. An atomic reactor was constructed with U.S. assistance south of Tel Aviv, and a second one was built with French help in the Negev.

Agriculture, forestry, and fishing. The expansion in the amount of irrigated land has been a major factor in raising the value of agricultural production. There has also been a great expansion in the cultivation of citrus and of such industrial crops as peanuts (groundnuts), sugar beets, and cotton, as well as of vegetables and flowers. The number of milk cows has increased greatly. Agriculture also has become greatly mechanized.

Scarcity of water

The main agricultural problem is scarcity of water. Water from the Jordan and Yarqon rivers and from Lake Tiberias (Sea of Galilee) is diverted by pipeline to arid areas in the south. Because of utilization of practically all of the country's potential water resources, further development of agriculture involves intensifying the yield from land already irrigated or obtaining more water by cloud seeding, reducing the amount of evaporation, desalinizing sea water, and diverting water from the occupied territories.

Because only a limited quantity of fish is available off Israel's Mediterranean and Red Sea coasts, Israeli trawlers sail to the rich fishing grounds off the Ethiopian coast and engage in deep-sea fishing in the Atlantic. Inland, fishpond production meets much of the domestic demand.

Industry. *Mining and quarrying.* The mining industry supplies local demands for fertilizers, detergents, and drugs, and also produces some exports. The Timna Copper Mines near Elat exported copper ore until they were closed in 1976; a plant in Haifa produces potassium nitrate and phosphoric acid, both for local consumption and for export. Products of the Haifa Oil Refineries include polyethylene and carbon black, which are used by the local tire and plastic industries. The electrochemical industry also produces food chemicals and a variety of other commodities. Oil pipelines run from the port of Elat to the Mediterranean. Israel has producing oil wells but continues to import some oil.

Manufacturing. Industrial growth has been especially rapid in electronics, weapons, transportation, machinery, and metals. The largest share of manufacturing output is accounted for by the food industry, after which the principal products are textiles, chemicals, and metals. The diamond-cutting and polishing industry ranks among the largest in the world. One of the largest industrial enterprises is Israel Aircraft Industries. Industries manufacturing military supplies and equipment have expanded considerably since the 1967 war—a circumstance that has stimulated the development of the electronics industry. The largest share of Israel's industrial exports are marketed in Europe, followed by the United States and Canada. The great majority of industries are privately owned.

Industrial expansion has been stimulated by the growth of local demand, which in turn has resulted from the growth of population and from the rise in the standard of living. Industry enjoys a high degree of protection against competitive imports. The government also assists industry by making loans available from the development budget at low rates of interest. The main limitations experienced by industry are the scarcity of raw materials and sources of energy, and the restricted size of the local market.

Tourism and shipping. Tourism is a growing industry and is one of the largest sources of foreign exchange. Shipping is a vital factor both in the economy and in communications with other countries. As a result of the closing of the land frontiers following the Arab blockade of Israel, shipping has played a major role in the transportation of supplies. It formerly was used to provide passage to Israel for immigrants. Israel's access routes to both the Atlantic and Indian oceans have stimulated a continuous growth of its merchant fleet.

Finance. Israel has commercial (deposit) banks and cooperative credit institutions, mortgage and investment credit banks, and other financial institutions that are supervised by the Central Bank of Israel. The banking system shows a high degree of specialization; commercial banks are, in general, restricted to short-term business. Medium- and long-term transactions are handled by institutions established to cater to the investment needs of the separate elements of the economy: agriculture, industry, housing, and shipping. These institutions are either fully owned by the government or are owned jointly by the banks and the government.

Israel has a managed floating currency. The Israeli shekel consists of 100 agorot. There have been numerous devaluations of Israeli currency, which led to the introduction of the shekel in February 1980 to replace the Israeli pound at the rate of 1 shekel for every 10 pounds. The Central Bank of Israel issues currency and acts as the government's sole fiscal and banking agent. Its major function is to regulate the money supply and short-term banking credit.

Trade. Imports are mainly raw materials, including rough diamonds, and capital goods. Exports include a variety of light industrial products, textiles, polished diamonds, fertilizer and chemical products, and agricultural produce (mainly citrus fruits).

Import–export balance

The central problem of foreign trade is the large and persistent deficit resulting from the imbalance of imports over exports. Free access to foreign markets is, therefore, vital for the further expansion of the economy. The fact that Israel is not a member of any of the regional economic groupings represents a considerable handicap. In 1964 Israel concluded a special agreement with the European Economic Community (EEC) and in 1975 attained a new agreement with the EEC. An attempt has also been made to ameliorate foreign trade problems by participating actively in the General Agreement on Tariffs and Trade (GATT), a specialized agency of the United Nations.

Administration of the economy. The management of Israel's economy has been conditioned by many dynamic, powerful, and often contradictory factors. The large imports of capital that have been derived from public and semipublic sources have passed through government channels or through public organizations. This has resulted in the enlargement of the sector of the economy that includes public and semipublic enterprises. At the same time, the government's policy has been directed toward liberalizing the economy. The socioeconomic structure of the economy is, therefore, diversified. The governmental, cooperative, and private sectors coexist in an economy that is subordinated to the broad objectives of state policy.

Taxation. The rates of taxation are among the highest in the world. Income, customs and excise, land, and luxury taxes are the main sources of revenue. The distribution between direct and indirect taxation has been altered over the years. Since the late 1950s the proportion of indirect taxation has increased.

Trade unions and employer associations. The General Federation of Labour in Israel (the Histadrut) is the largest labour union and the largest voluntary organization in the country. Since 1960 Arab workers have been admitted with full membership rights. The National Labour Federation is another major labour organization. The Manufacturers' Association of Israel and the Farmers' Union represent a large number of the country's employers.

Transportation. Road transport is of more significance than rail in internal communications. A combined road and rail system extends to the port of Elat.

Deepwater ports

Three modern deepwater ports—Haifa and Ashdod on the Mediterranean, and Elat on the Red Sea—are maintained and developed by the Israel Ports Authority.

The international airport at Lod is the country's largest. Regular flights are maintained by several international airlines, with EL AL Israel Airlines Ltd., Israel's national airline, accounting for the largest share of the traffic. Domestic aviation, operated by Arkia Israeli Airlines Ltd., has developed greatly in the late 20th century, with both planes and helicopters being used. Jerusalem (Ataroth), Tel Aviv (Sdeh-Dov), Elat, Rosh Pinna, and Haifa airfields serve the country's domestic air traffic.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. *Constitutional framework.* Israel is a democratic republic with a parliamentary system of government. It has a strong cabinet, a multiparty system with two major parties, and a marked tendency toward political and administrative centralization. Israel does not have a formal written constitution. The foundation on which the system of government has been built is composed of legislation, administrative acts, and parliamentary practice.

The Knesset, or assembly, is a 120-member, single-chamber legislature that is elected every four years. In its internal organization and parliamentary procedure it has followed continental rather than Anglo-Saxon practices. Members exercise important functions in standing committees. Hebrew and Arabic, the country's two official languages, are used in all proceedings.

The president, who is the head of state, is elected by the Knesset for a five-year term, which can only be renewed once. The president has no veto powers and exercises only ceremonial functions.

The cabinet is the main policy-making body. Its members may be, but need not be, members of the Knesset. Following a general election or the resignation of a government, the president, after consultation with representatives of all parties, entrusts a member of the Knesset with the task of forming a cabinet. The prime minister is the leading figure in the cabinet and the government.

The state controller, an independent officer appointed by and responsible only to the Knesset, is the auditor of the government's financial transactions and is empowered to inquire into the efficiency of its activities.

The civil service is gradually developing into a politically neutral and professional body; previously, it tended to support the party in power. The extensive functions of the government have tended to result in a growing bureaucracy.

Local and regional government. Administratively, the country is divided into six districts—the Central, Jerusalem, Haifa, Northern, Southern, and Tel Aviv districts—and into 13 subdistricts. There are three types of local government councils—municipalities, local councils (for smaller settlements), and regional rural councils. The bylaws of the councils, as well as their budgets, are subject to approval by the Ministry of the Interior.

Israeli-occupied Arab territories. After the 1967 war Arab territories occupied by Israeli forces were placed under military administration. These territories included Jordanian territory on the west bank of the Jordan River (the West Bank), the Gaza Strip, the Sinai Peninsula region of Egypt, and the Golan Heights region of Syria. East Jerusalem was also occupied by Israeli forces, and Israeli authorities took over administration of the city as a single municipality; in 1967 East Jerusalem and adjoining villages were incorporated into the State of Israel—an action that is disputed abroad. Israel completed its withdrawal of civilian settlers and military personnel from the Sinai Peninsula in April 1982. In April 1981 the Golan Heights, however, was in effect annexed to Israel by the extension of Israeli law to the area.

Elections

The political process. Elections, which are nationwide, are by universal, direct suffrage, with secret balloting. The system of election is by proportional representation. All resident Israeli citizens are enfranchised upon completion of their 18th year; candidates for election must be at least 21 years old. Similar conditions govern elections to local government bodies.

Israel's party system is complex and volatile; splinter groups are commonly formed and party alliances often change. Political parties are both secular and religious; the secular parties are Zionist and range in orientation from Marxist to capitalist.

Cabinets are based on coalitions of varying political composition. For almost thirty years after the creation of Israel, the dominant party was the Israel Labour Party, a moderate social-democratic and Zionist group. While in power it held a majority in the cabinet, the premiership, and the major ministries. In 1977 the Israel Labour Party suffered its first defeat when the Likud bloc, an alliance of the extreme nationalist Herut Party and several other parties, gained ascendacy in the Knesset.

Israeli citizens take an active interest in public affairs. The pattern of Israel's social and economic organization favours the participation in state and public affairs of both trade unions and employers' organizations. The Arab community is allowed to play a full role in national politics, as long as it does not forcefully oppose the basic concepts of Zionism.

Justice. Municipal, religious, and military courts exercise a jurisdiction almost identical with that exercised by such courts during the period of the Palestine Mandate before the birth of Israel. Regional labour courts were established in 1969. Matters of marriage and divorce are dealt with by the religious courts of the various recognized communities. Capital punishment has been maintained only for genocide and crimes committed during the Nazi period. The judges of the magistrates', district, and supreme courts are appointed by the president, and every judge holds office for life.

Law is derived from a variety of sources, including Ottoman and British legislation and precedent, religious court opinion, and Israeli parliamentary enactments. Special investigative panels have been formed on unusual occasions—such as the war of 1973 and the massacres of Palestinians in Israeli-controlled sectors of Beirut in 1982—to issue reports and allocate responsibility among political and military leaders.

The armed forces. The Israel Defense Forces (IDF) is an integrated organization controlling land, sea, and air forces. A special force (Nahal) combines military and agricultural training and also engages in the establishment of new defense settlements on the borders. Youth Battalions deal with premilitary training of youth both in and out of school. There is compulsory military service for men and for women; Arabs, both Muslim and Christian, are exempted, but the IDF includes a minorities unit in which Druze, Circassian, Bedouin, and Christian Arabs may serve. After the period of conscription men and childless women undergo a period of regular reserve training.

The IDF is based essentially on the reserve service of the population, and thus continues to be a popular militia rather than a professional army. Civilian-military relations are based on the subordination of the army to civilian control. The defense of the country is based upon a regional defense system. The basic unit in the IDF is the brigade group. Any number of brigade groups can be combined under the command of divisional groupings in time of war. The rank of *rav-alluf* (lieutenant general) is held only by the chief of staff of the IDF. He is the senior military authority commanding all armed forces and is appointed by the government on the recommendation of the minister of defense.

The police services are a centralized agency under a ministry of the Interior. They are controlled by national headquarters and commanded by an inspector general. Prisons are administered by the Ministry of Police and are linked to a system for the rehabilitation of prisoners into society on their release.

Education. By enactment, education is obligatory and free for children between the ages of five and 15 and free, but not compulsory, for those 16 and 17. Young people between the ages of 14 and 18 who have not completed schooling are obliged to attend special classes. Parents may choose whether the children receive state lay education or state religious education. The school syllabus includes radio and television lessons in both Hebrew and Arabic. Special attention is given to agricultural and technical training. Adult education for immigrants assists them in their cultural integration. There are teachers' training colleges, including two for Arabs. The Ministry of Education, in conjunction with the military authorities, also has assumed responsibility for education in Israeli-occupied Arab territories.

Institutions of higher learning

In addition to the Hebrew University of Jerusalem (opened in 1925) and the Technion–Israel Institute of Technology in Haifa (opened in 1924), Israel has several institutions of higher learning that have been founded since 1948. These include the Weizmann Institute of Science in Reḥovot, the universities of Tel Aviv, Bar-Ilan, and Haifa, and Ben Gurion University of the Negev. Everyman's University, in Tel Aviv, was opened in 1974. The language of instruction in the universities is Hebrew, while the teaching system represents a mixture of European and American methods. Academic freedom in the universities is protected by Israeli law.

Health and welfare. The Ministry of Health maintains its own public and preventive services and supervises those of nongovernmental institutions. There are many voluntary organizations in the country dealing with first aid, children's health, and with the aged, crippled, and blind.

The Ministry of Social Welfare controls the service bureaus that deal with family, youth, and community welfare, as well as with rehabilitation of the handicapped. Social insurance is compulsory.

Wages and cost of living

Wage policy is determined by the trade unions and the employers' associations but is also influenced by the government, which not only applies moral suasion to both groups but is itself the largest employer in the country. Wage policy finds expression in cost-of-living allowance agreements, which are intended to safeguard the real value of wages against rises in consumer prices, and in wage agreements between employers and employees.

Israel's Jewish community is a dynamic society still in the process of social and economic formation. The most significant social divisions among Israelis are of a predominantly community nature. Eastern or Oriental Jews (Sefardim) tend to be poorer, less educated, and underrepresented in higher offices as compared to Western Jews (Ashkenazim). The social and economic divisions within the Arab community (especially among Muslims and Druzes) and between Arabs and Jews are still strong. Arabs are generally in the lower ranges of socioeconomic categories and consider themselves to be the victims of prejudice at the hands of Jewish Israelis. Changes in the late 20th century, however, have diminished some of the antagonisms between Jews and Arabs, as well as differences between urban and rural Arabs.

CULTURAL LIFE

The cultural milieu. Jews arriving from communities in many parts of the world have brought with them both their own cultural inheritance and aspects of individual majority cultures that they have absorbed over the centuries. The intermingling of the Ashkenazi, Sefardi, and Middle Eastern traditions has been of profound importance, although the arrival of immigrants from the Soviet Union has slowed the trend, common among immigrants from central Europe and America, toward creating a cultural synthesis embracing both East and West. There has been little cultural interchange between the Jewish and Arab sections of Israel's population, and the impact of Arab culture on Israeli cultural life has been insignificant. The revival of the Hebrew language has been of great importance. Jewish tradition, both religious and historical, and the Hebrew language together constitute the foundation of cultural life in Israel.

The state of the arts. The Israel Philharmonic Orchestra has an international reputation. Folk dancing and popular singing combine foreign elements with original creative manifestations. Different folk traditions, such as folk songs, musical instruments, and other expressions of popular culture, have been preserved mainly among the Oriental Jewish communities and among the rural Arab population. Painting and sculpture are still largely influenced by European schools, but local schools have begun to emerge. In literature and drama a concentration on themes of the Diaspora is giving way to an interest in national themes. Among Israel's most distinguished writers is Shmuel Yosef Agnon (1888–1970), who was awarded the Nobel Prize for literature in 1966.

Cultural institutions. In 1954 the Hebrew Language Academy was established as the supreme authority on all questions related to the language and its usages. The Israel Academy of Sciences and Humanities was founded in 1960. There are several hundred libraries in the country. The Jewish National and University Library in Jerusalem is particularly notable. Habima, Israel's national theatre, was founded in Moscow in 1917 and moved to Palestine in 1932. There are a number of other theatres in the country, some of them in the kibbutzim. There are many art galleries and museums; foremost among them is the Israel Museum in Jerusalem, which also houses part of the archaeological collection of the government's department of antiquities. Archaeological activities are initiated by the government and by Israeli academic institutions, as well as by foreign archaeological organizations. The discovery of the Dead Sea Scrolls in 1947 gave a powerful stimulus to biblical and historical research.

Archaeological activities

Press and broadcasting. Tel Aviv is the centre of newspaper publishing. Newspapers are often associated with a political party. Although most newspapers are written in Hebrew, there is a considerable circulation of papers published in Yiddish, English, German, Hungarian, French, Bulgarian, and Romanian. There are hundreds of other periodicals, of which more than half are in Hebrew.

Broadcasting is vested in a broadcasting authority whose members are appointed by the president. Languages of broadcast include Hebrew, Arabic, English, French, Yiddish, Russian, Georgian, Portuguese, Hungarian, Romanian, Ladino (a Spanish dialect of the Sefardic Jews), Moghrabit, (a dialect spoken by Jews in the Maghrib), and Persian. Television, which was introduced in 1966, consists of Hebrew and Arabic programs. There is also an educational television service. Radio programs are broadcast to many foreign countries. For statistical data, see the "Britannica World Data" section in the current *Britannica Book of the Year.* (E.E./W.L.O.)

History

ZIONISM

The Jewish nationalist movement, Zionism, has had as its goal the creation and support of a Jewish national state in Palestine, the ancient homeland of the Jews, which is called in Hebrew *Eretz Israel* (Land of Israel). Though Zionism originated in eastern and central Europe in the latter part of the 19th century, it is in many ways a continuation of the ancient and deep-felt nationalist attachment of the Jews and of the Jewish religion to Palestine, the promised land where one of the hills of ancient Jerusalem was called Zion. This attachment to Zion continued to inspire the Jews throughout the Middle Ages and found its expression in many important parts of their liturgy.

Early history. At the end of the Middle Ages a number of "messiahs" came forward with the claim to lead the Jews back to Palestine, and they were generally received with great enthusiasm by their fellow Jews. The most important of these messiahs were David Reubeni and his disciple Solomon Molcho in the first part of the 16th century and in the 17th century Shabbetai Tzevi, who proclaimed himself the Messiah in Turkey in 1648. Most European Jews believed in Shabbetai's mission, and many continued to do so even after he had become a Muslim.

The age of the Enlightenment in the second half of the 18th century, with its growth of religious toleration and its general universal and liberal ideas, laid the foundations in western Europe and North America for the emancipation of the Jews and their participation as citizens in the

Moses Mendelssohn

life of the nations in the midst of which they lived and whose members they became. The consequence was less emphasis among the Jews on traditional religious attitudes and more on assimilation of Western secular culture. This movement emerged also in Germany, where Moses Mendelssohn (1729–86), a philosopher and a friend of the great German writer Gotthold Ephraim Lessing, emphasized the spiritual and universal aspects of Judaism. He and a group of like-minded fellow Jews, most of them residents of Berlin and Königsberg in Prussia, wished to win the Jews over to modern Western civilization. The younger Jewish generation gladly seized the opportunity of intellectual enrichment and civic freedom that the new movement, originally called by the Hebrew name Haskala (Enlightenment), offered. Many went the way of complete assimilation, including the abandonment of the faith of their fathers. Others found their place as citizens of the Jewish faith in the new liberal and egalitarian societies emerging in the 19th century in western Europe and North America. Still others applied the new scholarship learned from Western civilization to a study of the Jewish past and produced, especially in Germany, works of lasting value in the rediscovery and reinterpretation of the ancient heritage. Some wealthy Jews—among them Sir Moses Montefiore and the Rothschild family—tried to help their coreligionists in less fortunate lands, especially in eastern Europe and in the Middle East, by establishing schools and introducing them to agriculture and to various trades. For reasons of religious piety, a small number of Jews, supported by donations from outside, settled in Palestine.

The interest in a return of the Jews to Palestine was kept alive in the first part of the 19th century more by Christian millenarians, especially in Great Britain, than by Jews themselves. Among the few Jews pleading then for a Jewish settlement or state was the American Mordecai Manuel Noah (1785–1851), who in 1813 became U.S. consul in Tunis and later high sheriff and surveyor of the port of New York. In 1825 he acquired Grand Island in the Niagara River and invited the Jews of the whole world to create a Jewish state, Ararat, there. In 1844 he pleaded with the Christian world in *Discourse on the Restoration of the Jews* to help the Jews resettle in Palestine. More important but not more successful were the attempts by Lord Shaftesbury, Sir Laurence Oliphant, and others in Great Britain to create a Jewish state in Palestine. Some political writers thought of a Jewish state in the Holy Land as a means of assuring the overland route to India. Others were inspired by religious or mystic ideas, "anxious to fulfill the prophecies and bring about the end of the world," as was the eccentric Oliphant. He was accompanied on one of his visits to the Middle East by Naphtali Herz Imber (1856–1909), a Hebrew poet of Polish origin, famous in the history of Zionism as the author of "ha-Tiqva" ("The Hope"), which became the Zionist national anthem, and of "Mishmar Ha-Yarden" ("The Watch on the Jordan"), a popular nationalist song.

These early sympathies with the return to Zion in the English-speaking world found their best literary portrayal in George Eliot's novel *Daniel Deronda* (1876). A German socialist, Moses Hess (1812–75), influenced by the example set by the unification of Italy, gave the first theoretical expression to Zionism among Jews, *Rome and Jerusalem* (1862; Eng. trans. 1918). This short book, which contained many thoughts later widely accepted by Jewish nationalists, combined ethical socialism, fervent nationalism, and religious conservatism. Hess believed that the historical ideal of the Jewish people could be realized only in their own historic homeland. He insisted that a moral and spiritual regeneration must precede the settlement there. He hoped that France, which he venerated as the home of the Revolution, would protect the Jewish settlement because it would wish to see the bridge across the Middle East held by a friendly people. Hess's book attracted no attention when it appeared. Only decades later was it rediscovered by the Zionist movement, which had by then developed in eastern and central Europe.

Emigration of Russian Jews to Palestine

The Love of Zion movement. Whereas in western Europe the Jews became in the 19th century an integral part of the nations whose citizens they were and fully adopted native language and culture, the Jews in eastern Europe, then identical with the Russian Empire, lived as a separate community with their own language, Yiddish, their own civilization, and their own economic structure. They did not enjoy political or legal equality with the Russians. Further, under the reactionary regime that set in after the reform age of Tsar Alexander II (1855–81), all hopes for Jewish emancipation were dashed. A wave of bloody pogroms, instigated or tolerated by the government, threatened Jewish lives and property. As a result large-scale emigration to western Europe and to the United States started. A very small trickle of Jewish youth from Russia also went to Palestine and founded there the first agricultural settlements. In 1882 Leo Pinsker (1821–91), a physician in Odessa, published *Auto-Emancipation* (Eng. trans. 1884), an appeal in German to the western European Jews to save the Jewish people from persecution and the misery of dispersion. He applied the ideas of 19th-century European nationalism and secularism to the Jews and propagated the necessity of concentrating them territorially, in Palestine or elsewhere. He found no echo among the western European Jews. But in Russia a small group, which took the name Ḥovevei Ẓiyyon (Lovers of Zion), gathered around him and formed a committee in Odessa to promote the settlement of Jewish farmers and artisans in Palestine. Though these early settlements were able to survive only with the help of Baron Edmond de Rothschild of Paris, they laid the foundations of practical Jewish colonization in Palestine.

The most prominent among these early Zionists was Asher Ginzberg (1856–1927), whose essays written under the pen name Aḥad Haʿam (One of the People) became classics of the modern Hebrew language that they helped to create. Aḥad Haʿam denied that the majority of the Jewish people could be settled in Palestine; the smallness of the country and the fact that it was inhabited by a large native population seemed to him to be insurmountable obstacles. But though Palestine according to him could not become a Jewish state, he believed in the creation of a Jewish cultural centre there, a place for the regeneration of Judaism, from which spiritual influences would radiate into all the many lands where Jews continued to live and which would awaken in their hearts a true "love of Zion."

This early Hebrew Renaissance in Russia also produced several great Zionist poets, among them Hayyim Nahman Bialik and Saul Tchernichowsky (1875–1943). At the same time the Yiddish language, an eastern European derivation from medieval German not to be confused with the ancient Semitic Hebrew tongue, was raised from a people's vernacular to a medium of art by a number of writers, the first of whom was S.J. Abramovich (1835?–1917), who wrote under the pen name Mendele Mokher Sefarim (Mendele the Itinerant Bookseller).

Theodor Herzl

Political Zionism. A new impetus was given to Zionism by Theodor Herzl, an Austrian journalist. In the multinational pre-World War I empire of the Habsburgs with its violent nationality struggles, the Jews found themselves in a difficult position. Except for eastern Europe, anti-Semitism was nowhere as strong as among the Austrian Germans, especially in Vienna where Herzl lived. The Dreyfus affair of the 1890s and its attendant outburst of anti-Semitism, which Herzl witnessed in 1895 as a newspaper correspondent in Paris, caused him to write a pamphlet, *Der Judenstaat* (*The Jewish State,* 1896). He regarded assimilation as most desirable but, in view of anti-Semitism, impossible. Against their own wishes, he believed, the Jews were forced by pressure from outside to form a nation. As such they could lead a normal existence only through concentration in one territory. Herzl had no living ties with Jewish and Hebrew traditional values. He never desired the rebirth of Hebrew as the Jewish national language. In his novel *Altneuland* (1902) he depicted the future Jewish life in Palestine in terms of life as he had known it among the liberal, assimilated central European Jews. In this novel, his testament to the movement, he rejected all narrow nationalism and demanded above all brotherly consideration for, and closest cooperation with, the Palestinian natives in a common homeland.

Herzl molded Zionism into a political movement of worldwide significance. He became its indefatigable organizer, propagandist, and diplomat. He convened in August 1897 the first Zionist Congress at Basel, Switzerland, which drew up a constitution for the movement. His friend Max Nordau participated in drawing up the Basel program of the movement, which proclaimed that "Zionism strives to create for the Jewish people a home in Palestine secured by public law."

To that end the movement was to promote on suitable lines the colonization of Palestine by Jewish rural and industrial workers; to reorganize the whole of Jewry by means of appropriate local and international institutions, in accordance with the laws of each country; to strengthen and foster Jewish national sentiment; and to obtain government consent where necessary to the attainment of the Zionist aims.

The centre of the movement was established in Vienna, where Herzl published the official weekly *Die Welt* ("The World"). The congresses met every year until 1901 and then every two years. Meanwhile Herzl entered into negotiations with the Turkish government in order to receive a charter establishing Palestine's autonomy, but the Turkish government rejected the proposals. Only in England did Herzl find sympathy, and for that reason he established the financial instruments of the movement in London. In 1903 the British government offered an area of 6,000 square miles (15,540 square kilometres) to the Zionist organization in the uninhabited highlands of Uganda. This offer led to violent controversy and even a split in Zionist ranks. A minority under the leadership of Israel Zangwill was willing to accept the offer. Members of this minority founded in 1905 the Jewish Territorial Organization with the aim of finding an autonomous territory for those Jews who could not or did not wish to remain in the countries in which they lived. The majority of the Zionists, most of them from Russia, insisted on Palestine as the only field of activity for Zionism, and the seventh Zionist Congress in 1905 rejected any colonization outside Palestine and its neighbouring countries. In 1904 in the midst of this bitter debate Herzl, only 44 years old, died.

The Jewish Territorial Organization

Pre-World War I Zionism. With the death of Herzl the leadership moved from Vienna to Germany, first to Cologne and then to Berlin. Austrian and German Jews led the movement, but its mass strength came from Russia. At that time only a very small minority of the Jews were organized in the Zionist movement. There was much opposition to Zionism in Jewish ranks, based partly on the conviction that the Jews had to become, or were already, an integral part of the nations among whom they lived and to which they belonged; partly on religious orthodoxy, which expected the return to Palestine only under divine guidance and the strict application of traditional religious laws; and partly on the conviction that the Jewish people could exist as a distinct national group in the Diaspora (state of dispersion), especially in countries of Jewish mass settlement and Yiddish folk culture. Among the last group was the Bund, a Jewish socialist party, founded in Russia in 1897. Jewish socialists were also organized as Zionists under the name of Poale Zion (Workers of Zion). Their first world conference was held in The Hague in 1907.

Of greater importance was the question of the position of traditional orthodoxy in the Zionist movement. In 1902 a number of Russian Zionists founded Mizraḥi as the party of religiously orthodox Zionists who insisted on the strict observance of Jewish religious laws in Palestinian Jewish life. The ideal of Herzl, on the other hand, had been a modern secular movement. The Zionist organization declared itself neutral in matters of religion, but from time to time the insistence of the religious groups upon observance of Jewish religious precepts caused conflicts within the movement.

Though Zionism represented only a minority of Jews, and in the Western lands only a small minority, it was the only worldwide democratically organized part of Jewry. It developed an active propaganda through orators and pamphlets, created its own newspapers in many languages, and gave an impetus to what was called a "Jewish renaissance" in letters and the arts. At the same time the failure of the Russian Revolution of 1905 and the wave of pogroms and the repressive measures that followed it disillusioned many in regard to Jewish emancipation in eastern Europe. Again as in 1882, but in growing numbers, Russian Jewish youth emigrated to Palestine to live there as pioneers in newly founded agricultural settlements in which they hoped to realize their nationlist and socialist ideals. They fought against the employment of "foreign" (Arab) labour on Jewish settlements and insisted on the use of Hebrew as the spoken language.

Further emigration of Russian Jews to Palestine

The growth of the Jewish settlement in Palestine was due to the practical Zionists, who were opposed by the political Zionists, who insisted on the granting of a charter as an essential prerequisite for colonization. With the growing strength of the Young Turk nationalist movement in Turkey, especially after 1908, the prospects of obtaining a charter dimmed considerably. But in spite of small financial means, urban development and agricultural settlement among the Jews in Palestine made steady progress. In 1914 there were about 90,000 Jews in Palestine, where the large majority of the population was Arab. There were 43 Jewish agricultural settlements with 13,000 settlers, many of them supported by Baron Rothschild.

The situation changed with the outbreak of World War I. Zionist work in Palestine came to a standstill. Turkey and Britain were at war. An opportunity offered itself for political Zionism to reassert itself and to combine the old British sympathies for Zionism with the opportunities of political warfare. As a result the centre of the Zionist movement shifted from Germany, Turkey's ally, to London. The leadership passed to Jews of Russian origin living in London—among them Chaim Weizmann and Nahum Sokolow. As a result of the Russian Revolution of November 1917 and of the consequent civil war with its pogroms perpetrated by the White armies and because of the intensified nationalism of the various succession states of post-World War I Europe, great misery spread among eastern European Jews. From then on the financial and economic strength of Zionism came from Jews in the United States, and the masses of its adherents from 1920 to 1938 came from Poland.

The Balfour Declaration. Weizmann and Sokolow were instrumental in causing a letter to be written by Arthur James (later Lord) Balfour, then British foreign secretary, to Lord Rothschild on November 2, 1917, declaring that "His Majesty's Government view with favour the establishment in Palestine of a national home for the Jewish people, and will use their best endeavours to facilitate the achievement of this object, it being clearly understood that nothing shall be done which may prejudice the civil and religious rights to existing non-Jewish communities in Palestine, or the rights and political status enjoyed by Jews in any other country."

The British government hoped that a declaration in favour of Zionism would help to rally Jewish opinion, especially in the United States, to the side of the Allies, and that the settlement in Palestine of a Jewish population attached to Britain by ties of sentiment and interest might help to protect the approaches to the Suez Canal and the road to India. The Balfour Declaration fell short of the expectations of the Zionists, who had asked for the reconstitution of Palestine as *the* Jewish national home. Instead, the Balfour Declaration envisaged only the establishment *in* Palestine of *a* national home for the Jewish people. The declaration, nevertheless, aroused enthusiastic hopes among Zionists and seemed the fulfillment of Herzl's hopes. It was endorsed by the principal Allied powers, and through its acceptance by the Conference of San Remo in 1920, it became an instrument of British and international policy. The council of the League of Nations approved on July 24, 1922, a British mandate over Palestine that included the Balfour Declaration in the preamble and various provisions dealing with facilitating Jewish immigration. The mandate had been officially interpreted in a statement of June 3, 1922, in which Winston Churchill, the British colonial secretary, announced that the declaration meant not the "imposition of a Jewish nationality upon the inhabitants of Palestine as a whole, but the further development of the existing Jewish community, with the assistance of Jews of other parts of the world, in order that it may become a centre in which the Jewish people as a whole may take, on grounds of religion and race, an interest and a pride." His Majesty's government, he announced, had not contemplated at any time,

Endorsement of the Balfour Declaration

as appeared to be feared by the Arabs, "the disappearance or the subordination of the Arabic population, language, or culture in Palestine." (H.K.)

The period of the British mandate. Arab nationalists everywhere, fearing that the Zionists meant to make Palestine a Jewish state, declared their unconditional opposition to the Jewish ambitions there. The question was more immediately the concern of the Palestinian Arabs, who saw their interests as directly endangered by Zionist aims. Assured of general Arab support and somewhat misled by the pro-Arab sympathies of British mandatory officials, the Palestine Arabs saw no need to make concessions to Zionism beyond accepting the Jews already established in Palestine as a minority with guaranteed rights.

Concerns of the Palestinian Arabs

In September 1921 the British had promulgated a constitution that provided for the establishment of a Palestinian state in which Arabs and Jews would cooperate. The constitution, however, made concessions to Zionism that the Arabs were unwilling to accept, and it was not applied.

Since no constitutional Palestinian state could be established, the British continued direct rule. They also continued, however, to seek ways to create auxiliaries or partners in government, and in 1923 they offered each of the two major communities an "agency." The Jewish community accepted and formed the Jewish Agency in 1928, but the Arabs refused. As so often in the following years, the Arabs' political action amounted to abstention, while the Jewish community profited by creating a shadow government that enabled them to concentrate their efforts on achievement of their major objectives.

In August 1929 a dispute over the Jewish use of the Wailing Wall—the only remnant of Herod's Temple in Jerusalem, forming the outer wall of the Muslim Haram area—was followed by the first large-scale attacks upon Jews by Arabs. In the course of the troubles, the *muftī* of Jerusalem, Amin-el Husseini, emerged as the leader and champion of the Palestinian Arab cause.

The 1929 troubles led the British to establish a special commission under Sir Walter Shaw, which reported in March 1930 that the conflict was mainly the result of the disappointment of Palestinian Arab hopes for independence, and to the fact that Jewish expansion was creating a "landless and discontented" Arab class in the country. The Shaw Commission urged that restrictions be imposed on the management of the Jewish national home. A subsidiary commission under Lord Passfield was next sent to make concrete proposals, on the basis of which the Passfield White Paper, published in October 1930, recommended that Jews be forbidden to acquire more land while the Arabs were landless and that Jewish immigration be stopped as long as Arabs were unemployed. It renewed the offer of a legislative council that would serve as the basis of constitutional government, a proposal favoured by the Arab leaders but not by the Zionists. When the Arab leaders refused an invitation to discuss the constitutional issue with the Jewish leaders at a roundtable conference, the offer was allowed to lapse.

The Passfield White Paper

The Palestine revolt and the Peel Commission. Zionist pressure and British ambivalence worked against the strict application of the Passfield recommendations on Jewish immigration and purchase of land. When the mandatory government refused to take effective measures to forbid the sale of land to Jews and to stop the illegal Jewish immigration that increased with the persecution of the Jews in Germany after 1933, the Arab leaders announced a policy of noncooperation with the British and a boycott of British goods; at the same time, the existing restrictions on immigration, which were only partly effective, led to Jewish protests and riots.

In April 1936 an Arab High Committee was formed to unite the Palestinian Arabs in opposition to the Jews; its formation was followed by a renewal of Arab attacks on the Jews, soon developing into open war. The revolt of the Arabs continued during the next three years. In November 1936 a new commission, under Lord Peel, arrived to study the situation. The Arab leaders boycotted the commission until just before its departure. The commission's report, published in July 1937, emphasized that cooperation between Arabs and Jews in a Palestinian state was impossible; to the dismay of the Arabs, it recommended the partition of Palestine. The report made it clear that the establishment of a Jewish state would involve radical movements of population to secure the necessary Jewish majority, even in the parts of Palestine where the Jewish population was largest.

In September 1937 nonofficial representatives from the various Arab countries met at Blūdān in Syria and announced the complete rejection of the Peel proposals. In Palestine the publication of the Peel report was followed by renewed Arab terrorism and violence. The British thereupon disbanded the Arab High Committee and deported its leading members. The *muftī* and a few others escaped arrest and fled to Syria, which became the headquarters of a continuing Palestinian Arab insurrection. Before long, however, the insurrection lost its singleness of purpose and degenerated into an Arab civil war as the leaders of the revolt turned their energies against their political rivals.

The British government appointed a new commission under Sir John Woodhead to reconsider the partition plan suggested by the Peel Commission. The Woodhead report, published in November 1938, proposed a reduction of the Jewish share of Palestinian territory to about 400 square miles around Tel Aviv—the only area where the Jews constituted a majority. The Woodhead scheme was completely rejected by the Zionists and also rejected by the Arabs on principle. A conference of Jews and Arabs was next convoked, including Arab representatives from various countries. The conference met in London between February and March 1939, the British conferring with the Arab and Jewish delegations separately.

When no settlement between the two sides could be reached, the British government decided to impose its own terms, which were spelled out in a White Paper published in May 1939. International events were by then moving toward a second world war. This made British appeasement of the Arabs imperative, as German propaganda was gaining wide support in Arab nationalist circles; the Jews, persecuted in Germany, had no choice as to which side to support in the coming war. The White Paper of 1939 was a renewed concession to the Arab position. It stated that there would be no partition and that it was not British policy that the country should become either a Jewish state or an Arab state. It envisaged the establishment within 10 years of an independent "Palestine State." In the intervening period, Jews and Arabs would be invited to take an increasing share in the administration; and Jewish immigration into Palestine would be limited to a total of 75,000 during the next five years, after which no further immigration would be allowed without Arab consent. Land purchases by Jews from Arabs would be prohibited in some areas and restricted in others, in accordance with regulations to be published by the high commissioner.

The British White Paper of 1939

As a proposal for the final settlement of the Palestine question, the White Paper was opposed by both the Zionists and the Arabs. As a means for freezing the situation for the duration of the war, however, it succeeded. Between 1939 and 1945 Palestine was relatively quiet; only as World War II neared its end did the Arab–Jewish conflict resume.

The partition of Palestine and the emergence of Israel. Between 1922 and 1939 the Jewish population in Palestine had risen from 83,790 to 445,457 (30 percent of the total inhabitants). Tel Aviv had become a Jewish city of 150,000.

Until 1939 the plans to establish a Jewish state in Palestine had been compromised by the fact that there were not enough Jews there to secure a majority in a portion of Palestinian territory of the required size. As eastern Europe fell under German domination, however, Nazi persecution drove many more Jews to seek refuge in Palestine by illegal immigration; and when the systematic slaughter of the Jews of Europe began in 1942, the flow of immigrants broke all bounds.

Jewish militance

As the Jewish population swelled with this increase in refugees, the nature of Zionist activity in the country began to change and to tend more toward violence. Following the first Arab attacks on the Jews in 1921, a secret Jewish army called Haganah (Defense) had been formed.

Until 1936 Haganah restricted itself to purely defensive action, but during the years of the Arab revolt it became more aggressive; it also received some legal recognition when the British administration formed a Jewish Settlement Police drawn exclusively from Haganah and placed nominally under British command. In 1937 a more clandestine Jewish militia representing the extreme revisionist party within the World Zionist Organization had been formed—the Irgun Zvai Leumi (National Military Organization). During the early years of the war, Irgun followed the lead of the Jewish Agency and cooperated with Haganah; it soon resumed its extremist course, however, as Jewish refugees freshly arrived from Poland joined its ranks and took over its control. The Irgun leaders were convinced that Britain had betrayed the Zionist cause—an opinion that was shared by another terrorist organization, the Stern Group, or Gang, whose leader, Abraham Stern, was killed in a police raid in 1942. In the last years of the war, the Irgun and the Stern Group began resorting to terror against the British.

Meanwhile, the Jewish Agency, under its veteran leader Chaim Weizmann and younger leaders such as David Ben-Gurion, tried to maintain British goodwill by offers of help to the British war effort. They proposed formation of a Jewish Legion that would undertake the defense of Palestine. Though reluctant to sponsor a Zionist force, the British Army eventually formed a brigade of Jewish volunteers that was active late in the war in Africa and Europe.

Biltmore Resolution

Pressures for a Jewish state. A new Palestine policy was decided upon at a Zionist conference held in May 1942 at the Biltmore Hotel in New York City; it called for unrestricted Jewish immigration into Palestine and for the ultimate establishment of the country as a Jewish commonwealth. By the end of the war, Zionist political activity in the United States had succeeded in winning the U.S. government's support for Zionism, and Britain, unable to resolve the predicament on its own, was pleased to admit American involvement.

At the end of the war in Europe, the Jewish Agency addressed a memorandum to Britain demanding the full and immediate implementation of the Biltmore Resolution. Another memorandum followed in June 1945 demanding that immigration visas be issued for 100,000 European Jewish refugees awaiting admission into Palestine. By the time Japan surrendered in September, Haganah had gone into alliance with the Irgun and the Stern Group to present a united front in Palestine. The Jewish Agency stood ready to assume the provisional government of the Jewish state. In the absence of a unified Arab leadership in Palestine, Arab leaders from the neighbouring countries of Egypt, Syria, Lebanon, Transjordan, Iraq, Saudi Arabia, and Yemen, which in March 1945 had formed the League of Arab States, proclaimed their intent to take up the defense of the Arab cause in Palestine.

In the United States, Pres. Harry S. Truman took up the Zionist cause and urged that the European Jewish refugees be immediately admitted into Palestine. Beset by U.S. pressure, fighting a costly and unpopular war against Zionist guerrillas in Palestine, and seeing no practical solution, Britain participated in yet another commission, the Anglo-American Committee of Inquiry, which published its conclusions in April 1946. In essence, it recommended the immediate admission to Palestine of 100,000 Jewish refugees from Europe, the withdrawal of all restrictions on Jewish purchase of land, and the eventual incorporation of both communities in a binational state under United Nations trusteeship. The British government refused the central and immediate demand, admission of 100,000 refugees, and suddenly found itself involved in a war. Forced to maintain a large and costly military establishment when its electorate demanded demobilization and easement of the tax burden, with no end of the bloodshed in sight and with all its practical political and diplomatic ploys used and options closed, the British referred the question to the United Nations.

UNSCOP

The General Assembly voted on May 15, 1947, to create a Special Committee on Palestine (UNSCOP) to submit "such proposals as it may consider appropriate for the solution of the problem of Palestine." When it arrived in Jerusalem, UNSCOP was boycotted by the Arabs but actively aided by the Zionists. Few issues had been more studied than Palestine, and UNSCOP found nothing new but urgency. The only solution, it suggested, was partition, but it urged that the consequences of partition be mitigated by the maintenance of economic union. On November 29 the UN General Assembly approved, with slight frontier modifications, the UNSCOP recommendations.

The Palestine war. The UN decision was a major Zionist victory. Not only did it affirm the Zionist right—the fundamental point at issue and bitterly opposed by the Arabs—to establish a Jewish state in Palestine, but it also gave the state a territory that, although smaller than that proposed by the Jewish Agency, was far out of proportion to the relative numbers of Jews to Arabs in the country. It comprised more than half the territory of Palestine, including the greater part of the valuable coastal area, leaving only the narrow coastal strip of Gaza, half of Galilee, the Judean and Samarian uplands, and a bit of the Negev to the Arab state. Shocked and angry, the Arab leaders refused to recognize the validity of the UN decision and declared their determination to oppose it by force.

Deir Yāsīn

By January 1948 volunteers were arriving from the Arab countries to help the Palestinian Arabs, but they were soon overwhelmed by the Zionist forces. By May 13 the latter had secured full control of the Jewish share of Palestine and captured important positions in the areas allotted to the Arabs. The Irgun (whose distinction from the Haganah was not clear) stormed and captured the village of Deir Yāsīn and massacred much of the population. This highly publicized act terrorized the Arab villagers, who began a mass exodus from Palestine.

On May 14 the State of Israel was proclaimed and was immediately recognized by the Soviet Union and the United States. On the following day, as the British announced the end of their mandate in Palestine, troops of the modern Transjordanian army and their poorly trained and ill-equipped counterparts from Egypt, Syria, Lebanon, and Iraq entered the country. The Arab forces occupied the areas in the south and east, which were not yet controlled by the Jews, and tried to blockade Jewish Jerusalem.

The United Nations on May 20 appointed Count Folke Bernadotte af Wisborg as mediator to bring about a settlement between Israel and the Arab states. He obtained a brief cease-fire in June and a second one in July. On September 17, however, Bernadotte was assassinated by Jewish terrorists; he was suceeded by his deputy, Ralph J. Bunche of the UN Secretariat. Despite the orders of the United Nations, the truce was not observed faithfully by either side.

Between February and July 1949 the mediator secured separate armistice agreements between Israel and Egypt, Lebanon, Transjordan, and Syria. These agreements left Israel in possession of all the areas it had won by conquest: the whole of Galilee, the whole of the Palestinian coast minus a reduced Gaza Strip (occupied by Egypt), all of the Negev, and a strip of territory connecting the coastal region to the western section of Jerusalem. The remaining parts of Jerusalem (including the Old City), along with what remained of the Arab share of Palestine, were taken over by Transjordan, which then became the Hashemite Kingdom of Jordan. No entity remained that was officially called Palestine. The departure of hundreds of thousands of Palestinian Arabs had meanwhile left Israel with a substantial Jewish majority. (K.S.S./W.R.P./Ed.)

THE STATE OF ISRAEL

The emergence of Israel as a Jewish state on the former territory of Palestine was the central political issue of the Middle East after World War II. The energy, enthusiasm, and skill of the Zionists led to remarkable achievements. Essentially, Israel represented the coming of a modern European state into an underdeveloped area, and in that fact lay the crux of its achievement and its problems.

Deriving proportionally huge external financial and military support from Western governments (its relations with the Soviet Union, despite immediate recognition in 1948, have been strained), the Israelis also benefitted from a highly trained and motivated citizenry to create a unique nation-state. Ironically, they were assisted in this process by the surrounding Arab countries, because the state of siege in which they were forced to live helped to unite them against the common danger and gave them a sense of mission. And, finally, hanging as a tragic backdrop to the state was the memory of the Nazi holocaust, both a justification and a cause of the final push to success of the Zionist movement.

Israel after 1948. In 1948 David Ben-Gurion, the head of the Jewish Agency, became prime minister of the provisional government. Parliamentary elections in January 1949 yielded a Knesset (Parliament) in which the Mapai (Labour Party) of Ben-Gurion had the ascendancy. The new nation did not acquire a written constitution because of disagreement over what such a document should say about the relationship between religion and the state. In keeping with Zionist principle, the Knesset proclaimed Jerusalem the capital of Israel (even though only a section of the new city was actually held by Israel) and passed the Law of Return, which gave every Jew the right to immigrate.

Ben-Gurion formed a succession of coalition governments that kept him in power continuously (except in 1953–55, when his Mapai associate Moshe Sharett was prime minister) until his resignation in 1963. Levi Eshkol, also of the Mapai, followed as the head of another succession of coalition governments. Upon Eshkol's death in 1969, the Israel Labour Party, which had been formed in 1968 in order to unite the Mapai with the more leftist Ahdut Avodah and with the Rafi (a splinter party of the Mapai founded in 1965 by Ben-Gurion in association with Maj. Gen. Moshe Dayan), chose Golda Meir, former foreign minister and secretary general of the party, as prime minister.

Israel Labour Party

Gen. Yitzḥak Rabin, a native-born Israeli, became prime minister in 1974. Rapid inflation, the deep divisions inside the Labour coalition, Israel's relatively poor performance in the 1973 war, and the deep dissatisfaction of Oriental Jews with political domination by eastern Europeans helped bring about the victory of the Likud bloc, a right-wing coalition, and its candidate, Menachem Begin, in May 1977. Begin resigned in 1983, and the Likud bloc remained in power under Yitzḥak Shamir. In 1984 neither Labour nor Likud could form a coalition government by itself, so the two created a national unity government. Under the coalition agreement, Labour leader Shimon Peres served as prime minister and Shamir served as foreign minister until 1986, when the two officials switched portfolios. Because of sharp policy disagreements, the government was often deadlocked.

In the parliamentary elections of November 1, 1988, Likud and Labour won about the same number of seats, and in December the two parties renewed their coalition so that Shamir remained as prime minister and Peres served as finance minister. Major issues in the campaign were the Palestinian uprising that had begun in December 1987 and the differing approaches of the two parties toward the peace process. Likud opposed an international peace conference, encouraged building many more Jewish settlements in the occupied territories, and wished to retain control over all of the West Bank and Gaza Strip. Since Labour adopted the opposite position on these matters, it withdrew from the governing coalition, and for the first time in the history of the country the Knesset voted no confidence in the government on March 15, 1990.

The Arab-Israeli dispute. To large numbers of Jews throughout the world, the State of Israel symbolized the hopes and aspirations of the Jews as a people. To its neighbours, Israel remained an alien presence forcibly established on Arab soil and consequently unacceptable. At a minimum, they demanded a return to frontiers fixed by the 1947 UN partition and repatriation of the Palestinian refugees who had fled the country. Because of Arab hostility toward Israel, the government maintained restrictions on the civil liberties of Arab citizens of Israel.

The precarious frontiers of Israel were periodically exposed to guerrilla raids from the neighbouring countries. Israel reacted against such raids by joining with France and Great Britain to attack Egypt in 1956 but was forced by international pressure to return Sinai and Gaza to Egyptian control in 1957.

The Arab-Israeli war of 1967. While the direct causes of the Six-Day War of June 1967 were Egypt's provocations directed against Israel, the war was, in a more profound sense, another battle of the wars of 1948–49 and 1956. The immediate causes of the war began with Syria's announcement in May 1967 that Israel was massing troops on its border; Egypt felt obliged to come to the defense of Syria. Pres. Gamal Abdel Nasser thereupon called for the immediate withdrawal of the UN Emergency Force from the Israel–Egypt cease-fire lines and closed the Strait of Tiran to Israeli shipping. Moshe Dayan became the Israeli minister of defense.

Six-Day War

On May 30 King Hussein of Jordan signed a mutual defense pact with Nasser, convincing the Israeli cabinet that an Arab attack was imminent. Israel attacked first with a preemptive strike on June 5; the Israeli Air Force caught the Egyptian Air Force on the ground, largely destroying the Arab world's most effective military force. In Sinai in the following days the Israeli Army smashed the Egyptian troops. King Hussein chose to enter the war; the price Jordan paid was the loss of East Jerusalem and the West Bank. Syria in turn suffered an Israeli frontal assault that pushed through the Golan Heights. By the time the UN Security Council managed to effect a cease-fire on June 11, Israel had won a spectacular military victory over all three of its opponents, gaining by its success a new sense of security from the fear of defeat.

Consequences of the 1967 war. Israel immediately annexed the Old City of Jerusalem. Arabs who lived there could choose Israeli or Jordanian citizenship; almost all continued their Jordanian identity. In the West Bank and Gaza Strip, Israel, ruling through a military administration, began constructing a series of settlements for Jewish Israelis. The Arabs who lived in those areas had neither Israeli citizenship nor any real voice in determining their own fate.

Following the shock of the Arab defeat in the 1967 war, the Palestinians concluded that only they could regain what they viewed as their homeland. The Palestine Liberation Organization (PLO) began making raids into Israel and Israeli-occupied territory. Israeli forces retaliated by attacking the "host" countries, Jordan and Lebanon. Since the Palestinians increasingly insisted on their right to act independently, Jordan became an armed camp in which the royal government was but one among several powers. During "Black September" of 1970 King Hussein's army wiped out the Palestinian military. Driven from Jordan, the PLO focused its activities on Lebanon.

Israel, having fought Egypt to a stalemate in a war of attrition along the Suez Canal in 1969–70, now turned to the PLO in Lebanon. Israeli strategy was to induce the Lebanese to crush the PLO in their country as King Hussein had done in Jordan, but this policy did not succeed since the Lebanese armed forces were weak and the Lebanese polity was divided over what its reaction to Israeli pressure should be.

The war of 1973. When Nasser died in 1970, he was succeeded by Anwar el-Sādāt. As Sādāt became frustrated by his inability to change the diplomatic situation, he decided on a limited war with Israel. By the end of June 1973 Sādāt and Pres. Ḥafiz al-Assad of Syria had made plans for a coordinated surprise attack. Fighting started on the Syrian and Egyptian fronts on October 6—for the Jews, Yom Kippur, the holiest day of the year, and for the Muslims, the 10th day of Ramaḍān, the anniversary of a crucial battle fought by the Prophet Muḥammad. Despite initial setbacks, once Israel was fully mobilized the Israel Defense Forces assumed the offensive. As Israeli forces crossed the Suez Canal and moved to surround the Egyptian Third Army, the United States and the Soviet Union entered into a grave crisis that threatened to lead to a nuclear world war. A compromise was worked out between the two superpowers, and a UN-monitored cease-fire took effect on October 24. On the Syrian front, a similar cease-fire arrangement finally became fully effective only in May 1974. During the October fighting, Israel gained additional territory along the Suez Canal (Egyptian

Surprise attack

armed forces were present on both sides of the canal as well), and Israel took more land from Syria beyond the cease-fire lines of 1967, coming closer to Damascus.

The United States brokered disengagement agreements between Israel and Egypt and between Israel and Syria. As Israel implemented these arrangements, Israeli forces withdrew from designated territories, Egyptian or Syrian forces moved forward in limited numbers, and UN forces were stationed in between.

Moves toward peace. Most Israelis welcomed U.S. Secretary of State Henry Kissinger's "step-by-step" approach to peace, since Israel not only gained the prospect of a final end to the Arab-Israeli dispute but also secured increased American diplomatic and economic assistance. In 1977 the new U.S. administration of Pres. Jimmy Carter issued a joint note with the Soviet Union outlining a basis for peace in the Middle East. Israel and Egypt, fearing this initiative, moved rapidly toward ending 30 years of war with the dramatic trip of President Sādāt to Israel in November 1977. Sādāt told the Knesset that he believed peace could be established by breaking the psychological barriers of suspicion and fear between the two nations. Menachem Begin, the new Israeli prime minister, endorsed the concept of a peace of mutual reconciliation. In September 1978, under the direct supervision of President Carter, a framework for peace was announced at Camp David, Maryland.

The actual peace treaty between Egypt and Israel—signed in Washington, D.C., on March 26, 1979—gave Israel a complete peace, full diplomatic recognition by Egypt, and limits on Egyptian armaments in the Sinai and gave Egypt the reoccupation, in stages, of its lands lost in the 1967 war and the vague assurances of the creation of a self-governing authority for the Palestinians living in the West Bank and Gaza Strip.

The PLO and most Arab countries opposed the treaty, decrying the continued Israeli presence in the occupied territories and the failure to include the PLO in the negotiations. Israel continued to approve new Jewish settlements in the occupied territories despite the opposition of the United States and the danger such action posed to the peace process. Talks on Palestinian autonomy, provided for by the peace treaty, proved fruitless.

Despite the assassination of President Sādāt on October 6, 1981, Egypt, under its new president, Hosnī Mubārak, continued with the peace process. Israel fulfilled the 1979 peace by returning the last segment of the Sinai Peninsula to Egyptian control in April 1982. During that period, however, Israel's desire for peace with the other Arab states and autonomy for the Palestinians was called into doubt by several Israeli actions: the removal of elected Arab mayors from office in West Bank towns, the declaration of undivided Jerusalem as the eternal capital of Israel, the bombing of a nuclear reactor in Iraq so as to preserve Israel's nuclear warfare monopoly in the Middle East, and the annexation of the Golan Heights.

Declaration of Jerusalem as capital

Having neutralized Egypt, the largest and most powerful of the Arab states, Begin and Defense Minister Ariel Sharon, a hero of the 1973 war, planned an invasion of Lebanon to secure the elimination of the PLO and the selection of a new president of Lebanon who would sign a peace treaty with Israel along the lines of the Egyptian-Israeli treaty of 1979. On June 6, 1982, Israel invaded Lebanon and subsequently defeated the PLO, the Syrian armed forces, and assorted leftist Lebanese groups. By June 13, Israeli forces and their Phalangist Lebanese allies had encircled West Beirut, and the trapped PLO and Syrians were forced to agree to leave the city. The assassination of the pro-Israeli Lebanese president-elect, Bashir Gemayel, provoked Israeli troops to move into West Beirut, where they allowed Lebanese Christian Phalangists to massacre Palestinian civilians in two refugee camps.

Israel had succeeded in forcing the PLO out of most of Lebanon, but the Israeli-Lebanese troop-withdrawal agreement, which was tantamount to a peace treaty, failed since Syria and many Lebanese violently opposed it. Even with U.S. support, the new Lebanese president, Amin Gemayel, failed to unite his country behind a pro-Israeli foreign policy. Instead, the United States was forced to withdraw by February 1984, and Israel finished its slow retreat by June 1985. Lebanon slid even further into chaos while Israel retained a sphere of influence only in the extreme southern part of Lebanon.

Repeated attempts by the PLO and Jordan to form a joint negotiating team to meet with Israel were blocked by the Israelis, who refused to deal with the PLO. Instead of negotiations, raids and counterraids continued between the Israelis and the Palestinians. Perhaps the most spectacular of these was the Israeli attack on PLO headquarters in Tunis on October 1, 1985.

The intifada. A large-scale uprising by the Palestinians in the occupied territories began on December 8, 1987. This "shaking" (*intifada* in Arabic) came after 20 years of Israeli occupation; the Palestinians felt that there was no other way to accomplish their goals of self-government and national independence except by revolt. The *intifada* took a number of different forms: boycotts of Israeli goods, attacks against Israeli settlers, demonstrations to show public support for Palestinian nationhood, and rock throwing by youths against Israeli soldiers. In the first year alone, more than 300 Palestinians were killed, thousands were wounded, and 20,000 more were imprisoned. Israel's reaction was one of armed suppression of the revolt, including the use of rigorous tactics by the Israeli military, whose severity was condemned not only by the Palestinians but also by many Israelis.

Reaction to *intifada*

The Labour and Likud parties agreed that the *intifada* must be suppressed before changes took place in the status of the occupied territories. King Hussein of Jordan did not wait to see the outcome of the *intifada;* instead, he announced on July 31, 1988, that Jordan was severing its official claims to the West Bank and East Jerusalem. On November 15 the PLO National Council voted to declare the establishment of "a Palestinian state with Jerusalem as its capital" despite actual Israeli control over the territory claimed by the new state.

Israel was faced with a number of dilemmas: whether it should annex the territories and thereby dilute its Jewish majority, return the conquered lands to Jordanian control, hoping for an exchange of land for peace, or deal with the PLO and thereby accept the idea of Palestinian nationhood. (W.L.O.)

For later developments in the history of Israel, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 96/11, 971, and 978, and the *Index.*

BIBLIOGRAPHY

Physical and human geography: Landforms are described in EFRAIM ORNI and ELISHA EFRAT, *Geography of Israel,* 4th rev. ed. (1980); and *Geography* (1973), compiled from material originally published in the *Encyclopaedia Judaica* and published by Keter Books. LEO PICARD, *Structure and Evolution of Palestine* (1943); and "History of Mineral Research in Israel," *Israel Economic Forum,* vol. 6 (1954), are authoritative and comprehensive geologic studies. MICHAEL ZOHARY, *Plant Life of Palestine: Israel and Jordan* (1962), is a fundamental study. FRIEDRICH S. BODENHEIMER, *Animal Life in Palestine* (1935), and *Prodromus Faunae Palestinae* (1937), are classic works on Israel's fauna. Economic studies include DAVID HOROWITZ, *The Economics of Israel* (1967); NADAV HALEVI and RUTH KLINOV-MALUL, *The Economic Development of Israel* (1968); and MEIR HETH, *Banking Institutions in Israel* (1966; originally published in Hebrew, 1963), informative and reliable. Administrative and political aspects are explored by DON PERETZ, *The Government and Politics of Israel,* 2nd ed., updated (1983); EDWARD LUTTWAK and DAN HOROWITZ, *The Israeli Army* (1975); ASHER ZIDON, *Knesset: The Parliament of Israel* (1968; originally published in Hebrew, 1964); HENRY E. BAKER, *The Legal System of Israel,* rev. ed. (1968); and WILLIAM FRANKEL, *Israel Observed* (1980). DOV FRIEDLANDER and CALVIN GOLDSCHEIDER, *The Population of Israel* (1979), is a highly useful work on population policy. RAANAN WEITZ and AVSHALOM ROKACH, *Agriculture and Rural Development in Israel: Projection and Planning,* trans. from Hebrew (1963), and *Agricultural Development: Planning and Implementation* (1968), examine economic aspects. JOSEPH S. BENTWICH, *Education in Israel* (1965), is informative and comprehensive. WALTER PREUSS, *The Labour Movement in Israel: Past and Present,* 3rd ed. (1965; originally published in German, 1932–36), is a substantial historical volume.

History: Valuable general surveys of Israel's prehistory may be found in WILLIAM F. ALBRIGHT, *The Archaeology of Palestine* (1949, reprinted 1971); KATHLEEN M. KENYON, *Archaeology in the Holy Land,* 4th ed. (1979); and ARCHAEOLOGICAL INSTITUTE OF AMERICA, *Archaeological Discoveries in the Holy Land* (1967). Works describing the Zionist movement and the establishment and subsequent history of Israel include NAHUM SOKOLOW, *History of Zionism 1600–1918,* 2 vol. (1919, reprinted 2 vol. in 1, 1969); LEONARD STEIN, *Zionism* (1925); ELMER BERGER, *Judaism or Jewish Nationalism* (1957); NORMAN BENTWICH, *Palestine* (1934, reissued 1946); ALBERT M. HYAMSON, *Palestine Under the Mandate, 1920–1948* (1950, reprinted 1976); SAMUEL HALPERIN, *The Political World of American Zionism* (1961, reissued 1985); and BARNET LITVINOFF, *To the House of Their Fathers: A History of Zionism* (1965).

The material available on the Palestine question, Israel, and Arab-Israeli relations is vast, hardly any of it objective. Some of the few works that are objective include FRED J. KHOURI, *The Arab-Israeli Dilemma,* 3rd ed. (1985); CHARLES D. SMITH, *Palestine and the Arab-Israeli Conflict* (1988); BERNARD REICH, *Israel: Land of Tradition and Conflict* (1985); DON PERETZ, *Intifada: The Palestinian Uprising* (1990); and Howard M. Sachar, *A History of Israel,* 2 vol. (1979–87).

(E.E./H.K./W.L.O.)

Istanbul

Istanbul (formerly Constantinople, ancient Byzantium, Turkish İstanbul) is the largest city and seaport of Turkey. It was formerly the capital of the Byzantine Empire, of the Ottoman Empire, and—until 1923—of the Turkish Republic.

The old, walled city of Istanbul stands on a triangular peninsula between Europe and Asia. Sometimes as a bridge, sometimes as a barrier, Istanbul for more than 2,500 years has stood between conflicting surges of religion, culture, and imperial power. For most of those years it was one of the most coveted cities in the world.

The name Byzantium may derive from that of Byzas, who, according to legend, was leader of the Greeks from the city of Megara who captured the peninsula from pastoral Thracian tribes and built the city about 657 BC. In 196 BC, having razed the town for opposing him in a civil war, the Roman emperor Septimius Severus rebuilt it, naming it Augusta Antonina in honour of his son. In AD 330, when Constantine the Great dedicated the city as his capital, he called it New Rome. The coinage, nevertheless, continued to be stamped Byzantium until he ordered the substitution of Constantinopolis. In the 13th century Arabs used the appellation Istinpolin, a "name" they heard Byzantines use—*eis tēn polin*—which, in reality, was a Greek phrase that meant "in the city." Through a series of speech permutations over a span of centuries, this name became Istanbul. Until the Turkish Post Office officially changed the name in 1926, however, the city continued to bear the millenary name of Constantinople.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. The old city contains about nine square miles (23 square kilometres), but the present municipal boundaries stretch for more than 98 square miles, including areas on both sides of the Bosporus and the Sea of Marmara. The original peninsular city has seven hills requisite for Constantine's "new Rome." Six are crests of a long ridge above the Golden Horn; the other is a solitary eminence in the southwest corner.

By long tradition, the waters washing the peninsula are called "the three seas": they are the Golden Horn, the Bosporus, and the Sea of Marmara. The Golden Horn is a deep, drowned valley about 4$^1/_2$ miles (seven kilometres) long. Early inhabitants saw it as being shaped like a deer horn, but modern Turks call it the Haliç (Canal). The Bosporus (Boğazici) is the channel connecting the Black Sea (Kara Deniz) to the Mediterranean (Ak Deniz) by way of the Sea of Marmara (Marmara Deniz) and the straits of the Dardanelles. The narrow Golden Horn separates the old city of Stamboul to the south from the "new" city of Beyoğlu to the north; the broader Bosporus divides European Istanbul from the city's districts on the Asian shore—Üsküdar (ancient Chrysopolis) and Kadiköy (ancient Chalcedon).

Like the forces of history, the forces of nature impinge upon Istanbul. The great rivers pouring off the plains of Russia and middle Europe—the Danube, Don, Dnestr, and Dnepr—make the Black Sea colder and less briny than the Mediterranean. The Black Sea waters thrust southward through the Bosporus, but beneath them the salty warm waters of the Mediterranean push northward as a powerful undercurrent running through the same channel.

Climate. The prevailing wind, the northeast wind, or *poyraz,* comes from the Black Sea, giving way at times during the winter to an icy blast from the Balkans—the northwest wind, known as the *karayel,* or Black Veil, capable of freezing solid the Golden Horn and even the Bosporus. When the *lodos,* or southwest wind, blows, it can raise storms on the Sea of Marmara.

Istanbul's disasters

Fire, earthquake, riot, and invasion have ravaged the city many times. More than 60 conflagrations have been important enough to be recorded in history, and there remain scorched stretches of the old city that have never been rebuilt. Fifty major earthquakes and innumerable less serious temblors have shaken the city since the time of Constantine the Great. The fall of each empire has also been followed by devastation and a period of decay for the capital.

The city plan. Many of the burned-out neighbourhoods have slowly been rebuilt, while a continuing program of street improvement has pushed wide avenues through some of the meanest quarters of the old city. There remain, however, numbers of unpaved alleys overhung with decrepit wooden houses.

The Golden Horn, Istanbul. The two mosques in the background are (centre left) Hagia Sophia and (centre right) the Mosque of Süleyman.

Ara Gule

The city walls

Stamboul is still a walled city. The land walls, which isolate the peninsula from the mainland, were breached only once, by the cannon of Mehmed II (the Conqueror) in 1453, at the spot since called Cannon Gate (Top Kapısı). The walls are four and a half miles long and consist of a double line of ramparts—the inner built in 413, the outer in 447—protected by a moat. The higher inner wall is about 30 feet (9 metres) high and 16 feet thick and is studded with 60-foot towers about 180 feet apart. Of 92 turrets originally on the outer wall, 56 are still standing.

The sea walls were built in 439. Only short sections of their 30-foot-high masonry still remain along the Golden Horn. Intact, these walls had 110 towers and 14 gates. The walls along the Sea of Marmara, which stretch about five miles from Seraglio Point, curving around the bottom of the peninsula to join the land walls, had 188 towers; they were, however, only about 20 feet high, because the Marmara currents provided good protection against enemy landings. Most of these walls still stand.

Within the city walls are the seven hills, their summits flattened through the ages, their slopes still steep and toilsome. Geographers number them from the seaward tip of the peninsula, proceeding inland along the Golden Horn, the last hill standing alone where the land walls reach the Sea of Marmara.

The Golden Horn

The Galata and Atatürk bridges cross the Golden Horn to Beyoğlu. Each day before dawn their centre spans are swung open to allow passage to seagoing ships. The shores of the Horn, served by water buses, are a jumble of docks, warehouses, factories, and occasional historical ruins. Ferries to the Asian side of Istanbul leave from under the Galata Bridge. Istanbul's Bosporus Bridge (completed in 1973), with a main span of 3,524 feet, is one of the world's longest suspension bridges.

Beyoğlu, considered to be "modern Istanbul," remains, as it has been since the 10th century, the foreign quarter. Warfare and fires have left standing only a few structures that were built earlier than the 19th century.

The approach from the Golden Horn is steep, and a funicular railway runs between the Galata waterfront and the Pera Plateau. On the heights are the big hotels and restaurants, the travel bureaus, theatres, the opera house, the consulates, and many Turkish government offices.

From the 10th century onward, Galata was an enclave for foreign traders—principally the Genoese—who enjoyed extraterritorial privileges behind their walls. After the Ottomans took the city in 1453, all foreigners who were not citizens of the empire were restricted to this quarter. Around palatial embassies were compounds that included schools, churches, and hospitals for the various nationalities. Eventually Galata became too crowded, so that the tide of building moved higher up the slope to the open country of Pera. For centuries, foreigners who wished to visit Stamboul, where the Court was installed, could do so only if accompanied by one of the sultan's Janissaries.

Remains of Constantine's city

Architecture. *Byzantine monuments.* Nothing remains of the Byzantium that Constantine chose as the site of New Rome, and almost nothing is left of the mighty city he built there. Constantine's column, the Burnt Column (Çemberlitaş), a shaft of porphyry drums bound by metal laurel leaves, still stands near the Mosque of Nûruosman, but there is no proof that any building in the city dates from his period. He completed the Hippodrome that Septimus Severus had restored, but it was enlarged and rebuilt by his successors until the 5th century. Only its curved end remains, with three columns along the central spina—an obelisk removed from Egypt by the Roman emperor Theodosius I, a masonry obelisk of Constantine VII Porphyrogenitus (AD 905–959), and a Delphic column formed by three entwined serpents (now headless) cast after the Battle of Plataea, when the Greeks defeated the Persians in 479 BC.

Of the myriad columns that decorated Constantinople, there remain standing only the base of the column of the emperor Arcadius (reigned 383–408) in the Cerrahpaşa quarter; a column of the emperor Marcian (reigned 450–457) that the Turks call Kıztaşı; the Column of the Virgin, in the Fatih quarter; and, in the grounds of the Topkapı Palace, a perfectly preserved Corinthian column thought to be from the reign of another emperor, Claudius II Gothicus (268–270).

Spanning the valley between the third and fourth hills is the two-story limestone aqueduct built in 366 by the emperor Valens. Some of the enormous open-water cisterns of the Byzantine epoch now serve as market gardens. The closed cisterns, of which there are more than 80 remaining, include one of the most beautiful and mysterious structures of Istanbul, the Basilican Cistern, or Yerebatan Sarayı (underground palace), near the Hagia Sophia; its 336 columns rise from the still, black waters to a vaulted roof.

The Golden Gate is a triumphal arch from about 390 and was built into the defenses of Theodosius II, near the junction of the land and sea walls. The marble-clad bases of its two large towers still stand, and three arches decorated with columns stretch between them.

Monuments on the Seven Hills and Their Slopes

First Hill
Hagia Sophia
Church of St. Irene
Mosque of Kiiçük Ayasoya
(Church of SS. Sergius and Bacchus)
Mosque of Sokullu Mehmed Paşa
Mosque of Ahmed I (Blue Mosque)
Fountain of Ahmed III
The Museums
Çinili Kiosk (Pavilion of Tiles)
Basilican Cistern
(Yerebatan Sarayı)
Hippodrome
Topkapı Palace (Seraglio)
Marmara Sea Walls

Second Hill
Mosque of Nûruosman
The Burnt Column (Çemberlitaş)
The Great Bazaar (Kaplı Çarşı)

Third Hill
Mosque of Vefa Kilise
(Church of St. Theodore Tiro)
Mosque of Bayezid II
Mosque of Laleli
Mosque of Şehzade
Mosque of Süleyman, and tombs
Mosque of Bodrum
(Monastery of Myrelaion)
Mosque of Kalendarhane
(Monastery of Akataleptos)
Aqueduct of Valens

Fourth Hill
Mosque of the Fatih (Conqueror)
Mosque of Mollazeyrek
(Church of Christ Pantokrator)
Mosque of Eski İmaret
(Church of Christ Pantepoptes)

Fifth Hill
Mosque of Ahmed Paşa (Church of St. John the Baptist in Trullo)
Mosque of Gül
(Church of St. Theodosia)
Mosque of Fethiye (Church of the Pammakaristos Virgin)
Church of St. Mary of the Mongols
Greek Patriarchal Church of St. George

Sixth Hill
Mosque of Kari
(Church of St. Saviour in Chora)
Mosque of Mihrimah
Adrianople Gate (Edirne Gate)
Tekfur Sarayı (Palace of Constantine)

Seventh Hill
Mosque of Hekimoğlu Ali Paşa
Mosque of Ramazan Efendi
Seven Towers Castle (Yedikule)
Mosque of Koca Mustafa Paşa
(Church of St. Andrew in Krisei)
Mosque of İmrahor (Church of St. John of the Stoudion)

The only well-preserved example of Byzantine palace architecture is the shell of a three-story rectangular building of limestone and brick, laid in patterns and stripes. Dating from about 1300, it is called the Palace of Constantine (Tekfur Sarayı) and is attached to the land walls not far from the Golden Horn.

The Hagia Sophia

The largest legacy from the capital of the vanished empire is constituted by 25 Byzantine churches. Many of these are still in use—as mosques. The largest of the churches is considered one of the great buildings of the world. This is the Hagia Sophia, whose name means "Divine Wisdom." Its contemporary and neighbour, St. Irene, was dedicated to "Divine Peace." Many art historians deem the dome (105 feet in diameter) of Hagia Sophia to be the most beautiful in the world. The church, which shared its clergy with St. Irene, is said to have been built by Constantine in 325 on the foundations of a pagan temple. It was enlarged by the emperor Constans and rebuilt after the fire of 415 by the emperor Theodosius II. The church was burned again in the Nika Insurrection of 532 and reconstructed by Justinian. The structure now standing is essentially the 6th-century edifice, although an earthquake tumbled the dome in 559, after which it was rebuilt to a smaller scale and the whole church reinforced from the outside. It was restored again in the mid-14th century. In 1453 it became a mosque with minarets, and a great chandelier was added. In 1935 it was made into a museum. The walls are still hung with Muslim calligraphic disks.

The Church of SS. Sergius and Bacchus was erected by Justinian between 527 and 536 as a thank offering. The two soldier-saints allegedly appeared to the emperor Anastasius I to intercede for Justinian, who had been condemned to death for conspiracy. The church is built as a domed octagon within a rectangle, with a columned and galleried Byzantine interior. It is also called the Mosque of Küçük Ayasofya (Little Sophia) and can be considered an architectural parent of Justinian's reconstruction of the Hagia Sophia. The Church of St. Saviour in Chora (now called the Mosque of Kari) is near the Adrianople Gate. It was restored in the 11th century and remodelled in the 14th; the building is now a museum renowned for its 14th-century mosaics, marbles, and frescoes. Over the central portal is a head of Christ with the inscription, "The land of the living." When it was made a mosque, it acquired the narthex (an enclosed passage between the main entrance and the nave), portico, and minarets.

A massive tower that dominates the Galata district was built by the Genoese traders in 1349 as a watchtower and a fortification for their walled enclave.

The Sinan architectural tradition

Turkish monuments. When the Turks took possession of Constantinople, they covered the spines of the seven hills with domes and minarets, changing the character of the city. Like the Greeks, the Romans, and the Byzantines, the new rulers loved the city and spent much of their treasure and energy on its embellishment. The Ottoman dynasty, which lasted from 1300 to 1922, continued to build new important structures almost until the end of their line. The most imposing of their mosques were constructed from the mid-15th to the mid-16th century, and the greatest of the architects all bore the name of Sinan. They were Atik Sinan (the Elder), Sinan of Balıkesir, and Mimar Koca Sinan (Great Architect Sinan). Although the building was deeply influenced by the Persian-born traditions of the Seljuqs (once masters of the Ottomans), the style was blended with prevailing Hellenic and Byzantine traditions of the city. Mimar Koca Sinan's masterpiece—and his burial place—is the Mosque of Süleyman (1550–57), inspired by, but not copied from, the Hagia Sophia. It ranks as another of the world's great buildings. Probably the most popularly known of all the mosques in Istanbul is the Blue Mosque, that of the Mosque of Ahmed I (Ottoman sultan from 1603 to 1617), which has six minarets instead of the customary four.

The mosques of the 18th century and later show the deleterious effects of importing European architects and craftsmen, who produced baroque Islāmic architecture (such as the Mosque of the Fatih, rebuilt between 1767 and 1771) and even Neoclassical styles, as in the Dolmabahçe Mosque of 1853, now the Naval Museum.

The big mosques were built with ancillary structures, such as a Qur'ānic school (*medrese*), baths (*hamam*) for purification, a hostel and kitchen for the poor (*imaret*), or tombs of royalty and distinguished persons.

There are more than 400 fountains in Istanbul. Some simply flow from wall niches, but others, erected as public philanthropies, are pavilions. The most magnificent of these was built by the sultan Ahmed III in 1728, behind the apse of Hagia Sophia. It is square, with marble walls and bronze gratings, a mixture of the Turkish with the Western rococo style.

Topkapı Palace

To the north of it, toward the Golden Horn and occupying the whole tip of the promontory, is the sultan's Seraglio (Topkapı Palace), enclosed in a fortified wall. It was begun in 1462 by Mehmed II and served as the residence of the sultans until the beginning of the 19th century. It was to this palace that foreign ambassadors were accredited, and they were admitted through the Imperial Gate, or Bab-i-Hümayun, mistranslated by Westerners as "Sublime Porte." The Seraglio consists mostly of small buildings grouped around three courts. The most significant buildings are the Çinili Kiosk (Pavilion of Tiles) built in 1472, the Audience Chamber (Arzodası), the Hirkaiserif, a sanctuary containing relics of the Prophet Muḥammad, and the elegant Baghdad Kiosk commemorating the capture of Baghdad in 1638. The Seraglio houses the sultan's treasure and has important collections of manuscripts, china, armour, and textiles. After the abandonment of the Old Seraglio, the sultans built for themselves palaces along the Bosporus, such as the Beylerbey Palace (1865), the lavish Dolmabahçe Palace (1853), the Çeragan Palace built in 1874 and burned in 1910, and the Yıldız Palace, which was the residence of Abdülhamid II, Ottoman sultan from 1876 to 1909.

The Great Bazaar (Kaplı Çarşi), founded early in the Turkish regime, but often subject to fire and earthquake, had 4,000 shops around two central distributing houses. The district is laid out on a grid plan. It still boils with life and the pursuit of piasters.

THE PEOPLE

Istanbul, like other major cities, attracts an increasing number of migrants from the countryside. These migrants have contributed to the use of shantytowns called *gecekondu* (literally "set down by night") that have no sanitation facilities. Some migrants work as porters, bearing upon their backs burdens of immense size and weight. The Turkish Muslim majority continues to grow, and the Christian and Jewish minorities continue to shrink both in percentage of the whole and in numbers.

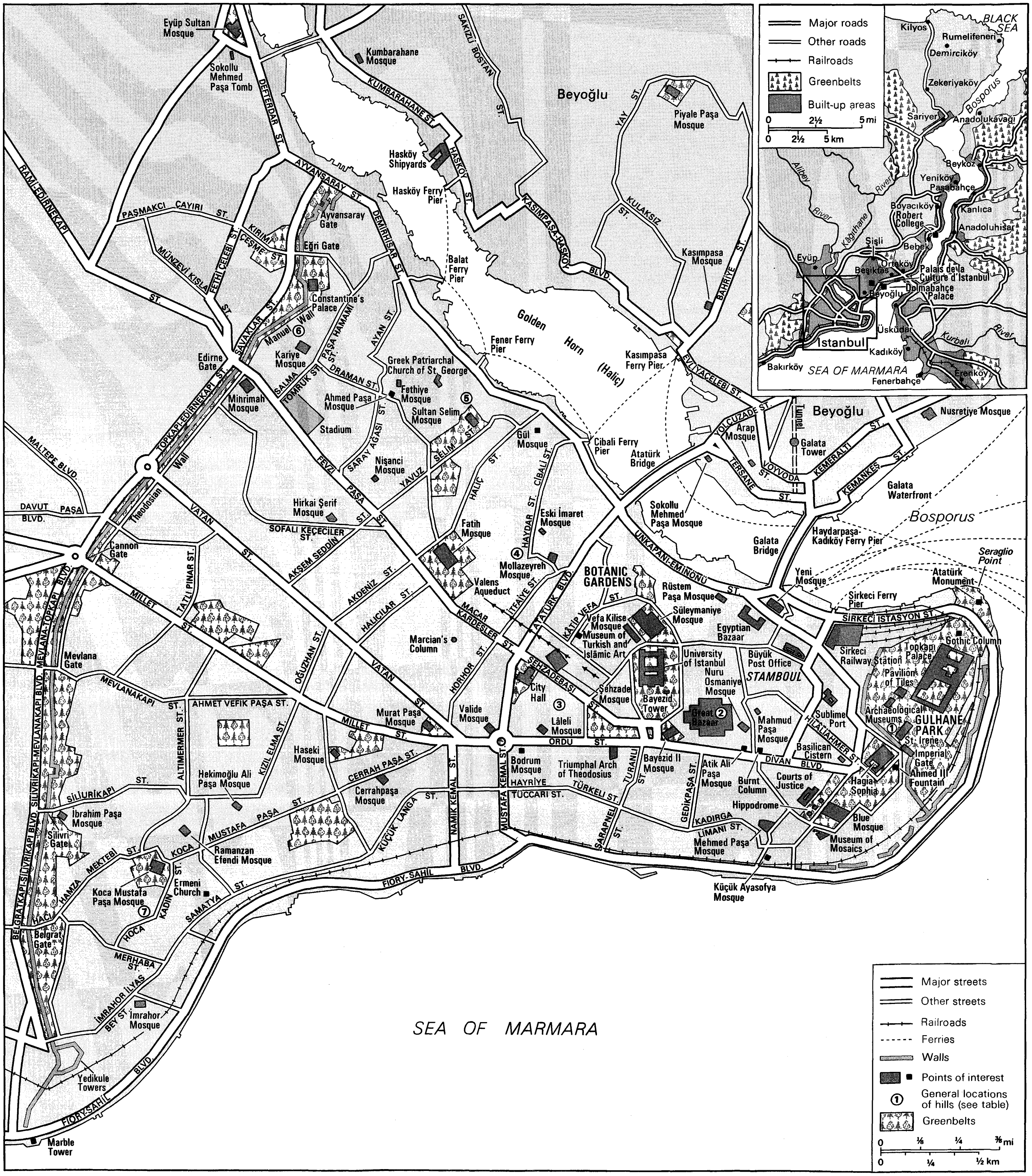
THE ECONOMY

Industry. Istanbul is Turkey's largest port and is the hub of its industry. Textiles, flour milling, tobacco processing, cement, and glass are the city's principal manufactures.

Transportation. Tourism is a growing source of income for Istanbul. There is rail service along the walls of the old city, and Haydarpaşa station, on the Asian side of the city, is the starting point of the Baghdad Railway. Maritime services include many forms of transport, from harbour dinghies and small ferries to international liners. Yeşilköy Airport is about 17 miles to the west of the city. Buses provide internal urban transportation, and the ferries range as far as the Kızıl Adalar (Princes Islands), several hours sailing to the south.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The mayor, appointed by the president of the republic, serves as prefect of Istanbul city and gover-



Istanbul and (inset) its metropolitan area.

nor of Istanbul *il* (province). The municipality, which was organized by Constantine as 14 districts in imitation of Rome, is now divided into 12 circumscriptions (*kazas*), each governed by a *kaymakam*. These are, in the old city, Eminönü and Fatih; on the European side above the Golden Horn, Eyüp, Gakirkoy, Beyoğlu, Şişli, Beşiktas, and Sarıyer; and across the Bosporus on the Asiatic side, Beykoz, Üsküdar, Kadıköy, and Adalar.

Public utilities. While Istanbul has a chlorinated and filtered water supply and sewage disposal system, these facilities are not sufficient to meet the increased need created by the influx of rural migrants to the city. Water supply is a problem particularly in the summer when rivers run dry; at this season tap water is likely to flow only sporadically, except in luxury hotels. Electric power supplies have been increased to help promote industrial expansion.

Health. Most of the health services of Istanbul *il* (province) are concentrated in the municipality. There are more than 70 hospitals, about half of which are public.

Istanbul's universities

Education. The first University of Istanbul was founded in 425 by Theodosius II and was succeeded by Istanbul University (İstanbul Ünıversitesi), founded in 1453. The university now includes faculties of letters, science, law, medicine, and forestry located in the former Seraskerat (war ministry) between the Great Bazaar and the Mosque of Süleyman. There is also a technical university on the Galata side of the Horn as well as an Academy of Fine Arts and schools of technology, commerce, and economics. Foreign educational institutions include the American Robert College for boys (founded in 1863) and the American College for girls (founded in 1871), both on the Bosporus.

CULTURAL LIFE

The Palais de la Culture d'Istanbul is an important centre for the arts. Facilities include a concert hall, art gallery, and two theatres. It is the home of the Istanbul Municipal Symphony Orchestra and the Istanbul City Opera. The municipal theatre operates several playhouses, and there are many theatre companies.

A large number of learned societies and research institutes are headquartered in the city, including the Geographical Institute (Coğrafya Enstitüsü), German and French archaeological institutes, and the Turkish Folklore Society (Türk Halk Bilgisi Derneği). There is a nuclear research centre at Küçük Çekmece.

There are many public and private libraries. The small, specialized Köprülü Library (1677) has books from early Ottoman presses and handwritten works more than 1,000 years old. Many of the city's mosques, palaces, and monuments, as mentioned earlier, contain museums; other museums include the Archaeological Museums of Istanbul (Istanbul Arkeoloji Müzeleri), the Museum of Turkish and Islāmic Art (Türk ve Islam Eserleri Müzesi), and the Museum of the Janissaries (Türkiye Askeri Müzesi).

The Hippodrome is now a public garden; there are also numerous other public parks. A unique feature of the city is its market gardens; these kitchen gardens are associated with the open cisterns that formed early Constantinople's water-supply system. The cisterns have been partially built over and are called Çukur Bostan (Hollow Gardens).

Football (soccer) is a popular sport, and Istanbul has three stadiums—Mithatpaşa, Fenerbahçe, and the indoor Spor ve Sergi Sarayı. There are facilities for tennis, fencing, mountain climbing, riding, golf, and water sports. Florya and Ataköy are popular beaches on the Sea of Marmara.

History

THE EARLY PERIOD

The founding of Byzantium

Byzantium. Byzantium was one of the many colonies founded from the end of the 8th century onward along the coasts of the Bosporus and the Black Sea by Greek settlers from the cities of Miletus and Megara.

The Persian king Darius I took the settlement in 512 BC; it slipped from Persian grasp during the Ionian revolt of 496, only to be retaken by the Persians. In 478 an Athenian fleet captured the city, which then became a rich and important member of the Delian League. As Athenian power waned during the Peloponnesian War, Byzantines acknowledged Spartan overlordship. Although Alcibiades besieged and retook the city, Sparta reasserted its domination after defeating Athens in 405 BC.

In 343 BC Byzantium joined the Second Athenian League, throwing off the siege of Philip II of Macedon three years later. The lifting of the siege was attributed to the divine intervention of the goddess Hecate and was commemorated by the striking of coins bearing her star and crescent. Byzantium accepted Macedonian rule under Alexander the Great, regaining independence only with the eclipse of Macedonian might. In the 3rd century BC, the city's treasury was drained to buy off marauding Gauls. A free city under Rome, it gradually fell under imperial control and briefly lost its freedom under the emperor Vespasian. When, in AD 196, it sided with the usurper Pescennius Niger, the Roman emperor Septimus Severus massacred the populace, razed the walls, and annexed the remains to the city of Perinthus (or Heraclea, modern Marmaraereğlisi), in Turkey.

Subsequently, Septimus Severus rebuilt the city on the same spot but on a grander scale. Although sacked again by Gallienus in 268, the city was strong enough two years later to resist a Gothic invasion. In the subsequent civil wars and rebellions that broke out sporadically in the Roman Empire, Byzantium remained untouched until the arrival of the emperor Constantine I—the first Roman ruler to adopt Christianity. Overcoming the army of the rival emperor, Licinius, at nearby Chrysopolis, on September 18, 324, Constantine became head of the whole Roman Empire, east and west. He decided to make Byzantium his capital.

The New Rome

Constantinople. Within three weeks of his victory, the foundation rites of New Rome were performed, and the much-enlarged city was officially inaugurated on May 11, 330.

It was an act of vast historical portent. Constantinople was to become one of the great world capitals, a font of imperial and religious power, a city of vast wealth and beauty, and the chief city of the Western world. Until the rise of the Italian maritime states, it was the first city in commerce, as well as the chief city of what was, until the mid-11th century, the strongest and most prestigious power in Europe.

Constantine's choice of capital had profound effects upon the ancient Greek and Roman worlds. It displaced the power centre of the Roman Empire, moving it eastward, and achieved the first lasting unification of Greece.

Culturally, Constantinople fostered a fusion of Oriental and Occidental custom, art, and architecture. The religion was Christian, the organization Roman, and the language and outlook Greek. The concept of the divine right of kings, rulers who were defenders of the faith—as opposed to the king as divine himself—was evolved there. The gold *solidus* of Constantine retained its value and served as a monetary standard for more than a thousand years. As the centuries passed—the Christian Empire lasted 1130 years—Constantinople, seat of empire, was to become as important as the empire itself; in the end, although the territories had virtually shrunk away, the capital endured.

Constantine's new city walls tripled the size of Byzantium, which now contained imperial buildings, such as the completed Hippodrome begun by Severus, a huge palace, legislative halls, several imposing churches, and streets decorated with multitudes of statues taken from rival cities. In addition to other attractions of the capital, free bread and citizenship were bestowed on those settlers who would fill the empty reaches beyond the old walls. There was, furthermore, a welcome for Christians, a tolerance of pagan beliefs, and benevolence toward Jews.

Constantinople as a religious capital

Constantinople was also an ecclesiastical centre. In 381 it became the seat of a patriarch who was second only to the bishop of Rome; the patriarch of Constantinople is still the nominal head of the Orthodox Church. Constantine inaugurated the first ecumenical councils; the first six were held in or near Constantinople. In the 5th and 6th centuries emperors were engaged in divising means to keep the Monophysites attached to the realm. In the 8th and 9th centuries Constantinople was the centre of the battle between iconoclasts and the defenders of icons. The matter was settled by the seventh ecumenical council against the iconoclasts, but not before much blood had been spilled and countless works of art destroyed. The eastern and western wings of the church drew further apart, and after centuries of doctrinal disagreement between Rome and Constantinople a schism occurred in the 11th century. The Pope originally approved the sack of Constantinople in 1204, then decried it. Various attempts were made to heal the breach in the face of the Turkish threat to the city, but the divisive forces of suspicion and doctrinal divergence were too strong.

By the end of the 4th century, Constantine's walls had become too confining for the wealthy and populous metropolis. St. John Chrysostom, writing at the end of that century, said many nobles had 10 to 20 houses and owned from one to 2,000 slaves. Doors were often made of ivory, floors were of mosaic or were covered in costly rugs, and beds and couches were overlaid with precious metals.

The pressure of population pressing from within, and the barbarian threat from without, prompted the building of walls further inland at the hilt of the peninsula. These new walls of the early 5th century, built in the reign of Theodosius II, are those that stand today.

In the reign of Justinian I (527–565) medieval Constantinople attained its zenith. At the beginning of this reign the population is estimated to have been about 500,000. In 532 a large part of the city was burned and many of the population killed in the course of the repression of the Nika Insurrection, an uprising of the Hippodrome factions. The rebuilding of the ravaged city gave Justinian the opportunity to engage in a program of magnificent construction, of which many buildings still remain.

In 542 the city was struck by a plague that is said to have killed three out of every five inhabitants; the decline of Constantinople dates from this catastrophe. Not only the capital but the whole empire languished, and slow recovery was not visible until the 9th century. During this period the city was frequently besieged—by the Persians and Avars (626), the Arabs (674 to 678 and again from 717 to 718), the Bulgars (813 and 913), the Russians (860, 941, and 1043), and by a wandering Turkic people, the Pechenegs, (1090–91). All were unsuccessful.

In 1082 the Venetians were allotted quarters in the city itself (there was an earlier cantonment for foreign traders at Galata across the Golden Horn) with special trading privileges. They were later joined by Pisans, Amalfitans, Genoese, and others. These Italian groups soon obtained a stranglehold over the city's foreign trade—a monopoly that was finally broken by a massacre of Italians. Not for some time were Italian traders permitted once more to settle in Galata.

Crusader rule

In 1203 the armies of the Fourth Crusade, deflected from their objective in the Holy Land, appeared before Constantinople—ostensibly to restore the legitimate Byzantine emperor, Isaac II. Although the city fell, it remained under its own government for a year. On April 13, 1204, however, the crusaders burst into the city to sack it. After a general massacre, the pillage went on for years. The crusading knights installed one of themselves, Baldwin of Flanders, as emperor, and the Venetians—prime instigators of the crusade—took control of the church. While the Latins divided the rest of the realm among themselves, the Byzantines entrenched themselves across the Bosporus at Nicaea (now İznik) and at Epirus (now northwestern Greece). The period of Latin rule (1204 to 1261) was the most disastrous in the history of Constantinople. Even the bronze statues were melted down for coin; everything of value was taken. Sacred relics were torn from the sanctuaries and dispatched to religious establishments in western Europe.

In 1261 Constantinople was retaken by Michael VIII Palaeologus, Greek emperor of Nicaea. For the next two centuries the shrunken Byzantine Empire, threatened both from the West and by the rising power of the Ottoman Turks in Asia Minor, led a precarious existence. Some construction was carried out at the end of the 13th and the beginning of the 14th centuries, but thereafter the city was in a state of decay, full of ruins and tracts of deserted ground, contrasting with the prosperous condition of Galata across the Golden Horn, which had been granted to the Genoese by the Byzantine ruler Michael VIII. When the Turks crossed into Europe in the mid-14th century, the fate of Constantinople was sealed. The inevitable end was retarded by the defeat of the Turks at the hands of Timur (Tamerlane) in 1402; but in 1422 the Ottoman sultan of Turkey, Murad II, laid siege to Constantinople. This attempt failed, only to be repeated 30 years later. In 1452 another Ottoman sultan, Mehmed II, proceeded to blockade the Bosporus by the erection of a strong fortress at its narrowest point; this fortress, called Rumeli Hisarı, still forms one of the principal landmarks of the straits. The siege of the city began in April 1453. The Turks had not only overwhelming numerical superiority but also cannon that breached the ancient walls. The Golden Horn was protected by a chain, but the Sultan succeeded in hauling his fleet by land from the Bosporus into the Golden Horn. The final assault was made on May 29, and, in spite of the desperate resistance of the inhabitants aided by the Genoese, the city fell. The last Byzantine emperor, Constantine XI Palaeologus, was killed in battle. For three days the city was abandoned to pillage and massacre, after which order was restored by the Sultan.

Capture by the Turks

CENTURIES OF GROWTH

When Constantinople was captured, it was almost deserted. Mehmed II began to repeople it by transferring to it populations from other conquered areas such as the Peloponnese, Salonika (modern Thessaloníki), and the Greek islands. By about 1480 the population rose to between 60,000 and 70,000. Hagia Sophia and other Byzantine churches were transformed into mosques. The Greek patriarchate was retained, but moved to the Church of the Pammakaristos Virgin (Mosque of Fethiye), later to find a permanent home in the Fener (Phanar) quarter. The Sultan built the Old Seraglio (Eski Saray), now destroyed, on the site occupied at present by the university, and a little later the Topkapı Palace (Seraglio), which is still in existence; he also built the Eyüp Mosque at the head of the Golden Horn and the Mosque of the Fatih on the site of the Basilica of the Holy Apostles. The capital of the Ottoman empire was transferred to Constantinople from Adrianople (Edirne) in 1457.

After Mehmed II, Istanbul underwent a long period of peaceful growth, interrupted only by natural disasters—earthquakes, fires, and pestilences. The sultans and their ministers devoted themselves to the building of fountains, mosques, palaces, and charitable foundations so that the aspect of the city was soon completely transformed. The most brilliant period of Turkish construction coincides with the reign of the Ottoman ruler Süleyman the Magnificent (1520–66).

Westernization of the city

The next major change in the history of Istanbul occurred at the beginning of the 19th century, when dismemberment of the Ottoman Empire was approaching. This period was known as the era of internal reforms (Tanzimat). The reforms were accompanied by serious disturbances, such as the massacre of the Janissaries in the Hippodrome (1826). With the triumph of the progressive Ottoman sultan Mahmud II over the conservative opposition, the westernization of Istanbul started apace. There was an ever-growing influx of European visitors who, since the 1830s, could reach Istanbul by steamship. The first bridge across the Golden Horn was built in 1838. In 1839 the Ottoman sultan Abdülmecid I issued a charter guaranteeing to all his subjects, whatever their religion, the security of their lives and fortunes. The process of westernization was further accelerated by the Crimean War (1853–56) and the quartering of British and French troops in Istanbul. The latter part of the 19th and the beginning of the 20th centuries were marked by the introduction of various public services: the European railroad extending to Istanbul was begun in the early 1870s. The underground tunnel joining Galata to Pera was completed in 1873; a regular water supply for Istanbul and the settlements on the European side of the Bosporus was brought from Lake Terkos on the Black Sea coast (29 miles from the city) by the French company, La Compagnie des Eaux, after 1885; electric lighting was introduced in 1912 and electric street cars and telephones in 1913 and 1914. An adequate sewage system had to wait until 1925 and later.

MODERN ISTANBUL

In the first quarter of the 20th century, there were various disruptions marking the death of the Ottoman Empire and the birth of modern Turkey. In 1908 the city was occupied by the army of the Young Turks who deposed the hated sultan Abdülhamid II. During the Balkan Wars (1912–13) Istanbul was nearly captured by the Bulgarians. Throughout World War I the city was under blockade. After the conclusion of the Armistice (1918) it was placed under British, French, and Italian occupation that lasted until 1923. The Greco-Turkish War in Asia Minor, as well as the Russian Revolution, brought thousands of refugees to Istanbul. With the victory of the Nationalists under Mustafa Kemal Atatürk, the sultanate was abolished, and the last Ottoman sultan, Mehmed VI, fled from Istanbul (1922). After the signing of the Lausanne Treaty, Istanbul was evacuated by the Allies (October 2, 1923), and Ankara was chosen as the capital of Turkey (October 13, 1923). On October 29, the Turkish Republic was proclaimed. Because of Turkey's neutrality during most of World War II, Istanbul suffered no damage, although a German invasion was feared after the Balkans had been conquered by the Axis. The influx of automobiles brought acute traffic problems to Istanbul, and large tracts of the city were demolished or cleared to make way for modern highways.

BIBLIOGRAPHY

Antiquities: A. VAN MILLINGEN, *Byzantine Constantinople: The Walls of the City and Adjoining Historical Sites* (1899); PHILIP SHERRARD, *Constantinople: Iconography of a Sacred City* (1965); PHILIP GRIERSON, *The Tombs and Obits of the Byzantine Emperors, 337–1042* (1962); BERNARD LEWIS, *Istanbul and the Civilization of the Ottoman Empire* (1963); DEAN A. MILLER, *Imperial Constantinople* (1969); MICHAEL MACLAGAN, *The City of Constantinople* (1968), well illustrated, with good annotated bibliography and index.

Churches: W.R. LETHABY and H. SWAINSON, *The Church of Sancta Sophia* (1894); T. WHITTEMORE, *The Mosaics of St. Sophia at Istanbul,* 4 vol. (1933–52); PAUL ATKINS UNDERWOOD, *The Kariye Djami* (1966).

Contemporary descriptions: E. MAMBOURY, *Istanbul touristique* (1951); ROBERT BOULANGER, *Istanbul et ses environs* (1957; Eng. trans. 1960); PETER MAYNE, *Istanbul* (1967); JOHN FREELY, *Istanbul* (1983), in the "Blue Guide Series."

(B.E.)

Italian Literature

Vernacular Italian literature had its beginnings in the 13th century. Until that time nearly all literary work was written in Latin, was predominantly practical in nature, and was produced by writers trained in ecclesiastical schools. In literary quality and variety it fell short of the standard set by France. Only small fragments of Italian vernacular verse before the end of the 12th century have been found, although a number of legal documents were written in the vernacular.

This article traces the development of Italian literature from early French-inspired works to the diverse writings of present-day authors.

The article is divided into the following sections:

EARLY VERNACULAR LITERATURE

Franco-Italian literature. French prose and verse romances were popular in Italy from the 12th to the 14th century. Stories from the Carolingian and Arthurian cycles, together with free adaptations from the classics, were read by the literate, while French minstrels recited verse in public places throughout northern Italy. By the 13th century a "Franco-Italian" literature had developed; Italians copied French stories, often adapting and extending various episodes and sometimes creating new romances in French about characters from French works. In this literature, though the language used was French, the writers often unconsciously introduced elements from their own dialects, according to their varying knowledge of French. Writers of important prose works such as Martino da Canale and Brunetto Latini—who wrote *La Cronique des Veniciens* (1275; "Chronicle of the Venetians") and *Li Livres dou trésor* (*c.* 1260; "The Books of the Treasure"), respectively—were much better acquainted with French, while poets such as Sordello of Mantua wrote lyrics in Provençal revealing an exact knowledge of the language and of Provençal versification. Provençal love lyrics were, in fact, as popular as the French romances, and Italian writers carefully studied anthologies of the verse.

The Sicilian school. In the cultured environment of the Sicilian court of the Holy Roman emperor Frederick II, who ruled the Sicilian kingdom from 1208 to 1250, lyrics modeled on Provençal forms and themes were written in

The origins of the vernacular lyric

the vernacular. Poets were careful to eliminate narrowly local elements from their language and used words characteristic of the troubadour tradition. Poetry was considered an escape from serious matters of life, and it is significant that it was the love poetry of Provence—and not the political poetry—that was imitated by the Sicilian school.

The Tuscan poets. Sicilian poetry continued to be written after the death of Frederick II, but the centre of literary activity moved to Tuscany, where interest in the Sicilian lyric had led to several imitations by Guittone d'Arezzo and his followers. Although Guittone experimented with elaborate verse forms, his language mingled dialect elements with Latinisms and Provençalisms and had none of the beauty of the southern school.

The new style. While Guittone and his followers were still writing, a new development appeared in love poetry, marked by a concern for precise and sincere expression and a new, serious treatment of love. It became customary to speak of a new school of poets of the *dolce stil novo,* or *nuovo* ("sweet new style"), an expression used by Dante Alighieri in *La divina commedia* (*Purgatorio,* Canto XXIV, line 27) in a passage where he emphasized delicacy of expression suited to the subject of love. The major *stil novo* poets were Guido Guinizelli of Bologna, Guido Cavalcanti, Dante (in his poems in *La vita nuova*), and Cino da Pistoia, together with the lesser poets Lapo Gianni, Gianni Alfani, and Dino Frescobaldi.

Dante's *Vita nuova*

These poets seem to have been influenced by each other's work. Guido Guinizelli was best known for his canzone, or poem, beginning "Al cor gentil ripara sempre amore" ("Love always finds shelter in the gentle heart"), which posed the question of the relationship between love of woman and love of God. His poetry was immediately appreciated by Cavalcanti, a serious and extremely talented lyric poet. Most of his poems were tragic and denied the ennobling effect of love suggested by Guinizelli. Dante greatly admired Cavalcanti, but his concept of love, inspired by his love for Beatrice, who died young (in 1290), had much more in common with Guinizelli's. Dante's *Vita nuova* (*c.* 1293; *The New Life*) is the story of his love in poems linked by a framework of eloquent prose: God is the "root" of Beatrice, and she is able to mediate God's truth and love and inspire love of God—but her death is necessary for her lover to reach a state of purification. Cino da Pistoia used the vocabulary of the *stilnovisti,* as these poets were called, in an original and lighthearted way. A comparison of their language with the earlier Tuscan poets reveals extensive refinement of Tuscan dialects. Purely local characteristics were removed, and the standard literary language of Italy had been created.

Comic verse. Giocoso, or comic, verse was a complete contrast to serious love poetry. The language was often deliberately unrefined, colloquial, and sometimes obscene, in keeping with the themes dealt with in the poetry. This kind of verse belongs to a European tradition, owing something to the satirical Goliard poets of the 12th and 13th centuries, who wrote Latin verses in praise of pleasure or in vituperation of women or personal enemies or the church. The comic poets—whose usual verse form was the sonnet—were all cultured men. The earliest of them was Rustico di Filippo, who produced both courtly love poetry and coarse, sometimes obscene verse of the "realistic" kind. The best known and most versatile was Cecco Angiolieri, whose love poetry often skillfully parodied the *stil novo* writers and whose favourite subject was his father's meanness. Folgore di San Gimignano was known for his sonnets (following Latin models) on the worldly pleasures he considered as suitable to different months and different days of the week.

Religious poetry. The beautiful and famous *Cantico di frate sole* (*c.* 1225; "Canticle of the Sun") of St. Francis of Assisi was one of the earliest Italian poems. It was written in rhythmical prose that used assonance in place of rhyme and was in the Umbrian dialect; in it God is praised through all the things of his creation. It is probable that St. Francis also composed a musical accompaniment, and after his death the *lauda* became a common form of religious song used by the confraternities of lay people who gathered to sing the praises of God and the saints and to recall the life and Passion of Christ. The one real poet of the *laude* was Jacopone da Todi, a Franciscan and a mystic. His *laudi* were mostly concerned with the theme of spiritual poverty. Though Jacopone did not write with conscious art, some of the *laudi* were very fine.

The pessimism of northern Italian poetry

In northern Italy religious poetry was mainly moralistic and pervaded by a pessimism rooted in heretical ideas derived from Manichaeism, which saw the world and the body as being evil and under Satan's control. Giacomino da Verona, a Franciscan, author of *De Jerusalem celesti* (*c.* 1250; "On Heavenly Jerusalem") and *De Babilonia civitate infernali* (*c.* 1250; "On the Infernal Babylonian State"), was the liveliest and most imaginative of this group.

Prose. Literary vernacular prose began in the 13th century, though Latin continued to be used for writings on theology, philosophy, law, politics, and science.

The founder of Italian rhetorical prose style, Guido Faba, a rhetorician, illustrated his teaching by examples of prose styles in Bolognese. Guittone d'Arezzo, his most notable follower in epistolography, tended toward an extravagant style. In contrast with Guittone's style is the clear scientific prose of Ristoro d'Arezzo's *Della composizione del mondo* (1282; "On the Composition of the World") and the simple narrative of the Florentine collection of tales *Il novellino* (written in the late 13th century, published in 1525 as *Le ciento novelle antike; Il Novellino, the Hundred Old Tales*). The masterpiece of 13th-century prose is Dante's *Vita nuova.* Though not yet completely at ease in vernacular prose, Dante combined simplicity with great delicacy and a poetic power that derived from the mysterious depth beneath certain key words.

THE 14TH CENTURY

The literature of 14th-century Italy dominated Europe for centuries to follow and may be regarded as the starting point of the Renaissance. Three names stand out: Dante, Petrarch, and Boccaccio.

Dante. Dante Alighieri is one of the most important and influential names in all European literature, but it was only after his exile (1302) that he set out to write more ambitious works. *Il convivio* (*c.* 1304–07; "The Banquet"), revealing his detailed knowledge of scholastic philosophy, was the first great example of a treatise in vernacular prose: its language avoided the ingenuousness of popular writers and the artificiality of the translators from Latin. *De vulgari eloquentia* ("On Vernacular Eloquence"), written about the same time, contained the first theoretical discussion and definition of the Italian literary language. Both these works remained unfinished. In a later doctrinal work, *De monarchia* (written *c.* 1313; Eng. trans., *De Monarchia*), Dante expounded his political theories, which demanded the coordination of the two medieval powers, pope and emperor.

La divina commedia

Dante's genius found its fullest development in *La divina commedia* (written *c.* 1310–21; *The Divine Comedy*), an allegorical poem in terza rima (stanzas of three lines of 11 syllables each, rhyming *aba, bcb, cdc,* etc.), the literary masterpiece of the Middle Ages and one of the greatest products of any human mind. The central allegory of the poem was essentially medieval, taking the form of a journey through the world beyond the grave, with, as guides, the Roman poet Virgil and Beatrice, who symbolize reason and faith, respectively. The poem is divided into three *cantiche,* or narrative poems: *Inferno, Purgatorio,* and *Paradiso.* They are subdivided into various schematically symbolic areas, and each contains 33 cantos, with one canto as an overall prologue. Dante, through his experiences and encounters on the journey, gains understanding of the gradations of damnation, expiation, and beatitude, and the climax of the poem is his momentary vision of God. The greatness of the poem lies in complex imaginative power of construction, inexhaustible wealth of poetry, and continuing significance of spiritual meanings.

Petrarch. The intellectual interests of Petrarch (Francesco Petrarca, died 1374) were literary rather than philosophical; his political views were more realistic than Dante's and his poetic technique more elaborate though less powerful. Petrarch's influence on literature was enormous and lasting—stretching through the Italian humanists of the following century to poets and scholars throughout western Europe. He rejected medieval Scholasticism and took as his models the classical Latin authors and the Church Fathers. This convergence of interests is apparent in his philosophical and religious works. Humanist ideals inspired his Latin poem *Africa* (begun *c.* 1338) and his historical works, but the autobiographical *Secretum meum* (written 1342–58; *Petrarch's Secret*) is most important for a full understanding of his conflicting ideals. The *Rime,* or *Canzoniere*—a collection of sonnets, songs, six-line verses, ballads, and madrigals, dating from 1330 until his death—gave these ideals poetic expression. Although this collection of vernacular poems intended to tell the story of his love for Laura, it was in fact an analysis and evocation not of present love but of the passion that he had overcome. The main element of this poetry was therefore in the elaboration of its art, even if it always reflected the genuine spiritual conflicts exposed in the *Secretum.* In addition to the *Canzoniere* Petrarch wrote a vernacular allegorical poem, *Trionfi* (1351–74; *Triumphs*), in the medieval tradition, but it lacked the moral and poetical inspiration of Dante's great poem.

The literary phenomenon known as Petrarchism developed rapidly within the poet's lifetime and continued to grow during the following three centuries, deeply influencing the literatures of Italy, Spain, France, and England. His followers did not merely imitate but accepted his practice of strict literary discipline and his forms, including that for the sonnet—without which the European literary Renaissance would be unthinkable.

Boccaccio. Boccaccio's writings were purely literary, without any ideological implications. His first romance, *Il filocolo* (*c.* 1336; "Love's Labour"), derived from the French romance *Floire et Blancheflor,* was little more than literary experiment. Inability to write on an epic scale was evident in his two narrative poems *Il filostrato* (*c.* 1338; "Frustrated by Love") and *Teseida* (*c.* 1340; *The Book of Theseus*), while his *Ameto* (1341–42), a novel written in prose and verse, and his *Fiammetta* (*c.* 1343; *Amorous Fiammetta*), a prose novel, showed the influence of classical literature on the formation of his style. The *Decameron* (1348–53), a prose collection of 100 stories divided into 10 "days," was Boccaccio's most mature and important work. Its treatment of contemporary urban society ranged from the humorous to the tragic. Stylistically the most perfect example of Italian classical prose, it had enormous influence on Renaissance literature.

The *Decameron*

Boccaccio shared the humanist interests of his age, as shown in his Latin epistles and treatises. An admirer of Dante, he also wrote a *Trattatello in laude di Dante* (written *c.* 1360; "Treatise in Praise of Dante"; Eng. trans., *The Life of Dante*) and a commentary on the first 17 cantos of the *Inferno.* He contributed to allegorical poetry with *L'amorosa visione* (written 1342–43).

Popular literature and romances. During the second half of the 14th century, Florence remained a centre of culture, but its literature developed a more popular character. The best known representative of this development was Antonio Pucci (died 1388), whose vast production included the *Centiloquio* (written after 1348; "One Hundred Tales"), a versification of Giovanni Villani's *Cronica.* Florentine narrative literature was represented by the *Pecorone* (written *c.* 1378; "Dullard"), stories by Ser Giovanni Fiorentino after a pattern established by Boccaccio, and Franco Sacchetti's *Trecentonovelle* (written *c.* 1390; "300 Short Stories"), which provide colourful and lively descriptions of people and places.

The recasting of the Carolingian and Arthurian cycles continued along lines established during the 13th century. Compilations in prose and verse became commoner, and Franco-Venetian literature gained in literary value. Epic legends were turned into romantic stories, which appealed more to their audiences in town squares and other public places. Novels by Andrea da Barberino, *cantari* with legendary subjects by Antonio Pucci, and the anonymous *Pulzella gaia, Bel Gherardino, Donna del Vergiù,* and *Liombruno* were written in the popular style consistent with their practical aim.

Religious and historical literature. The most important author of religious literature was Jacopo Passavanti, whose *Specchio di vera penitenza* ("The Mirror of True Penitence") was a collection of sermons preached in 1354. Less polished, but of greater literary value, were translations of legends collected in the anonymous *Fioretti di San Francesco* (*The Little Flowers of St. Francis of Assisi*).

Vernacular historiography of this period could be described as popular literature, with Florence as its main centre, whose two principal chroniclers were Dino Compagni and Giovanni Villani. Compagni wrote his chronicle between 1310 and 1312, after having taken part in the political struggles of his town; his dramatic account of the episodes and the liveliness of his prose made it the most original work of medieval historiography. Villani's *Cronica* in 12 books, written from 1308 to 1348, was less personal; it followed the medieval tradition by beginning with the building of the Tower of Babel and included many legends. The last six books, which cover the period from Charles of Anjou's Italian expedition (1265) to the author's own time, are of importance to historians. His prose lacked the dramatic power of Compagni's, but his work may be described as the greatest achievement of Italian vernacular historiography during the Middle Ages.

The Florentine chroniclers

From Boccaccio's death to about the middle of the 15th century, Italian poetry suffered a decline. The following period was to be characterized by critical and philological activity rather than by original creative work.

THE RENAISSANCE

The age of humanism. The European Renaissance had really begun in 14th-century Italy with Petrarch and Boccaccio. The 15th century, devoid as it was of major poetic works, was nevertheless of very great importance because it was the century in which a new vision of human life, embracing a different conception of man and life as well as more modern principles of ethics and politics, gradually found its expression. This was the result, on the one hand, of the rediscovery of classical antiquity, and, on the other, of political conditions quite different from those of previous centuries. With regard to the second point, nearly all Italian princes competed with each other in the 15th century to promote culture by patronizing research, offering hospitality and financial support to literary men of the time, and founding libraries. As a consequence, their courts became centres of research and discussion, thus making possible the great cultural revival of the period. The most notable courts were that of Florence, under Lorenzo de' Medici, "the Magnificent"; that of Naples, under the Aragonese kings; that of Milan, first under the Visconti and later the Sforza family; and finally the papal court at Rome, which gave protection and support to a large number of Italian and Byzantine scholars. As for the first point, the search for lost manuscripts of ancient authors, begun in the second half of the previous century, led to an extraordinary revival of interest in classical antiquity: in particular, much research was devoted to Plato and Greek philosophy in general, a fact that was to have profound influence on the thinking of the Renaissance as a whole.

By and large, the new culture of the 15th century was a revaluation of man. Humanism opposed the medieval view of man as a being with relatively little value and extolled him as the centre of the universe, the power of his soul as linking the temporal and the spiritual, and earthly life as a realm in which the soul applies its powers. These concepts, which mainly resulted from the new interest in Plato, were the subject of many treatises, the most important of which were Giannozzo Manetti's *De dignitate et excellentia hominis* (completed in 1452; *On the Dignity of Man*) and Pico della Mirandola's *Oratio de hominis dignitate* (written 1486; *Oration on the Dignity of Man*). The humanist vision evolved during this period condemned many religious tenets of the Middle Ages still widely prevalent: monastic ideals, for example, were attacked by Leonardo Bruni, Lorenzo Valla, and Poggio Bracciolini. Forthright though these attacks were, humanism was not essentially anti-Christian, for it generally remained faithful to Christian beliefs, and the papal court itself regarded humanism as a force to be assimilated rather than defeated.

Italian humanist view of man

In the first half of the century humanists, with their enthusiasm for Latin and Greek literature, had a disdain for the Italian vernacular. Their poetic production, inspired by classical models and written mostly in Latin or occasionally Greek, was abundant but of little value. Writing in a dead language and closely following a culture to which they had enslaved themselves, they rarely showed originality as poets. Among the few notable exceptions are Giovanni Pontano, Michele Marullo Tarcaniota, Angelo Ambrogini Poliziano (Politian), and Jacopo Sannazzaro. These poets sometimes succeeded in creating sincere poetry in which the conventional themes of 15th-century lyrics were expressed with new, original intimacy and fervour.

The rise of vernacular literature. Toward the middle of the 15th century Italian began to oust Latin as the literary language. The Certame Coronario, a public poetry competition held in Florence in 1441 with the intention of proving that the spoken language was in no way inferior to Latin, marked a definite change. In the second half of the century there were a number of works of merit inspired either by Carolingian legends or by the new humanist culture.

The chivalrous epic of Carolingian legends, which had degenerated into clichés, was given a new lease on life by two poets of very different temperament and education: Matteo Maria Boiardo, whose *Orlando innamorato* (1483; "Orlando in Love") reflected past chivalrous ideals as well as contemporary standards of conduct and popular passions; and Luigi Pulci, whose *Morgante,* published before 1480, was pervaded by a new bourgeois and popular morality.

The new ideals of the humanists were most complete in Angelo Ambrogini Poliziano (Politian), Jacopo Sannazzaro, and Leon Battista Alberti, three outstanding figures who combined a wide knowledge of classical antiquity with a personal and often profound inspiration. Politian's most important Italian work was *Stanze per la giostra* (1475–78; "Stanzas for the Joust"), which created a mythical world in which concepts of classical origin were relived in a new way. The same could be said of Sannazzaro's *Arcadia* (1504), a largely autobiographical work in verse and prose that remained widely influential up to the 18th century. A more balanced view of contemporary reality was given in Alberti's literary works, which presented a gloomy picture of human life, dominated by man's wickedness and the whims of fortune. As for Lorenzo de' Medici, patron of many men of letters, he himself had a vast poetic output, though this is more notable for documentary than for literary value.

Pietro Bembo's view of language

Pietro Bembo of Venice published his *Prose della volgar lingua* ("Writings in the Vulgar Tongue") in 1525. In this work, which was one of the first Italian grammars, Bembo demanded an Italian literary language based on 14th-century Tuscan models, particularly Petrarch and Boccaccio. He was opposed by those who thought that a literary language should be based on existing linguistic developments, particularly by Gian Giorgio Trissino, who developed Dante's theories on Italian as a literary language. In practice the problem was both linguistic and stylistic, and there were in the first half of the 16th century a great number of other contributors to the question, but Bembo's theories did finally triumph in the second part of the century. This was due to a large extent to the activities of the Florentine Accademia della Crusca, and this more scientific approach to the language question resulted in the academy's first edition of an Italian dictionary in 1612.

During the first decades of the 16th century, treatises on poetry were still composed according to humanist ideas and the teachings of the Roman Augustan poet Horace. It was only after 1536, when the original classical Greek text of Aristotle's *Poetics* was first published, that a gradual development became apparent in aesthetic theory. The traditional principle of imitation was now better analyzed, emphasis being given to the imitation of classical authors rather than to that of nature. The three unities of tragedy (time, space, action) were among the rhetorical rules then reestablished. The classical conception of poetry as a product of imagination supported by reason was at the basis of 16th-century rhetoric, and it was this conception of poetry, revived by Italian literature, that triumphed in France, Spain, and England during the following century.

Political, historical, biographical, and moral literature. Niccolò Machiavelli's works reflected Renaissance thought in its most original aspects, particularly in the objective analysis of human nature. Machiavelli has been described as the founder of a new political science: politics divorced from ethics. His own political experience was at the basis of his ideas, which he developed according to such general principles as the concepts of *virtù* ("power") and *fortuna* ("chance"). He considered *virtù* to be power with a practical aim that should struggle against *fortuna,* which represented the forces of violence and irresponsibility. His famous treatise *Il principe* (*The Prince*), composed in 1513, revealed the author's prophetic attitude, based on observation of contemporary political affairs. Its description of a model ruler became a code for the wielding of absolute power throughout Europe for two centuries. Machiavelli's *Discorsi sopra la prima deca di Tito Livio* (*c.* 1513–21; *Discourse on the First Ten Books of Titus Livius*), showed the same realistic attitude: public utility was placed above all other considerations, and political virtue was distinguished from moral virtue. His seven books on *Dell'arte della guerra* (1521; *The Art of War*), concerning the creation of a modern army, were more technical, while his historical works, including the *Istorie fiorentine* (1520–25; *Florentine History*), exemplified theories expounded in his treatises. Machiavelli also holds a place in the history of imaginative literature, above all for *La Mandragola* (1518), one of the outstanding comedies of the century.

The Prince

Although more of an individualist and pragmatist than Machiavelli, Francesco Guicciardini was the only 16th-century historian who could be placed within the framework of the political theories he constructed. He drew attention to the self-interest of those involved in political actions and made Machiavelli's theories appear idealistic by contrast. One of Guicciardini's main works, his *Ricordi* (1512–30; "Remarks"), has a place among the most original political writings of the century. Guicciardini was also the first, in his *Storia d'Italia* (1537–40), to compose a truly national history of Italy, setting it in a European context and attempting an impartial analysis of cause and effect.

Giorgio Vasari's *Vite de' più eccellenti architetti, pittori et scultori italiani da Cimabue insino a' tempi nostri* (1568; *Lives of the Painters, Sculptors, and Architects*) not only contained more than 200 biographies but also was the first proper critical and historical appraisal of Italian art. The autobiography of the sculptor and goldsmith Benvenuto Cellini (written 1558–66, published 1728) was remarkable for its spontaneity and its use of popular Florentine language.

Castiglione's *Cortegiano*

The highest moral aspirations of the Renaissance are expressed in Baldassare Castiglione's *Cortegiano* (published 1528; *The Courtier*), which deals with the perfect courtier, the noble lady, and the relationship between courtier and prince. It became one of the most influential books of the century. Giovanni della Casa was the author of another famous treatise, the *Galateo* (*c.* 1551–54; "Manners"; Eng. trans., *Galateo*), a book on courtesy in which the author's witty mind and the refinement of contemporary Italian society found full expression. The life of the period was also vividly reflected in the work of Pietro Aretino, who was called "the scourge of princes" by Ariosto. His *Ragionamenti* (1534-36; "Discussions") were written in a spontaneous style and showed a sensuous and unscrupulous nature.

Poetry. Lyric poetry in the 16th century was dominated by the model of Petrarch mainly because of the acceptance of the Renaissance theory of imitation and the teaching of Bembo. Almost all the principal writers of the century wrote lyric poems in the manner of Petrarch. Some originality was to be found in Della Casa's poems, and Galeazzo di Tarsia stood out from contemporary poets by virtue of a vigorous style. Also worthy of note are the passionate sonnets of the Paduan poet Gaspara Stampa and those of Michelangelo.

The tradition of autobiographical, humorous, and satirical verse was kept alive during the 16th century, when it reached some real stature with Francesco Berni, whose burlesque poems, mostly dealing with indecent or trivial subjects, showed his stylistic skill. Didactic poetry, already cultivated by humanist writers, was also continued during this period, chiefly by Giovanni Rucellai, who recast in *Le api* (1539; "The Bees") the fourth book of the Roman poet Virgil's *Georgics,* and Luigi Alamanni, in six books about rustic life called *La coltivazione* (1546).

The most refined expression of the classical taste of the Renaissance was to be found in Ludovico Ariosto's *Orlando furioso* (1516; Eng. trans., *Orlando Furioso*), which embodied many episodes derived from popular medieval and early Renaissance epics; but the poem's unique qualities derived from Ariosto's sustained inspiration and technique and detached ironical attitude to his characters. *Orlando furioso* was the most perfect expression of the literary tendencies of the Italian Renaissance at this time, and it exercised enormous influence on later European Renaissance literature. Ariosto also composed comedies that, by introducing imitation of Latin comedy, marked the beginning of Renaissance drama in the vernacular.

There were also attempts to renew the epic by submitting the tradition of chivalry to Aristotle's rules of composition. Gian Giorgio Trissino, a theorist on language, wrote according to the strictest Aristotelian rules, while Luigi Alamanni tried to focus the narrative on a single character in *Girone il cortese* (1548; "Girone the Courteous") and *Avarchide* (1570), an imitation of the *Iliad* of Homer. Giambattista Giraldi, while more famous as a tragic playwright, was a literary theorist who tried to apply his theories to his own poem *Ercole* (1557; "Hercules").

Fidenziana and macaronic poetry

Two burlesque medley forms of verse were invented during the century. *Fidenziana* poetry derives its name from a work by Camillo Scroffa, a poet who wrote in a combination of Latin words and Italian form and syntax. Macaronic poetry, on the other hand, is a term given to verse consisting of Italian words used according to Latin form and syntax. Teofilo Folengo, a Benedictine monk, was the best representative of macaronic literature, and his masterpiece was a poem in 20 books called *Baldus* (1517). A tendency to parody, ridiculing the excesses of humanist literature, was present in both the *fidenziana* and macaronic verse.

Torquato Tasso was the last great poet of the Italian Renaissance and one of the greatest of Italian literature. In his epic *Gerusalemme liberata* (1581; *Jerusalem Delivered*) he summed up a literary tradition typical of the Renaissance: the classical epic renewed according to the spiritual interests of his own time. Whereas most of Tasso's work shows a conflict between a desire to express himself according to classical ideals and a tendency to moralize, some of his works reflected his spontaneous inspiration. *L'Aminta* (1573), a joyous and uninhibited drama, was the best example of Tasso's youthful poetry and belonged to the new literary genre of pastoral (dealing with idealized rural life). *Gerusalemme liberata*, however, was the result of a balance in the poet's conflicting aspirations: a Christian subject dealt with in a classical way. In the subsequent *Gerusalemme conquistata* (1593; "Jerusalem Vanquished"), Tasso recast his poem according to strict Aristotelian rules and the ideals of the Roman Catholic Church's reaction against the Protestant Reformation, known as the Counter-Reformation. Tasso's conflict had ended in the victory of the moralistic principle: poetically the new poem was a failure. Tasso also wrote shorter verses throughout his life, including religious poems, and his prose works show a style no longer exclusively dominated by classical models.

Drama. Trissino's *Sofonisba* (written 1514–15) was the first tragedy of Italian vernacular literature; its structure derived from Greek models, but its poetic qualities were somewhat mediocre. Toward the middle of the 16th century Giambattista Giraldi (Cinzio) reacted against imitation of Greek drama by proposing the Roman tragedian Seneca as a new model, and in nine tragedies and tragicomedies—written between 1541 and 1549—he showed some independence from Aristotelian rules. He greatly influenced European drama, particularly the English of the Elizabethan period.

Italian comedies

Italian comedies of the century, inspired by Latin models, possessed greater artistic value than tragedies, and they reflected contemporary life more fully: they could be considered as the starting point for modern European drama. To the comedies of Ariosto and Machiavelli should be added a lively play, *La Calandria* (first performed 1513; *The Follies of Calandro*), by Cardinal Bernardo Dovizi da Bibbiena, and five equally amusing comedies written by Pietro Aretino. Giordano Bruno, a great Italian philosopher who wrote dialogues in Italian on his new cosmology and antihumanist ideas, also wrote a comedy, *Il candelaio* (1582; *The Candlemaker*).

Since the mid-20th century Angelo Beolco ("Il Ruzzante") has become generally recognized as the most powerful dramatist of the 16th century. His works, written in rustic Paduan dialect, treat the problems of the countryside with profound seriousness. Another dialect playwright of the same century, now also more widely appreciated, is the Venetian Andrea Calmo, who showed a nice gift for characterization in his comedies of complex amorous intrigue.

Narrative. The classicist trend established by Pietro Bembo also affected narrative literature, for which the obvious model was Boccaccio's *Decameron.* Originality and liveliness of expression were to be found in the 22 stories called *Le cene* (written after 1549; "The Suppers") of the Florentine apothecary Anton Francesco Grazzini. The worldly monk Agnolo Firenzuola produced several stories, including the fable *Asino d'oro* (1550), a free version of Apuleius' *Golden Ass.* The cleric and short-story writer Matteo Bandello started a new trend in 16th-century narrative with 214 stories that were rich in dramatic and romantic elements while not aiming at classical dignity. This trend was partly followed also by Giraldi in his collection of 112 stories called *Gli ecatommiti* (1565; "The Hundred Stories").

17TH-CENTURY LITERATURE

The 17th century in Italian literature is usually described as a period of "decadence" in which writers who were devoid of sentiment resorted to exaggeration and tried to cloak an utter poverty of matter beneath an exuberance of form. (In this period, it is said, freedom of thought was fettered by the Accademia della Crusca of Florence, whose aim it was to maintain the purity of the Tuscan tongue; by the Counter-Reformation; and by the political supremacy of Spain.) This style of writing was not, however, simply an Italian phenomenon. It was at this time that "Gongorism" (the ingenious, metaphorical style of the poet Luis de Góngora) flourished in Spain, and the witty, figurative verse of the Metaphysical poets was popular in England. Far from being exhausted, indeed, this was an extremely vital period, so much so that a new and more comprehensive understanding of the literature of the Italian Baroque has been formulated by scholars conversant with the changing attitude toward this phase of civilization in Germany, France, and England.

Poetry and prose. The popularity of satire was a reaction against prevailing conditions. Prominent in this genre was Salvator Rosa, who attacked in seven satires the vices and shortcomings of the age. Alessandro Tassoni acquired great fame with *La secchia rapita* (1622; *The Rape of the Bucket*), a mock-heroic poem that is both an epic and a personal satire. The greatest poet of the period was Tommaso Campanella, a Dominican friar, less well-known for his rough-hewn, philosophical verses than for the *Città del sole* (1602; *Campanella's City of the Sun*), a vision of political utopia, in which he advocated the uniting of humanity under a theocracy based on natural religion.

The influence of Marino

The principal representative of Italian writing during this period was Giambattista Marino, author of a large collection of lyric verse (*La lira* [1608–14; "The Lyre"] and *La sampogna* [1620; "The Syrinx"]), and a long mythological poem, *Adone* (1623). Marino derived inspiration from the poetry of the late 16th century, but his aim—typical of the age—was to excite wonder by novelty. His work is characterized by "conceits" of fantastic ingenuity, farfetched metaphor, sensuality, extreme facility, and a superb technical skill. His imitators were innumerable, and most 17th-century Italian poets were influenced by his work.

Gabriello Chiabrera, soberer in style than Marino, was successful in imitating the metres of classical poetry (especially of the Greek Pindar) and excelled in the composition of musical canzonets (short lyrical verses). Toward the end of the century a patriotic sonneteer, Vincenzo da Filicaia, and Alessandro Guidi, who wrote exalted odes, were hailed as major poets, though Guidi's verse is now seen as little more than rhetoric.

Among prose writers the satirist Traiano Boccalini stood out with *Ragguagli di Parnasso* (1612–13; *Advertisements from Parnassus*) in the fight against Spanish domination. A history of the Council of Trent (which defined Catholic doctrines in reaction to the Reformation) was written by Paolo Sarpi—advocate of the liberty of the Venetian state against papal interference—and a history of the rising of the Low Countries against Spain by Guido Bentivoglio. The Venetian novels of Girolamo Brusoni are still of interest, as are the travels of Pietro della Valle and the tales of Giambattista Basile. All the restless energy of this period reached its climax in the work of Galileo, a scientist who laid the foundations of mathematical philosophy and earned a prominent place in the history of Italian literature through the vigour and clarity of his prose.

The music drama; Accademia dell'Arcadia. With the rise of the music drama and the opera, Italian authors worked to an increasing extent with the lyric stage. Librettos written by poets such as Ottavio Rinuccini were planned with dramatic and musical artistry. During the 17th century a popular spirit entered the opera houses: intermezzi (short dramatic or musical light entertainments) were required between the acts, a practice that undermined the dramatic unity of the performance as a whole, and toward the end of the century every vestige of theatrical propriety was abandoned. The spread of Marino's influence was felt by many to be an abuse. In 1690 the Accademia dell'Arcadia was founded in Rome for the express purpose of eradicating "bad taste." The purpose of Arcadia was in tune with a genuinely felt need. Many of its members were rationalist followers of Descartes with severe classical sympathies, but their reaction consisted mainly in imitating the simplicity of the nymphs and shepherds who were supposed to have lived in the golden age, and thus a new artifice replaced an old one. A typical exponent of the Arcadian lyric was Pietro Metastasio, the 18th-century reformer of the operatic libretto.

18TH-CENTURY DEVELOPMENTS

Reform of the tragic theatre. In 1713 Francesco Scipione Maffei, an antiquary of Verona, produced *Merope*—a tragedy that met with great success and pointed the way toward reform of the Italian tragic theatre. Between 1726 and 1747 Antonio Conti—an admirer of Shakespeare—wrote four Roman tragedies in blank verse. It was not until 1782 and the writing of *Saul,* however, that an important Italian tragedian finally emerged in the person of Vittorio Alfieri. In strong contrast with Pietro Metastasio's and Paolo Rolli's *melodrammi*—librettos set to music or sometimes performed as plays in their own right—Alfieri's tragedies are harsh, bitter, and unmelodious. He chose classical and biblical themes, and through his hatred of tyranny and love of liberty he aspired to move his audience with magnanimous sentiments and patriotic fervour. Alfieri's influence in the Romantic period and the Risorgimento was immense, and, like Carlo Goldoni, he wrote an important autobiography, which gives a revealing account of his struggles to provide Italy with a corpus of drama comparable with that of the other European nations.

Goldoni's reform of the comedy. Metastasio's reform of the operatic libretto was paralleled in the mid-18th century by Goldoni's reform of comedy. Throughout the 17th century the commedia dell'arte—a colourful pantomime of improvisation, singing, mime, and acrobatics, often performed by actors of great virtuosity—had gradually replaced regular comedy, but by the early 18th century it had degenerated into mere buffoonery and obscenity with fixed characters and mannerisms. The dialogue was mostly improvised, and the plot—a complicated series of stage directions, known as the scenario—dealt mainly with forced marriages, star-crossed lovers, and the intrigues of servants and masters. Goldoni succeeded in replacing this traditional type of theatre with written works whose wit and vigour are especially evident when the Venetian scene is portrayed in a refined form of the local dialect. Perhaps because of his prolific output his work has sometimes been thought of as lacking in depth. His social observation is acute, however, and his characters are beautifully drawn. *La locandiera* (1753; "The Innkeeper"; Eng. trans., *Mirandolina*), with its heroine Mirandolina, a feminist before feminism's time, has things to say about class and the position of women that can still be appreciated today. Goldoni's rival and bitter controversialist, fellow Venetian Carlo Gozzi, also wrote comedy and satirical verse.

The world of learning. Giambattista Vico, Ludovico Antonio Muratori, Apostolo Zeno, and Scipione Maffei were writers who reflected the awakening of historical consciousness in Italy. Muratori collected the primary sources for the study of the Italian Middle Ages; Vico, in *Scienza nuova* (1725–44; *The New Science*), investigated the laws governing the progress of the human race and from the psychological study of man endeavoured to infer the laws by which civilizations rise, flourish, and fall. Giovanni Maria Mazzuchelli and Gerolamo Tiraboschi devoted themselves to literary history. Literary criticism also attracted attention; Gian Vincenzo Gravina, Vico, Maffei, Muratori, and several others, while advocating the imitation of the classics, realized that it should be cautious and thus anticipated critical standpoints that were later to come into favour.

The Enlightenment. With the end of Spanish domination and the spread of the ideas of the Enlightenment from France, reforms were gradually introduced in various parts of Italy. The new spirit of the times led men—mainly of the upper middle class—to enquire into the mechanics of economic and social laws. The ideas and aspirations of the Enlightenment as a whole were effectively voiced in such organs of the new journalism as Pietro Verri's periodical *Il Caffè* (1764–66; "The Coffeehouse"). A notable contributor to *Il Caffè* was the philosopher and economist Cesare Beccaria, who in his pioneering book *Dei delitti e delle pene* (1764; *On Crimes and Punishments*) made an eloquent plea for the abolition of torture and the death penalty.

Parini and the literary revival

More than anyone else, Giuseppe Parini seems to embody the literary revival of the 18th century. In *Il giorno* (published in four parts, 1763–1801; "The Day"), a long social satire of the rights of blood, he described a day in the life of a young Milanese patrician and revealed with masterly irony the irresponsibility and futility of a whole way of life. His *Odi* (1795; "Odes"), which are imbued with the same spirit of moral and social reform, are among the classics of Italian poetry.

The satire in the *Sermoni* (1763; "Sermons") of Gasparo Gozzi (elder brother of Carlo) is less pungent, though directed at similar ends, and in his two periodicals—*La Gazzetta veneta* and *L'Osservatore*—he presented a lively chronicle of Venetian life and indicated a practical moral with much good sense. Giuseppe Baretti—an extremely controversial figure who published a critical journal called *La Frusta letteraria* ("The Literary Whip"), in which he castigated "bad authors"—had learned much through a lengthy sojourn in England, where his friendship with Samuel Johnson helped to give independence and vigour, if not always accuracy, to his judgments. The *Viaggi di Enrico Wanton* (1749–64), a philosophical novel in the form of an imaginary voyage by the Venetian Zaccaria Seriman, was the most all-embracing satire of the time.

LITERARY TRENDS OF THE 19TH CENTURY

The 19th century was a period of political ferment in Italy, and many outstanding writers were involved in public affairs. Much of the literature written with a political aim, even when not of intrinsic value, became part of Italy's national heritage and inspired not only those for whom it was written but all who valued freedom.

Romanticism. Foremost among writers in early struggles for his country's unity and freedom from foreign domination was Ugo Foscolo, who reconciled passionate feeling with a formal perfection inspired by classical models. His *Ultime lettere di Jacopo Ortis* (1802; "Jacopo Ortis' Last Letters") was an epistolary story of a young man forced to suicide by frustrated love for both a woman and his fatherland. It was extremely moving and popular, as was a poem, "Dei sepolcri" (1807; "On Sepulchres"), where, in fewer than 300 lines, he wrote lyrically on the theme of inspiration to be found at the tombs of the great, exhorting Italians to be worthy of their heritage. This poem influenced the Italian Risorgimento, or national revival, and a passage in which Florence was praised because it kept in Sta. Croce the ashes of Michelangelo, Machiavelli, and Galileo is still very popular in Italy. Two odes celebrating the divine quality of beauty, 12 sonnets ranking with the best of Petrarch's and Tasso's, and an unfinished poem, "Le grazie" ("The Graces"), also testified to Foscolo's outstanding poetic merit. As an exile in England from 1816 until his death in 1827, he wrote remarkable critical essays on Italian literature for English readers.

In Foscolo patriotism and classicism united to form almost one passion, but Vincenzo Monti was outstanding for mobility of feeling. He saw danger to his country in the French Revolution and wrote "Il pellegrino apostolico" (1782; "The Apostolic Pilgrim") and *La bassvilliana* (1793; *The Penance of Hugo*); Napoleon's victories aroused his praise in "Prometeo" (*c.* 1805; "Prometheus"), "Il bardo della selva nera" (1806; "The Bard of the Dark Wood"), and "La spada di Federico II" (1806; "The Sword of Frederick II"); in "Il fanatismo" and "La superstizione" (1797) he attacked the papacy; later he extolled the Austrians. Thus every great event made him change his mind, through lack of political conviction, yet he achieved greatness in *La bellezza dell'universo* (1781; "The Beauty of the Universe"), the lyrics inspired by domestic affections, and in a translation of the *Iliad,* a masterpiece of Neoclassical beauty.

Classicism, purism, and the reaction of Romanticism

Melchiorre Cesarotti occupied a prominent position in the world of learning at the end of the 18th century, and his translations of James Macpherson's Ossian poetry, *Poesie di Ossian* (1763–72), influenced Foscolo, Leopardi, and others by their mysterious and gloomy fantasy, so alien from classical inspiration; *Saggio sulla filosofia delle lingue* (1785; "Essay on the Philosophy of Languages") was an important essay in the dispute on the Italian language. The trend was toward pedantic classicism as a reaction against an excessive Gallicism favoured by some 18th-century writers. Among the purists was Antonio Cesari, who brought out a new enlarged edition of the *Vocabolario della Crusca* (the first Italian dictionary, published by the Accademia della Crusca in 1612). He wrote *Sopra lo stato presente della lingua italiana* (1810; "On the Present State of the Italian Language") and endeavoured to establish the supremacy of Tuscan and of Dante, Petrarch, and Boccaccio as models. But a Lombard school opposed this Tuscan supremacy. Monti, its leader, issued *Proposta di alcune correzioni ed aggiunte al vocabolario della Crusca* (1817–26; "Proposal for Some Corrections and Additions to the Crusca Dictionary"), which attacked the Tuscanism of the Crusca. By contrast, the patriot Pietro Giordani—for a time a journalistic colleague of Monti—was a great exponent of *purismo.* His views did not stem from literary pedantry, however, but from a concern that all social groups throughout Italy should have a common means of communication. In this respect he was linguistically opposed to the poet Carlo Porta, who lampooned the aristocracy and clergy and expressed sympathy with the humble and wretched in a lively Milanese dialect. All Italy took part in the disputes about language, literature, and politics.

The publication of *Il Conciliatore*

An artificial form of classicism was associated with the Napoleonic domination of Italy, so that when Napoleon fell, forces antagonistic to classicism arose. Literary Romanticism had already won favour with the French, who erroneously thought themselves akin to German Romantics. Between 1816 and 1818 a battle was fought for Romanticism, particularly in Milan, where a Romantic periodical, *Il Conciliatore* (1818–19; "The Peacemaker"), was published. Giovanni Berchet (patriotic poet and author of *La lettera semiseria di Grisostomo,* a manifesto of Romanticism published in 1816), Silvio Pellico, Ludovico di Breme, Giovita Scalvini, and Ermes Visconti were among its contributors. Their efforts were silenced in 1820 when several were arrested by the Austrian police because of their liberal opinions; among them was Pellico, who later wrote a famous account of his experiences, *Le mie prigioni* (1832; *My Prisons*).

Alessandro Manzoni (grandson of Cesare Beccaria) was the chief exponent of Italian Romanticism, but perhaps an even higher claim to fame was his contribution to the resolution of the language problem. In 1821 he started working on a panoramic novel about the lives of simple people placed against a background of major historical events, and, in order that this should be accessible to a wide readership, he decided to write it in an idiom as close as possible to modern educated Florentine speech. This was a formidable enterprise for someone whose first languages were French and Milanese dialect—and to whom spoken Florentine was virtually a foreign tongue—and for the first draft (completed in 1823) he had to resort to Francesco Cherubini's Italian-Milanese dictionary. The second draft was published in 1825–27 under the title *I promessi sposi* (*The Betrothed*); and the final definitive edition came out in 1840–42 after a long, painstaking process of revision aimed at making the text conform more closely with colloquial Florentine usage. The result of this effort was clear, expressive prose—neither pretentious nor provincial—and the way in which the novel caught the public's imagination attested to Manzoni's success in addressing the sort of people to whom conventional literary Italian was almost as remote as Latin. Ironically, Manzoni the innovator became, in his turn, the model for a new kind of purism, with "Manzonians" composing works in an affected Tuscan, and it required authors with fresh ideas—not poor imitators—to continue the task of disencumbering and modernizing written Italian.

Manzoni's genius as a poet showed in the odes "Il cinque maggio" (1821; "The Fifth of May"; Eng. trans., "The Napoleonic Ode"), written on the death of Napoleon, and "Marzo 1821," and in passages of *Inni sacri* (1812–22; *Sacred Hymns*), five poems in celebration of church festivals, describing human affections. His tragedies, *Il conte di Carmagnola* (performed 1820; "The Count of Carmagnola") and *Adelchi* (1822), marked a victory of Romanticism over classicism; they contained passages of great lyrical beauty but lacked strong dramatic power.

Leopardi's pessimism

The foremost Italian poet of the age was Giacomo Leopardi, an outstanding scholar and thinker whose philological works together with his philosophical writings, *Operette morali,* would alone place him among the great writers of the 19th century. Embittered by solitude, sickness, and near penury, from the age of 20 he realized the vanity of hope. Though he developed a doctrine of universal pessimism, seeing life as evil and death as the only comfort, the poetry based on these bitter, despairing premises was far from depressing. Most of Leopardi's poems were contained in one book, *I canti* ("Songs"; Eng. trans., *The Poems of Leopardi*), first published in 1831. Some were patriotic and were once very popular; but the best came from deeper lyrical inspiration. Among them were a meditation on infinity; "A Silvia," on the memory of a girl who died when he was 20; "Le ricordanze," an evocation of his childhood; "Il passero solitario," comparing the lonely poet and a sparrow that sings by itself; and "La quiete dopo la tempesta" and "Il sabato del villaggio," two pictures of village life. They balance depth of meaning and formal beauty, simplicity of diction, intensity, and verbal music.

The Risorgimento and after. Circumstances made it inevitable that Italian Romanticism should become heavily involved with the patriotic myths of the Risorgimento; yet, while this served a useful civic purpose at the time, it did not encourage literature of consistent artistic merit or enduring readability. Of the writings produced by figures associated in some way with Italy's struggle for nationhood, it tends to be the less typical ones that attract attention today: the dialect poetry of Giuseppe Gioacchino Belli describing the life of contemporary papal Rome; verses by

Giuseppe Giusti satirizing petty tyrants, political turncoats, and coarse parvenus; or the works of the republican Roman Catholic from Dalmatia, Niccolò Tommaseo. The undoubted masterpiece of Risorgimento narrative literature is Ippolito Nievo's *Confessioni di un italiano* (published posthumously in 1867; "Confessions of an Italian"; Eng. trans., *The Castle of Fratta*), which marks Nievo as the most important novelist to emerge in the interval between Manzoni and Verga. Giuseppe Mazzini's letters can still be studied with profit, as can the memoirs of Luigi Settembrini (*Ricordanze della mia vita* [1879–80; "Recollections of My Life"]) and Massimo D'Azeglio (*I miei ricordi* [1868; *Things I Remember*]). D'Azeglio's historical novels or those of Francesco Guerrazzi now have a rather limited interest; and Mazzini's didactic writings—of great merit in their good intentions—are generally regarded as unduly oratorical. Giovanni Prati and Aleardo Aleardi, protagonists of the "Second Romanticism," wrote poetry of a sentimentality that helped to provoke a variety of reactive movements, including *scapigliatura* and *verismo*.

The works of Carducci

Giosuè Carducci was an outstanding figure whose enthusiastic support for the national cause during the struggle of 1859–61 was changed to disillusionment by the difficulties in which the new kingdom was involved. The bitterness of some of his poetry revealed frustration and rebelliousness. *Rime nuove* (*The New Lyrics*) and *Odi barbare* (*The Barbarian Odes*), both of which appeared in the 1880s, contained the best of his poetry: memories of childhood, evocations of landscape, laments for domestic sorrows, an inspired representation of historical events, an ambitious effort to resuscitate the glory of Roman history, and an anachronistic but sincere cult of pagan civilization. He tried to adapt Latin prosody to Italian verse, which sometimes produced good poems, but his opposition to Romanticism and his rhetorical tirades provoked a strong reaction, and his metrical reform was short-lived. He was also a scholarly historian of literature, and his literary essays had permanent value, although philosophical criticism such as that of De Sanctis was uncongenial to him. Both his poetry and his criticism were cited when he was awarded the Nobel Prize for Literature for 1906.

A figure connected politically with the Risorgimento but remembered chiefly for his critical writings was Francesco De Sanctis, whose most important works consisted of various critical essays and *Storia della letteratura italiana* (1870–71; *History of Italian Literature*). His main tenet was that literature was to be judged not on its intellectual or moralistic content so much as by the spirit of its "form," and the role of the critic was to discover how this form had been unconsciously and spontaneously conceived by studying its creator's temperament and background and the age in which he lived. De Sanctis was not properly appreciated in his day but came into his own at the turn of the century when Benedetto Croce rescued his works from oblivion.

While Carducci was still alive, Giovanni Pascoli acquired a reputation and succeeded him in the chair of Italian literature at the University of Bologna. His art was often impressionistic and fragmentary, his language occasionally laborious, but his lyricism, at first timid in inspiration in *Myricae* (1891; "Tamarisks"), rose to fuller tones when he attempted the loftier themes of antiquity: Roman heritage and greater Italy. His original vein still found expression in *Canti di Castelvecchio* (1903; "Songs of Castelvecchio") and in the classicism of *Poemi conviviali* (1904; "Convivial Poems"). Later he produced—both in humanistic Latin and in self-consciously elaborate Italian—heroic hymns in honour of two sacred cities, Rome and Turin.

Scapigliati

The veristi and other narrative writers. The patriotic niceties and sentimental Romanticism of much Risorgimento writing inevitably provoked a reaction. The first serious opposition came from the *scapigliati* ("libertines" or "bohemians"), adherents of an antibourgeois literary and artistic movement that flourished in Milan and Turin during the last four decades of the 19th century and whose declared aim was to link up with the most advanced Romantic currents from abroad. Unfortunately the movement—perhaps by its very nature—lacked intellectual cohesion and tended to cultivate the eccentric as an end in itself. The *scapigliati,* however, made a useful contribution in social criticism and in their informal linguistic approach. Among the foremost *scapigliati* were Giuseppe Rovani, whose monumental novel about Milanese life, *I cento anni* (*The Hundred Years*), was issued in installments (1856–58 and 1864–65); Emilio Praga, a poet tormented by contradictions; and Arrigo Boito, poet, musician, and librettist for Giuseppe Verdi's *Falstaff* and *Otello.*

Verismo and *veristi*

A more lasting and fruitful successor to conventional Italian Romanticism was *verismo* ("realism"; first theoretically expounded by Luigi Capuana in 1872), a movement initially inspired by the French Naturalist writers and influenced by positivist and determinist ideas. The *veristi* were not concerned with sermons or noble sentiments but with observable phenomena. When they dealt with the Risorgimento, they showed it warts and all. The greatest of *verismo* narrators was without a doubt Giovanni Verga, who explained in a preamble to a short story, "L'amante di Gramigna" (1880; Eng. trans., "Gramigna's Lover"), that in a perfect novel the sincerity of its reality would be so evident that the hand of the artist would be absolutely invisible and the work of art would seem to have matured spontaneously without any point of contact with its author. At times Verga almost seems to have achieved this unattainable goal, and in his two great narrative works dealing with the victims of social and economic change, *I malavoglia* (1881; "The Unwilling"; Eng. trans., *The House by the Medlar Tree*) and *Mastro-don Gesualdo* (1889), the reader often has the sensation of being put down in an unfamiliar milieu and—as would happen in real life—left to pick up the threads from gossip and chance remarks. Another *verista,* Federico De Roberto, in his novel *I vicerè* (1894; *The Viceroys*), has given a cynical and wryly funny account of an aristocratic Sicilian family that adapted all too well to change. Luigi Capuana, the founder of *verismo* and most rigorous adherent to its impersonal method of narration, is known principally for his dramatic psychological study, *Il marchese di Roccaverdina* (1901; "The Marquis of Roccaverdina").

In their search for documentary exactitude the *veristi* paid close attention to background. For Verga, De Roberto, and Capuana, this was Sicily. Matilde Serao, on the other hand, has given a detailed and colourful reportage of the Neapolitan scene, while Renato Fucini conveyed the atmosphere of traditional Tuscany. Emilio De Marchi, another writer in the realist mold, has Milan for his setting and in *Demetrio Pianelli* (1890) has painted a candid but essentially kindly portrait of the new Milanese petite bourgeoisie. Antonio Fogazzaro was akin to the *veristi* in his powers of observation and in his descriptions of minor characters; but he was strongly influenced by Manzoni, and his best narrative work, *Piccolo mondo antico* (1895; *The Little World of the Past*), is a nostalgic look back to a supposedly less individualistic age when inner tranquillity was seemingly achieved by devotion to a shared ideal. The *veristi* had a leavening effect on Italian literature generally, and their influence can be discerned, among others, in the early novels of the Sardinian Grazia Deledda (awarded the Nobel Prize for Literature for 1926) and in the distinguished narrative works of the Sienese writer Federigo Tozzi, including *Con gli occhi chiusi* (1919; "With Closed Eyes") and *Tre croci* (1920; *Three Crosses*).

THE 20TH CENTURY

Gabriele D'Annunzio's nationalism. After unification the new Italy was preoccupied with practical problems, and by the early 20th century a great deal of reasonably successful effort had been directed toward raising living standards, promoting social harmony, and healing the split between church and state. It was in this prosaic and pragmatic atmosphere that the middle classes—bored with the unheroic and positivist spirit of former decades—began to feel the need for a new myth. Thus it is easy to understand how imaginations across the political spectrum came to be fired by the extravagant personality of Gabriele D'Annunzio—man of action, nationalist, literary virtuoso, and (not least) exhibitionist—whose life and art seemed to be a blend of Jacob Burckhardt's "complete man" and the superman of Friedrich Nietzsche. At a distance from those times, it should be possible to evaluate D'Annunzio more clearly. There is, however, no critical consensus about his writings, although he is generally praised for his autobiographical novel, *Il piacere* (1889; *The Child of Pleasure*), for his mature poetry, and for his late memoirs.

Benedetto Croce's criticism. Although D'Annunzio's fame was worldwide, the function of modernizing intellectual life fell mainly to Benedetto Croce in almost 70 books and in the bimonthly review *La Critica* (1903–44). Perhaps his most influential work was literary criticism, which he expounded and continually revised in articles and books spanning nearly half a century.

Croce's beliefs implied condemnation of Fascism's ideology, but he was not seriously molested by the Fascist regime, and through the darkest days *La Critica* remained a source of encouragement to at least a restricted circle of freedom-loving intellectuals. Unfortunately, his highly systematized approach to criticism led to a certain rigidity and a refusal to recognize the merits of some obviously important writers, and this was undoubtedly one reason why after World War II his authority waned. His monumental corpus of philosophical, critical, and historical works of great scholarship, humour, and common sense remains, however, the greatest single intellectual feat in the history of modern Italian culture.

Literary trends before World War I. While Croce was starting his arduous task, literary life revolved mainly around reviews such as *Leonardo* (1903), *Hermes* (1904), *La Voce* (1908), and *Lacerba* (1913), founded and edited by relatively small groups. The two main literary trends were: *crepuscolarismo,* which favoured a colloquial style to express memories of sweet things past, as in the work of Guido Gozzano and Sergio Corazzini; and *futurismo,* loathing of traditional art and demanding complete freedom of expression, whose leader was Filippo Tommaso Marinetti, editor of *Poesia,* a fashionable cosmopolitan review. Both *crepuscolari* and *futuristi* were part of a complex European tradition of disillusionment and revolt, the former inheriting the sophisticated pessimism of French and Flemish "decadents," the latter taking part in an episode in the history of western European avant-garde developed from the French poets Stêphane Mallarmé and Guillaume Apollinaire to the Cubist, Surrealist, and Dada movements. Both shared a feeling of revulsion against D'Annunzian flamboyance and rhetoric, from which they attempted to free themselves and, paradoxically, from which they derived elements of their style (the "crepuscular" mood of D'Annunzio's *Poema paradisiaco* [1893] and most Futuristic "new theories"—identification of art with action, heroism, and speed, free use of words—being implied in his *Laus Vitae* [1903; "In Praise of Life"]).

Crepuscolarismo and *futurismo*

The "return to order." The end of World War I saw a longing for the revival of tradition, summed up in the aims of the review *La Ronda,* founded in 1919 by Vincenzo Cardarelli and others, which advocated a return to classical stylistic values. This led to an excessive cult of form in the narrow sense—as exemplified by the elegant but somewhat bloodless essays (*elzeviri*) published in Italian newspapers on page three—and obviously fitted in with the stifling of free expression under Fascism. The sterility of this period, however, should not be exaggerated. The 20 years of Fascist rule were hardly conducive to creativity, but in the dark picture there were a few glimmers of light. With 1923 came the publication of Italo Svevo's *Coscienza di Zeno* (*The Confessions of Zeno*), a gem of psychological observation and Jewish humour, which a few years later was internationally "discovered" through the mediation of James Joyce. The surreal writings of Massimo Bontempelli (*Il figlio di due madri* [1929; "The Son of Two Mothers"]) and of Dino Buzzati (*Il deserto dei Tartari* [1940; *The Tartar Steppe*]) were perhaps in part an escape from the prevailing political climate, but they stand up artistically nonetheless. Riccardo Bacchelli, with *Il diavolo a Pontelungo* (1927; *The Devil at the Long Bridge*) and *Il mulino del Po* (1938–40; *The Mill on the Po*), produced historical narrative writing of lasting quality. Aldo Palazzeschi, in *Stampe dell'Ottocento* (1932; "Nineteenth Century Engravings") and *Sorelle Materassi* (1934; *The Sisters Materassi*), reached the height of his storytelling powers. Meanwhile, the Florentine literary reviews *Solaria, Frontespizio,* and *Letteratura,* while having to tread carefully with the authorities, provided an outlet for new talent. Carlo Emilio Gadda had his first narrative work (*La Madonna dei filosofi* [1931; "The Philosophers' Madonna"]) published in *Solaria,* while the first part of his masterpiece, *La cognizione del dolore* (*Acquainted with Grief*), was serialized between 1938 and 1941 in *Letteratura.* Novelists such as Alberto Moravia, Corrado Alvaro (*Gente in Aspromonte* [1930; *Revolt in Aspromonte*]), and Carlo Bernari had to use circumspection in stating their views but were not completely silenced. Ignazio Silone, having chosen exile, could speak openly in *Fontamara* (1930). Antonio Gramsci, an unwilling "guest" of the regime, gave testimony to the triumph of spirit over oppression in *Lettere dal carcere* (1947; *Letters from Prison*).

Sterility and creativity under Fascism

Luigi Pirandello. Drama, which a few playwrights and producers were trying to extricate from old-fashioned realistic formulas and more recent superhuman theories, was increasingly dominated by Luigi Pirandello. His own experience of the "unreal," through his calamitous family life and his wife's insanity, enabled him to see the limitations of realism. From initial short-story writing, in which he explored the incoherence of personality, the lack of communication between individuals, the uncertain boundaries between sanity and insanity or reality and appearance, and the relativity of truth, he turned to drama as a better means of expressing life's absurdity and the ambiguous relationship between fact and fiction.

To multiply the fragmentation of the levels of reality, Pirandello tried to destroy conventional dramatic structures and to adopt new ones: a play within a play in *Sei personaggi in cerca d'autore* (1921; *Six Characters in Search of an Author*) and a scripted improvisation in *Questa sera si recita a soggetto* (1930; *Tonight We Improvise*). This was a way of transferring the dissociation of reality from the plane of content to that of form, thereby achieving an almost perfect unity between ideas and dramatic structure. Pirandello's plays, including perhaps his best, *Enrico IV* (1922; *Henry IV*), often contain logical arguments: several critics, including Croce, were misled into thinking that he intended to express in this way a coherent philosophy, whereas he used logic as a dramatic symbol. Pirandello was awarded the 1934 Nobel Prize for Literature.

Pirandello's play within a play

The Hermetic movement. Poetry in the Fascist period underwent a process of involution, partly influenced by French Symbolism, with its faith in the mystical power of words, and partly under the stress of changed political conditions after World War I, during which literature had declined. Many poets of the wartime generation, weary of tradition and rhetoric, had been seeking new expression: some, like the *futuristi,* had tried to work rhetoric out of their system by letting it run amok; others, such as Camillo Sbarbaro (*Pianissimo* [1914], *Trucioli* [1920; "Shavings"]), cultivated a style purified of unessential elements. Out of those efforts grew a poetry combining the acoustic potentialities of words with emotional restraint and consisting mainly of fragmentary utterances in which words were enhanced by contextual isolation and disruption of syntactic and semantic links. The resultant obscurity compensated poets for loss of influence in a society subservient to dictatorship by turning them into an elite and allowed some, notably Eugenio Montale (who won the Nobel Prize for Literature for 1975), to express their pessimism covertly. The name of this movement, *ermetismo* ("Hermeticism"), hinted at both its aristocratic ambitions and its esoteric theory and practice. Its leader, Giuseppe Ungaretti, tried to charge each word of his early poems with such intensity of meaning that concern with technical problems often overshadowed emotion, thus producing supremely stylized forms. Thus, what in the 1920s had appeared revolutionary proved later to be only another facet of the formalistic tradition. Against this background of refinement, obscurity, and unreality, only the simple and moving poems of Umberto Saba preserved an immediate appeal.

Experiments in obscurity

Social commitment and the new realism. During World War II the walls of the hermetic ivory tower began to crumble. Ungaretti's style became so articulate as to be almost unrecognizable. Salvatore Quasimodo adopted a new engagé, or committed, style, which won critical admiration, including the 1959 Nobel Prize for Literature, and others followed suit in a drift toward social realism.

Social realism

This development had been foreshadowed by some writers under Fascism. In 1929 Alberto Moravia had written a scathing indictment of middle-class moral indifference, *Gli indifferenti* (1929; *Time of Indifference*). Carlo Bernari wrote a novel about the working classes, *Tre operai* (1934; "Three Workmen"); Cesare Pavese produced *Paesi tuoi* (1941; "Your Lands"; Eng. trans., *The Harvesters*); and Elio Vittorini wrote *Conversazione in Sicilia* (1941; *Conversation in Sicily*); all definitely promised a new literary development. From these and from aspects of American literature (William Faulkner, Erskine Caldwell, John Steinbeck, John Dos Passos, and Ernest Hemingway, translated mainly by Elio Vittorini and Pavese) postwar writing took its cue. Certain English literature, the homegrown *veristi,* and the ideas of Marxism were also an influence on postwar authors, to whom in varying degrees the rather imprecise label of Neorealism was attached. It was a stimulating time in which to write, with a wealth of unused material at hand. There were the social and economic problems of the south, described by Carlo Levi in his poetic portrait of Lucania, *Cristo si è fermato a Eboli* (1945; *Christ Stopped at Eboli*), and by Rocco Scotellaro (*Contadini del sud* [1954; "Peasants of the South"]) and Francesco Jovine (*Le terre del Sacramento* [1950; "The Lands of the Sacrament"; Eng. trans., *The Estate in Abruzzi*]). Vivid pictures of the Florentine working classes were painted by Vasco Pratolini (*Il quartiere* [1945; "The District"; Eng. trans., *The Naked Streets*] and *Metello* [1955]) and of the Roman subproletariat by Pier Paolo Pasolini (*Ragazzi di vita* [1955; *The Ragazzi*] and *Una vita violenta* [1959; *A Violent Life*]). There were memories of the north's struggle against Fascist and Nazi domination from Vittorini and from Beppe Fenoglio (*I ventitrè giorni della città di Alba* [1952; "The 23 Days of the City of Alba"]). There were sad tales of lost war by Giuseppe Berto (*Il cielo è rosso* [1947; *The Sky Is Red*] and *Guerra in camicia nera* [1955; "A Blackshirt's War"]) and by Mario Rigoni Stern (*Il sergente nella neve* [1952; "The Sergeant in the Snow"; Eng. trans. in *The Lost Legions; Three Italian War Novels*]). By contrast, there were humorous recollections of provincial life under Fascism—for example, Mario Tobino's *Bandiera nera* (1950; "Black Flag") and Goffredo Parise's *Prete bello* (1954; "The Handsome Priest"; Eng. trans., *The Priest Among the Pigeons*). In contrast to the more topical appeal of these writings the great virtue of Pavese's narrative was the universality of its characters and themes. Among his finest works may be numbered *La casa in collina* (1949; *The House on the Hill*) and *La luna e i falò* (1950; *The Moon and the Bonfire*). Also of lasting relevance is Primo Levi's moving account of how human dignity survived the degradations of Auschwitz (*Se questo è un uomo* [1947; *If This Is a Man*]).

Other writings. Literary tastes gradually became less homogeneous. On the one hand, there was a rediscovery as an experimentalist of Carlo Emilio Gadda, whose best works had been written between 1938 and 1947. On the other, there was the runaway success of Giuseppe Tomasi di Lampedusa's old-fashioned historical novel *Il gattopardo* (1958; *The Leopard*), a soft-focused, flattering view of a family similar to the one described so pitilessly by Federico De Roberto in *I vicerè.* For this reason, it is easier to see Italian writing in terms of individual territory rather than general trends. Carlo Cassola's most memorable novels use the stillness of rural Tuscany as a background to the interior reality of its inhabitants, and in this his lineage can be traced to other Tuscan writers such as Romano Bilenchi (*La siccità* [1941; "The Drought"]) and Nicola Lisi (*Diario di un parroco di campagna* [1942; "Diary of a Country Priest"]) or in some respects back to Federigo Tozzi. Especially typical of Cassola's works are *Il taglio del bosco* (1953; *The Felling of the Forest*), *Un cuore arido* (1961; *An Arid Heart*), and *Un uomo solo* (1978; "A Man by Himself"). Giorgio Bassani's domain is the sadly nostalgic world of Ferrara in days gone by, with particular emphasis on its Jewish community (*Il giardino dei Finzi-Contini* [1962; *The Garden of the Finzi-Continis*]). Italo Calvino concentrated on fantastic tales (*Il visconte dimezzato* [1952; *The Cloven Viscount*], *Il barone rampante* [1957; *The Baron in the Trees*], and *Il cavaliere inesistente* [1959; *The Nonexistent Knight*]) and, later, on moralizing science fiction (*Le cosmicomiche* [1965; *Cosmicomics*] and *Ti con zero* [1968; *t zero*]). Paolo Volponi's province is the human aspect of Italy's rapid postwar industrialization (*Memoriale* [1962], *La macchina mondiale* [1965; *The Worldwide Machine*], and *Corporale* [1974]). Leonardo Sciascia's sphere is his native Sicily, whose present and past he displays with concerned and scholarly insight, with two of his better known books—in the format of thrillers—covering the sinister operations of the local Mafia (*Il giorno della civetta* [1963; *The Day of the Owl*] and *A ciascuno il suo* [1966; "To Each His Own"; Eng. trans., *A Man's Blessing*]). Giuseppe Berto, after a Neorealistic phase, plunged into the world of psychological introspection (*Il male oscuro* [1964; "The Dark Sickness"] and *La cosa buffa* [1966; "The Funny Thing"]). Natalia Ginzburg's territory is the family, whether she reminisces about her own (*Lessico famigliare* [1963; *Family Sayings*]), handles fictional characters (as in *Famiglia* [1977]), or ventures into historical biography (*La famiglia Manzoni* [1983]). Giovanni Arpino excelled at personal sympathies that cross cultural boundaries (*La suora giovane* [1959; *The Novice*] and *Il fratello italiano* [1980; "The Italian Brother"]). Fulvio Tomizza also tackled this theme in *L'amicizia* (1980; "The Friendship"). Meanwhile, Alberto Moravia and Mario Soldati defended their corners as never less than conspicuously competent writers. Moravia generally plowed a lone furrow. Of his mature writings, *Agostino* (1944), *Il conformista* (1951; *The Conformist*), and *La noia* (1960; "The Tedium"; Eng. trans., *Empty Canvas*) stand out as particular achievements. Soldati, in works such as *Le lettere da Capri* (1953; *The Capri Letters*) and *Le due città* (1964; "The Two Cities")—and in a later novel, *L'incendio* (1981; "The Fire"), which takes a quizzical look at the modern art business—showed himself to be a consistently skilled and entertaining narrator. There are many other accomplished authors who could be classified in this way, including Elsa Morante, who with *La storia* (1974; *History*) carved a unique niche for herself. Set in Rome during the years 1941–47, this combination of fact and allegory is a tour de force and one of the most remarkable narrative works to have come out of Italy since World War II.

Creative diversity

BIBLIOGRAPHY. GIULIANO PROCACCI, *History of the Italian People* (1970; 2nd Italian ed., 1968), provides an excellent general background. Histories of the literature include FRANCESCO DE SANCTIS, *History of Italian Literature,* 2 vol. (1931; reissued 1968; originally published in Italian, 1870); ROBERT ANDERSON HALL, *A Short History of Italian Literature* (1951); ERNEST HATCH WILKINS, *A History of Italian Literature,* rev. ed. (1974); JOHN H. WHITFIELD, *A Short History of Italian Literature,* 2nd ed. (1980); and ITALO DE BERNARDI and G. BARBERO, *Profilo storico della letteratura* (1983). PETER BONDANELLA and JULIA CONAWAY BONDANELLA (eds.), *Dictionary of Italian Literature* (1979), is a guide to authors, genres, schools, and periods of Italian literary history; see also the *Enciclopedia della letteratura Garzanti,* 4th ed. (1982), for cautious but sound views.

Studies that concentrate on specific periods or trends of Italian literary history include ADOLF GASPARY, *The History of Early Italian Literature to the Death of Dante* (1901; originally published in German, 1885); PIERO BOITANI (ed.), *Chaucer and the Italian Trecento* (1983); DAVID MARSH, *The Quattrocento Dialogue: Classical Tradition and Humanist Innovation* (1980); BERNARD WEINBERG, *A History of Literary Criticism in the Italian Renaissance,* 2 vol. (1961, reprinted 1974); JEFFERSON B. FLETCHER, *Literature of the Italian Renaissance* (1934, reprinted 1964); JOHN A. SYMONDS, *Renaissance in Italy,* 7 vol. (1875–86); VIOLET PAGET, *Studies of the Eighteenth Century in Italy,* new ed. (1887, reissued 1978); ANTONIO CIPPICO, *The Romantic Age in Italian Literature* (1918); LANDER MACCLINTOCK, *The Age of Pirandello* (1951, reprinted 1968); BENJAMIN CRÉMIEUX, *Panorama de la littérature italienne contemporaine* (1928), which contains still fascinating insights; SERGIO PACIFICI, *A Guide to Contemporary Italian Literature: From Futurism to Neorealism* (1962, reissued 1972), and *The Modern Italian Novel from Capuana to Tozzi* (1973); and JOHN GATT-RUTTER, *Writers and Politics in Modern Italy* (1978), a study of the relationship between politics and literature since World War II.

(G.A./S.Ra./G.P.Gi./D.M.Wh./F.Do./G.Car./Jo.M.)

Italy

With a shape that has been likened frequently to a high-heeled boot apparently about to prod its triangular subject island of Sicily, the peninsular home of the European nation of Italy (officially Italian Republic; Italian Italia, or Reppublica Italiana) juts deep into the Mediterranean Sea. Another important island, Sardinia, lies some 160 miles (260 kilometres) west of its "shin." The magnificent mountain barrier of the Alps forms a northern boundary which, historically, has hindered marauders less than might be supposed; these mountains separate Italy from France, Switzerland, Austria, and Yugoslavia and extend all the way down the Italian peninsula as a less elevated chain, the Apennines. Areas of plain, which are practically limited to the great northern oval of the Po Valley, cover a mere 23 percent of the total national area of 116,000 square miles (301,000 square kilometres); 42 percent is hilly and 35 percent mountainous, providing variations to the generally temperate climate.

The mountainous landscape of Italy has long influenced political and economic developments in the region by encouraging numerous independent states and by permitting in many regions only a meagre agriculture, providing grain sufficient only for a subsistence economy. Increased cultivation has caused deforestation. Since much of the land is mountainous, the population is dense. Since World War II, increasing numbers of Italians have abandoned the countryside for the rapidly industrializing cities, often creating severe dislocations in traditional ways of life.

The Italian economy, now ranked high in the world, blends areas as diverse as the "industrial triangle," formed by Milan, Turin, and Genoa, dating from around 1900, and the notoriously backward regions of the south and the islands, which are, however, being developed, mostly with state aid.

Agriculture, which operates often in difficult natural and economic conditions, contributes about 10 percent of the gross national product (GNP); industry, about 40 percent; and public and private services, nearly half. Sufficient wheat is grown for the population, and vegetables, fruit, wine, and oil are cultivated in suitable districts. Cattle raising, however, is less advanced; meat and dairy produce are imported.

Italian industry includes every type of production. Though mineral resources are scarce, imported raw materials since World War II have boosted siderurgy (the metallurgy of iron and steel), other metallurgy, and construction. The chemical industry also flourishes, and textiles constitute one of Italy's largest industries. Services, particularly tourism, are very important, and efforts have been made to provide comprehensive networks of *autostrade* (express highways). The heavy international exchange reflects an increasingly unfavourable balance of payments.

The peninsula has a proud tradition dating from the days of the ancient Roman Empire. From its unification in the second half of the 19th century until 1946, Italy was a monarchy. Then it became a parliamentary republic, operating under a constitution of 1948. The republic is subdivided into regions (*regioni*), provinces (*provincie*), and municipalities (*comuni*); these local bodies enjoy a certain autonomy, especially the regions, which differ widely in economic development. A similar diversity characterizes political life, which features a multiplicity of parties. The powerful Christian Democrat Party (Democrazia Cristiana, or DC) and the strong Italian Communist Party (Partito Comunista Italiano, or PCI) generally obtain strong showings in elections, and there are a number of smaller parties, representing groups of right-wing, centre, and left-wing persuasions. No party enjoys a dependable majority, and coalition governments have been notoriously unstable in the post-World War II decades.

Workers' unions have been increasingly important in national life. They are grounded in various confederations, principally the Confederazione Generale Italiana del Lavoro (General Confederation of Labour, or CGIL), controlled in effect by the Communist Party. Employers' groups and the great state bureaucracies also form important pressure groups in this often sharply polarized society.

Italy is part of the European Economic Community (EEC, or Common Market), and of the Council of Europe and belongs to many other international organizations. With its strategic geographical position on the southern flank of Europe, Italy has since World War II played a fairly important role in the North Atlantic Treaty Organization (NATO). (Li.L.)

The article is divided into the following sections:

PHYSICAL AND HUMAN GEOGRAPHY

The land

RELIEF

Italy is largely mountainous, with 35 percent of its territory occupied by ranges that are higher than 2,300 feet (702 metres), 42 percent by hills, and only 23 percent by plains. There are two mountain systems: the scenic Alps, part of which lie within the neighbouring countries of France, Switzerland, Austria, and Yugoslavia; and the Apennines, which form the spine of the entire peninsula and of the island of Sicily. A third mountain system exists in the two large islands to the west, Italian Sardinia and French Corsica.

Mountain ranges. The rugged Alps run in a broad west-to-east arc from the Colle di Cadibona (Cadibona Pass), near Savona, on the Gulf of Genoa, to the north of Trieste, at the head of the Adriatic Sea. The section properly called Alpine is the border district that includes the highest masses, made up of weathered Hercynian rocks, dating from the Carboniferous or the Permian periods (345,000,000–225,000,000 years ago). The Alps have rugged, very high peaks, reaching more than 13,000 feet in various spectacular formations, characterized as pyramidal, pinnacled, rounded, or needlelike. The valleys were heavily scoured by glaciers operating in the Quaternary Period (the last 2,500,000 years); there are still more than 1,000 glaciers left, though in a phase of retreat, more than 100 having disappeared in the last half century or so.

The three main Alpine groups

The Alpine mountain mass falls into three main groups: the Western Alps, running north to south in Italy from Aosta to the Cadibona Pass, with the Gran Paradiso (13,323 feet [4,061 metres]) and Monte Viso (12,602 feet [3,841 metres]); the Central Alps, running west to east, from the Western Alps to the Brenner Pass, leading into Austria and the upper Adige, also with high peaks, such as Monte Bianco (15,771 feet [4,807 metres]), the Matterhorn (14,692 feet [4,478 metres]), Monte Rosa (Italy's highest peak, shared with Switzerland, 15,203 feet [4,634 metres]), and Ortles (12,792 feet [3,899 metres]); and the Eastern Alps, running west to east from the Brenner to Trieste and including the Dolomites (Alpi Dolomitiche), with Monte Marmolada (10,964 feet [3,342 metres]). The Italian foothills of the Alps, which reach no higher than 8,200 feet, lie between these great ranges and the Po Valley. They are composed mainly of limestone and sedimentary rocks. A notable feature is the karst system of underground caves and streams that are especially characteristic of the Carso, the limestone plateau between the Eastern Alps and Illyria.

The Apennines are the long system of mountains and hills that run down the Italian peninsula from the Cadibona Pass to the tip of Calabria and continue in the island of Sicily. The range is about 745 miles long; it is only about 20 miles wide at either end but about 120 miles wide in the Central Apennines, east of Rome, where the Gran Sasso d'Italia group provides the highest peak (9,560 feet [2,914 metres]) and the only glacier on the peninsula, Calderone, the southernmost in Europe. The Apennines are predominantly of sandstone and limestone marl (clay) in the north; of limestone and dolomite (magnesian limestone) in the centre; and of limestone, weathered rock, and Hercynian granite in the south. On either side of the central mass are grouped two considerably lower masses, composed in general of more recent and softer rocks, such as sandstone. These are the sub-Apennines, which run in the east from Monferrato to the Golfo di Taranto and in the west from Florence southward through Tuscany and Umbria to Rome. This latter range is separated from the main Apennines by the valleys of the Arno and the Tiber rivers. At the outer flanks of the sub-Apennines two allied series of limestone and volcanic rock extend to the coast. They include, on the west, the Alpi Apuane (Apuan Alps), which are famous for their marbles; farther south, the Colline Metallifere (Ore Mountains; more than 3,400 feet), abundant in minerals; then various extinct volcanoes occupied by crater lakes, such as that of Bolsena; then cavernous mountains, such as Lepini and Circeo, and the partially or still fully active volcanic group of the Campi Flegrei and Vesuvius; and finally the limestone mountains of the peninsulas of Amalfi and Cilento. The extensions on the Adriatic coast are simpler, comprising only the small promontory of Monte Conero, the higher peninsula of Promontorio del Gargano (Gargano Plateau; 3,461 feet), and the Salentine Peninsula, in Puglia. All of these are limestone.

In Sardinia there are two mountain masses, separated by the long plain of Campidano, which runs from the Golfo dell'Asinara southeastward across the island to the Golfo di Cagliari. The group in the southwest is small and low, formed from sediments, mostly mineralized, of the early Paleozoic Era (perhaps 570,000,000 to 225,000,000 years ago). The northeastern mass reaches a height of more than 6,000 feet at Gennargentu; the underlying foundation is basically metamorphic (heat-altered) rock, and it is covered in the northeast by Paleozoic granite and partially covered in the northwest by Mesozoic limestones (65,000,000 to 225,000,000 years old) and by sandstone and Cenozoic (65,000,000 to 2,500,000 years old) clays. There are caves on the seacoast and inland where limestones predominate.

Present volcanic action had its first origins in the Pliocene Epoch and Quaternary Period (covering the last 7,000,000 years) and is represented by the Campi Flegrei, which is near Naples, and by the neighbouring islands, such as Ischia; by Vesuvius; by the Isole Eolie; and by Etna, which is on the island of Sicily. Phenomena that are related to volcanism include thermal springs in the Colli Euganei, *vulcanelli* (mud springs) at Viterbo, and emissions of gas at Pozzuoli.

Seismic activity, leading to earthquakes, is rare in the Alps and the Po Valley; it is infrequent but occasionally strong in the Alpine foothills; and it may be catastrophic in the central and southern Apennines and on Sicily. Seaquakes sometimes occur in Sicily, such as that at Messina in 1908.

The plains. Plains cover only 23 percent of the area of Italy. Some of these, such as the Po Valley and the Tavoliere di Puglia (Plain of Puglia), are ancient sea gulfs filled by alluvium. Others, such as the Tavoliere di Lecce (Plain of Lecce), in Puglia, flank the sea on rocky plateaus about 65 to 100 feet high and are formed of ancient land levelled by the sea and subsequently uplifted. Plains in the interior, such as the long Val di Chiana (Chiana Valley), are made by alluvial or other filling of ancient basins. The most extensive and important plain in Italy, that of the Po Valley, occupies more than 17,000 square miles of the 27,000 square miles of Italian plain land. It ranges in altitude from sea level up to 1,800 feet, the greater part being below 330 feet. Through it runs the Po River and all its tributaries and the Rivers Reno, Adige, Piave, and Tagliamento. The plain falls into several natural divisions. At its highest end, by the Alpine foothills, it is made up of parallel *ferretto* (red loam composed of ferrous clay) ridges, running from north to south, with areas of gravel and permeable sand between them. This section of the plain is terraced and unproductive, although the rainfall is high. Below this is the section where the rivers rise, their waters eventually providing vital irrigation both for the *marcite* (winter pastures) and for the intensive agriculture of the fertile lower plain. Other notable plains include the *maremme* of Tuscany and Lazio, reclaimed marshland with dunes at the edge of the sea; the Agro Pontino, a recently reclaimed seaward extension of the Roman countryside (*campagna*); the fertile Pianura Campana (Plain of Campania) around Vesuvius; and the rather arid Tavoliere di Puglia. In Sicily the Piana di Catania (Plain of Catania) is a good area for growing citrus fruit.

Coastal areas. The seacoasts are quite varied. Along the two Ligurian rivieras, on either side of Genoa, the coast alternates in rapid succession between high, rocky zones and level gravel. From Tuscany to Campania there are long, sandy, crescent beaches and abundant dunes, which are separated by rocky eminences. The coast of Calabria is high and rocky, though sometimes broken by short beaches. The coast of Puglia and, indeed, most of the Adriatic coast is level, although it is dominated by terraced gradients. The majestic delta of the Po, extending from Rimini to Monfalcone, is riddled with the lagoons that are familiar to visitors to Venice. The Carso, the limestone coastal region between Trieste and Istria, is rocky.

DRAINAGE AND SOILS

Rivers. Italian rivers are comparatively short; the longest, the Po, is only 400 miles long. Only three major rivers flow into the Ionian Sea, while Puglia has virtually only two rivers flowing to the Adriatic. Along the Adriatic coast a good number run parallel like the teeth of a comb down from the Apennines through Molise, Abruzzi, and the Marche regions. The rivers that flow into the Tyrrhenian Sea are longer and more complex and carry greater quantities of water. These include the Volturno, in Campania; the Roman Tiber; and the Arno, which flows through Florence and Pisa. The rivers of the Ligurian rivieras are mainly short and swift flowing; a few are important simply because cities, such as Genoa, or bathing resorts, such as Rapallo, are built on their deltas. But the prince of Italian rivers is the Po. Rising in the Monviso area, it runs across the Pianura Lombarda (Plain of Lombardy), through various important cities, such as Turin and Cremona, and is steadily enlarged by the numerous tributaries that join it, especially on its left bank. The Po debouches south of Venice, forming a large delta. In the Veneto there are also rivers that are not tributaries of the Po. One of these, the Adige, at 254 miles the second longest river in Italy, flows through Verona and debouches near Adria, south of Venice. The rivers in the south have imposing floods during winter storms, and those that run through zones of impermeable rock may become dangerous; yet during the summer many of these rivers are completely dry. The rivers of the centre and north are dry in the winter, because their headwaters are frozen, but they become full in the spring from melting snow and in the autumn from rainfall.

Lakes. There are about 1,500 lakes in Italy. The most common type is the small, elevated Alpine lake formed by Quaternary glacial excavation during the last 25,000 years. These are of major importance for hydroelectric schemes. Other lakes, such as Bolsena and Albano, in Lazio, occupy the craters of extinct volcanoes. There are also coastal lakes, such as those of Lesina and Varano, in Puglia; and lakes resulting from prehistoric faulting, such as the Lago di Alleghe, near Belluno. The best known, largest, and most important of the Italian lakes are those cut into valleys of the Alpine foothills by Quaternary glaciers. These, which are listed in order of size, are the Lago di Garda, Lago Maggiore, and the lakes of Como, Iseo, and Lugano. They have a semi-Mediterranean climate and are surrounded by groves of olive and citrus trees. Italy also has considerable areas in which, as a result of porous rock, the water systems run underground, forming subterranean streams, sinkholes, and lakes. These are often associated with caves, the most famous of which are those of Castellana, in Puglia.

The lakes

The soils. Varying climatic conditions in successive eras and differences in altitude and in types of rock have combined to produce in Italy a wide range of soils. Very common is dark-brown podzol, typical in mountains with a lot of flint, where the rainfall is heavy, as in the Alps above about 300 feet. In the Apennines, brown podzolic soils predominate, supporting forests and meadows and pastures. Brown Mediterranean soils are also characteristic of the Apennines and are suitable for agriculture. Renzinas, typically humus-carbonates, are characteristic of limestone and magnesian limestone mountain pastures and of many meadows and beech forests of the Apennines. Red earth—the famous terra rossa, derived from the residue of limestone rocks—is found not only in the extreme south (in Puglia, for instance) but also in Venetia, where it is the usual soil in vineyards, olive groves, and gardens. Sparse rocky earth, clays, dune sands, and gravel are found in the high mountains, in some volcanic zones, and in gullies in the sub-Apennines. There is also a red loam, or *ferretto,* composed of ferrous (iron) clay.

CLIMATE

Geographically, Italy lies in the temperate zone. Because of the considerable length of the peninsula, there is a variation between the climate of the north, attached to the European continent, and that of the south, surrounded by the Mediterranean. The Alps are a partial barrier against westerly and northerly winds, while both the Apennines and the great plain of northern Italy produce special climatic variations. Sardinia is subject to Atlantic and Sicily to African winds. In general, four meteorological situations dominate the Italian climate: the Mediterranean winter cyclone, with a corresponding summer anticyclone; the Alpine summer cyclone, with a consequent winter anticyclone; the Atlantic autumnal cyclone; and the eastern Siberian autumnal anticyclone. The meeting of the two last-mentioned air masses brings heavy and sometimes disastrous rains in the autumn.

The seven main climatic zones

Italy can be divided into seven main climatic zones. In the most northerly, the Alpine Zone, which has a continental mountain climate, temperatures are lower and rainfall higher in the east than in the west. The average temperature at Bardonecchia, in the west, is 45.3° F (7.4° C), and the rainfall is 26 inches (660 millimetres); at Val d'Ampezzo, in the east, the figures are 43.9° F (6.6° C) and 41.5 inches. In the Valle d'Aosta, in the west, the permanent snow line is at 10,200 feet, but in the Julian Alps it is as low as 8,350 feet. In autumn and in late winter the hot, dry wind that is known as the foehn blows from Switzerland or from Austria, and in the east the cold, dry bora blows with gusts of up to 125 miles per hour. Rain falls in the summer in the higher and more remote areas and in the spring and autumn at the periphery. Snow falls only in the winter, but the snowfall varies from about 10 to 32 feet in different years and in relation to the exact altitude or closeness of the sea. More snow falls in the foothills than in the mountains and more in the Eastern than in the Western Alps. Around the lakes the climate is milder, the average temperature in January at Milan being 34° F (1° C); at Salò, on Lago di Garda, it is 39° F (4° C).

The Po Valley has hot summers but severe winters, worse in the interior than toward the eastern coast. At Turin the winter average is 32.5° F (0.3° C) and the summer average 74° F (23° C). Rain falls mainly in the spring and autumn and increases with the altitude. There is scant snow, and that falls only on the high plain. The temperatures of places along the Adriatic coast rise steadily from north to south, partly because of the descending latitude and partly because the prevailing winds are easterly in the north but southerly in the south. The average annual mean temperature rises from 56.5° F (13.6° C) at Venice to 60° F (16° C) at Ancona and 63° F (17° C) at Bari. There is scant rain: Venice has 29.5 inches, Ancona 25.5 inches, and Bari 23.6 inches.

In the Apennines the winters vary in severity according to the altitude. Except at specific locations, there are but moderate amounts of both rain and snow, but in the cyclonic conditions of midwinter there may be sudden snowfalls in the south. The annual mean temperatures are 53.8° F (12.1° C) at Urbino, in the east, and 54.5° F (12.5° C) at Potenza, in Lucania; the annual rainfall is, respectively, 35 inches and 39.6 inches. Along the Tyrrhenian coast, on the Ligurian rivieras in the north, both temperature and rainfall are influenced by full exposure to the noonday sun; the nearness of the sea, with its prevailing southwesterly winds; and the Apennine Range that protects the area from the cold north winds. The eastern Riviera has more rain than the western: rainfall at San Remo, on the western Riviera, is 26.7 inches, but at La Spezia, on the eastern Riviera, it is 45.2 inches. Farther south, where the coastal area extends a great distance inland and is flatter, the mean temperature and annual rainfall are 58.6° F (14.8° C) and 30.3 inches at Florence and 61.9° F (16.6° C) and 31.4 inches at Naples. As a rule, the Tyrrhenian coast is warmer and more rainy than the Adriatic coast. Both Calabria and Sicily are mountainous regions that are surrounded by the Mediterranean, and they therefore have higher temperatures than the Italian mainland high regions farther north. Winter rains are scarce in the interior and heavier in the west and north of Sicily. At Reggio di Calabria the annual mean temperature is 64.7° F (18.2° C) and rainfall 22.4 inches; at Palermo, in Sicily, they are 64.4° F (18.0° C) and 38.2 inches. The sirocco, which is a hot, very humid, and depressing wind, blows frequently from Africa and the Near East. In Sardinia, conditions are more turbulent on the western side, and the island suffers from the cold mistral blowing from the northwest and also from the sirocco blowing from the southeast. At Sassari, in the northwest, the annual mean temperature is 62.6° F (17.0° C) and the rainfall 22.8 inches, while at Orosei, on the east coast, the temperature is 63.5° F (17.5° C) and the rainfall 21.2 inches.

PLANT LIFE

The native vegetation of Italy reflects the diversity in the prevailing physical environment in different parts of the country. There are at least three zones of differing vegetation, the Alps, the Po Valley, and the Mediterranean–Apennine area.

The three vegetation zones

From the foot of the Alps to their highest peaks, three bands of vegetation can be distinguished. First, around the Lombard lakes, the most common trees are the evergreen cork oak, the European olive, the cypress, and also the cherry laurel. Slightly higher, on the mountain plain, the beech is ubiquitous, giving place gradually to the deciduous larch and the Norway spruce. In the high-altitude zone, twisted shrubs, including rhododendron, green alder, and dwarf juniper, then give way to pastureland that is covered with grasses and sedges and wildflowers such as gentian, dryad, rock jasmine, campion, sea bindweed, primrose, and saxifrage. Farther up there is curved sedge, with the dwarf willow and the lovely anthophytes. On the snow line there are innumerable mosses, lichen, and flags, as well as a few varieties of hardy pollenating plants, such as saxifrage.

In the Po Valley almost nothing remains of the original forests; almost all of the vegetation has been planted or disposed by human activity. Poplars predominate where there is abundant water, but in the drier, more gravelly zones there are a few sedges. On the clayey upland plains, heather abounds, and there are forests of Scotch pine. There are the usual grasses beside the streams and in the bogs and water lilies and pondweed on the banks of the marshes. But the heavily predominant plants are the cultivated crops—wheat, corn (maize), potatoes, rice, and sugar beets. In the Apennine zone along the whole peninsula, a typical tree is the holm oak, while the area closer to the sea is characterized by the olive, oleander, carob, mastic, and the Aleppo pine. There is a notable development of pioneer sea grape on the coastal dunes. The Mediterranean foothill area is characterized by the cork oak and the Aleppo pine. Higher up, in southern Italy, there are still remaining traces of the ancient mountain forest, with truffle oak, chestnut, flowering ash, Oriental oak, white poplar, and Oriental plane. There are quite extensive beechwoods in Calabria (on La Sila and Aspromonte mountains) and Puglia and the silver fir and various kinds of pine in Abruzzi and Calabria. Where the forests have been destroyed in the strictly Mediterranean section of the Apennines, a scrub that is called macchia has grown up. On the island of Sardinia the destruction of the carob forests and on the Tavoliere di Puglia the decay of olive trees and shore vegetation have produced steppes of tough plants such as the various sorts of feather grass. Mountain meadowlands are found in Calabria and Basilicata, usually with vetch, bent grass, and the white asphodel. The Apennine pasturelands are very much like those of the Alps. The papyrus is quite common in Sicily as a freshwater plant.

ANIMAL LIFE

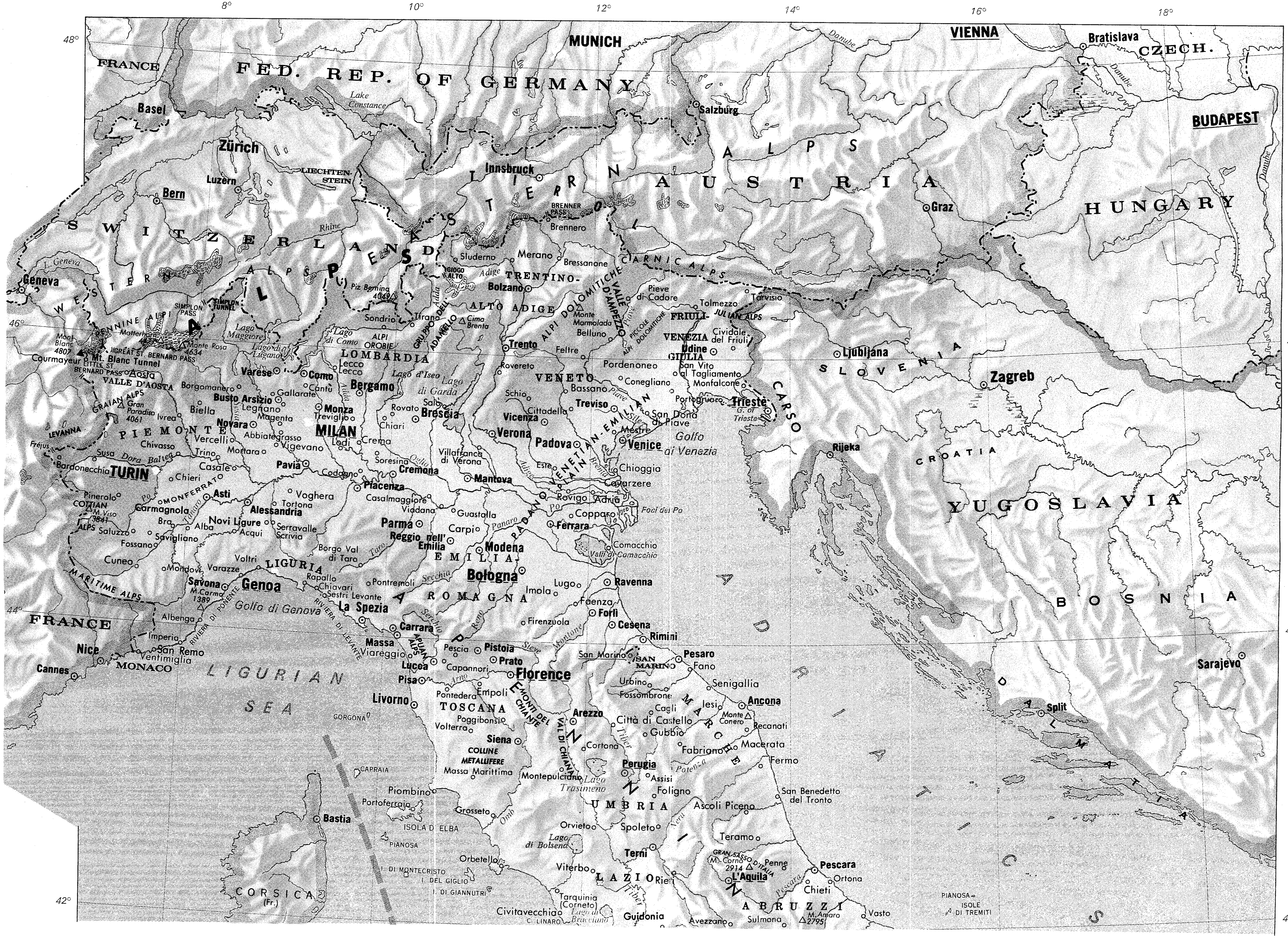
The extent of animal life in Italy has been much reduced by the long presence of human beings. In the Alps there are quite a number of animals, such as marmots, that hibernate and others that change their protective colouring according to the season, such as the ermine, the mountain partridge, and the Alpine rabbit. Larger mammals include the ibex, which is protected on the Gran Paradiso, the chamois in the Central Alps, and the roe in the Eastern Alps. The lynx, the stoat, and the brown bear (protected in Adamello and Brenta) are now rare. Alpine birds include the black grouse, the golden eagle, and, more rarely, the capercaillie, or wood grouse. Among the reptiles are vipers and among the amphibians the Alpine salamander and Alpine newt. Species that are found in the Alps also exist in other high mountain regions, where there are, however, more foxes and wolves. In Abruzzi the brown bear may be found and on the island of Sardinia the fallow deer, the mouflon sheep, and the wild boar. Among the freshwater fish are the brown trout, the sturgeon, and the eel. Among sea fish, besides common species such as the red mullet and the dentex, there are, especially in southern waters, the white man-eater shark, the bluefin tuna, and the swordfish. Among invertebrates, there is an abundance of red coral and commercial sponge on the rocks of the warm southern seas. In caves the greater horseshoe bat is found.

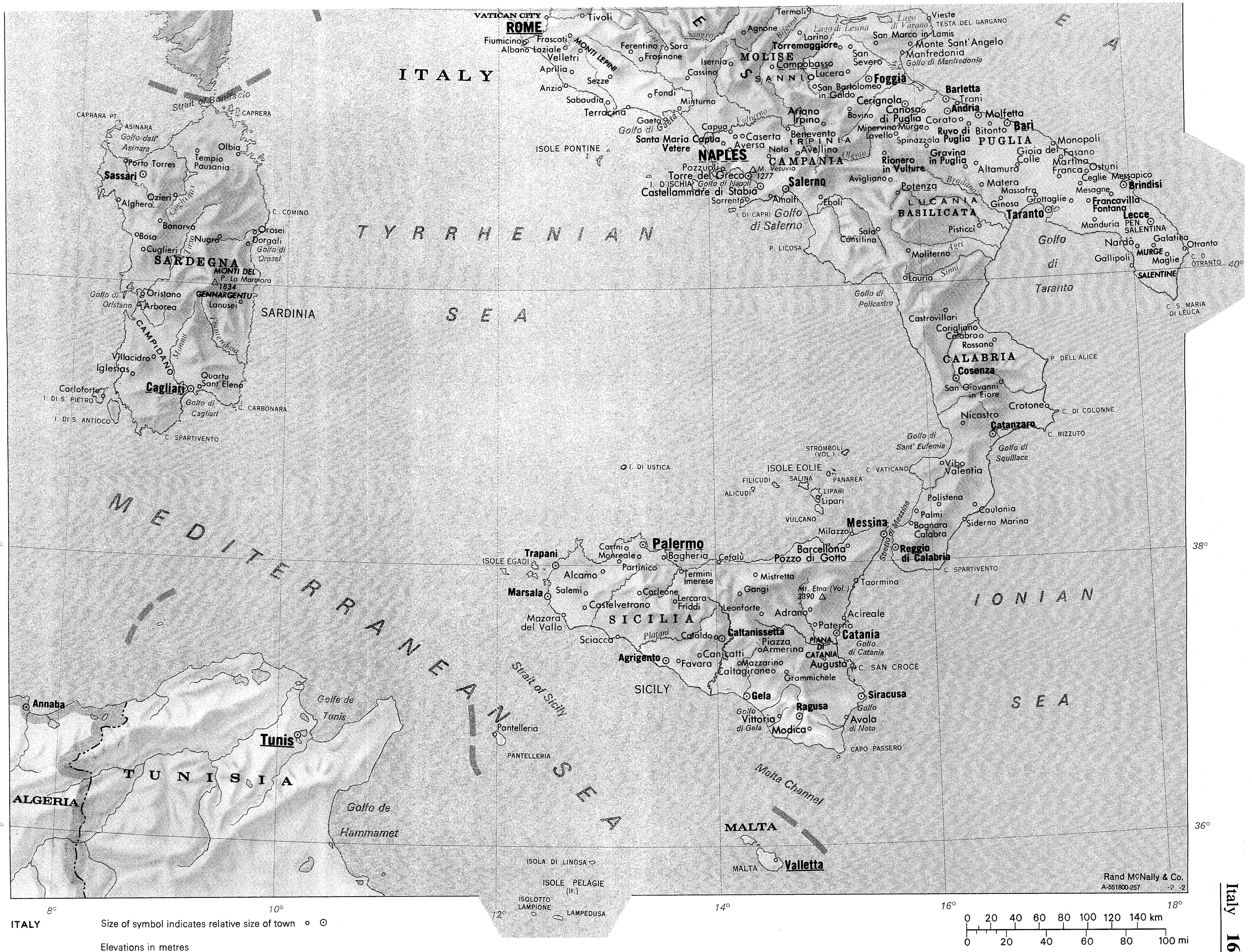
TRADITIONAL REGIONS

Italy is divided into 20 administrative regions, which correspond generally with historical traditional regions, though not always with exactly the same boundaries. A better known and more general way of dividing Italy is into four parts: the north, the centre, the south, and the islands.

Problems of regional division

The north includes such traditional regions as Piedmont, which is marked with some French influence and the seat of united Italy's royal dynasty, with Liguria extending southward around the Gulf of Genoa; the Milanese, which has been long celebrated for its productive agriculture and vigorously independent city communes and is now for its industrial output; and the Veneto, once the territory of the far-flung Venetian Empire and reaching from Brescia to Trieste in its greatest extent. The centre includes Emilia, with its prosperous farms; the Marche, on the Adriatic side; Tuscany and Umbria, which are treasured vestiges of Etruscan civilization and the great Renaissance traditions of art and culture; Latium (Lazio) and the Campagna, whose beautiful hills encircle the eternal city of Rome; and the Abruzzi and the Molise, regions of the highest central Apennines, which used to support a wild and remote people. The south (Mezzogiorno) includes Naples and its surrounding fertile Campania; the poorer regions of Puglia, with its great plain crossed by oleander-bordered roads leading to the low Murge Salentine (Murge Hills) and the heel of Italy, and Basilicata, Lucania, and Calabria, which was once brigand haunted. Finally, in the islands of Sicily and Sardinia are people who take pride in holding themselves apart from the inhabitants of mainland Italy.


8°
10°
12°
14°
16°
18°
48°
46°
44°
42°
FRANCE
FED. REP. OF GERMANY
MUNICH
VIENNA
Danube
Bratislava
CZECH.
BUDAPEST
Basel
Lake Constance
Salzburg
Zürich
Luzern
LIECHTEN-STEIN
Innsbruck
ALPS
AUSTRIA
HUNGARY
Graz
Bern
SWITZERLAND
Rhine
EASTERN ALPS
BRENNER PASS
Brennero
L. Geneva
Geneva
WESTERN ALPS
Sluderno
Merano
Bressanone
CARNIC ALPS
GIOGO ALTO
Adige
TRENTINO-
Bolzano
ALTO ADIGE
Piz Bernina 4049
ALPI DOLOMITICHE
VALLE D'AMPEZZO
Pieve di Cadore
Tolmezzo
Tarvisio
SIMPLON PASS
SIMPLON TUNNEL
PENNINE ALPS
Matterhorn 4478
Monte Rosa 4634
Lago Maggiore
Sondrio
Tirano
GRUPPO DELL' ADAMELLO
Cima Brenta
Monte Marmolada
PICCOLE ALPI DOLOMITICHE
FRIULI-
JULIAN ALPS
Mont Blanc 4807
GREAT ST BERNARD PASS
Mt. Blanc Tunnel
Courmayeur
LITTLE ST BERNARD PASS
Aosta
VALLE D'AOSTA
Lago di Lugano
Lago di Como
ALPI OROBIE
Belluno
VENEZIA GIULIA
Cividale del Friuli
Udine
Ljubljana
SLOVENIA
LOMBARDIA
Trento
Feltre
Lecco
Varese
Como
Lago d'Iseo
Rovereto
VENETO
Pordenone
San Vito al Tagliamento
Monfalcone
Zagreb
Borgomanero
Busto Arsizio
Gallarate
Cantù
Bergamo
Lago di Garda
Schio
Bassano
Conegliano
Portogruaro
Trieste
G. of Trieste
CARSO
GRAIAN ALPS
Gran Paradiso 4061
Ivrea
Biella
Legnano
Monza
Treviglio
Rovato
Salò
Brescia
Cittadella
Treviso
Piave
Sile
San Donà di Piave
LEVANNA
PIEMONTE
Novara
Magenta
Chiari
Vicenza
Verona
Padova
Mestre
Venice
Golfo di Venezia
Fréjus
Chivasso
Vercelli
Abbiategrasso
MILAN
Lodi
Crema
VENETIAN-EMILIAN PLAIN
Susa
Dora Baltea
Trino
Mortara
Vigevano
Soresina
Villafranca di Verona
Este
Chioggia
Rijeka
CROATIA
Bardonecchia
TURIN
Casale
Pavia
Codogno
Cremona
Oglio
Mantova
Brenta
Cavarzere
Chieri
MONFERRATO
Asti
Voghera
Piacenza
Pinerolo
COTTIAN ALPS
Po
Tortona
Casalmaggiore
Viadana
Guastalla
Rovigo
Adria
YUGOSLAVIA
M. Viso 3841
Carmagnola
Alessandria
Novi Ligure
Tanaro
Bra
Alba
Serravalle Scrivia
Acqui
Parma
Carpi
Panaro
PADAN
Copparo
Foci del Po
Ferrara
Saluzzo
Savigliano
Fossano
Reggio nell' Emilia
Modena
Taro
Borgo Val di Taro
EMILIA
Comacchio
Valli di Comacchio
Cuneo
Mondovì
Voltri
Varazze
LIGURIA
Rapallo
Chiavari
Sestri Levante
Pontremoli
Secchia
Bologna
Lugo
Ravenna
MARITIME ALPS
Savona
M. Carma 1389
Genoa
ROMAGNA
Imola
Faenza
Forlì
BOSNIA
Albenga
Golfo di Genova
RIVIERA DI PONENTE
RIVIERA DI LEVANTE
La Spezia
Firenzuola
Cesena
FRANCE
Imperia
San Remo
Ventimiglia
Nice
MONACO
Cannes
Carrara
Massa
Viareggio
APUAN ALPS
Serchio
Reno
Pescia
Pistoia
Sieve
Montone
Rimini
San Marino
SAN MARINO
Pesaro
Fano
Sarajevo
Lucca
Capannori
Prato
Florence
Pisa
Arno
Empoli
Urbino
Senigallia
DALMATIA
LIGURIAN SEA
Livorno
Pontedera
TOSCANA
MONTI DEL CHIANTI
APENNINES
Fossombrone
Cagli
Iesi
Ancona
Split
Monte Conero
Recanati
GORGONA
Poggibonsi
Volterra
Arezzo
Città di Castello
Gubbio
MARCHE
Siena
COLLINE METALLIFERE
VAL DI CHIANA
Tiber
Cortona
Fabriano
Macerata
Fermo
ADRIATIC SEA
CAPRAIA
Massa Marittima
Montepulciano
Lago Trasimeno
Perugia
Assisi
Potenza
Piombino
Portoferraio
Foligno
San Benedetto del Tronto
Bastia
Grosseto
Ombrone
UMBRIA
Ascoli Piceno
ISOLA D'ELBA
PIANOSA
Orvieto
Spoleto
Nera
Lago di Bolsena
Teramo
Orbetello
Terni
GRAN SASSO D'ITALIA
M. Corno 2914
Penne
Pescara
I. DI MONTECRISTO
I. DEL GIGLIO
I. DI GIANNUTRI
Viterbo
LAZIO
Rieti
L'Aquila
Chieti
Ortona
CORSICA (Fr.)
Tarquinia (Corneto)
Lago di Bracciano
Civitavecchia
C. LINARO
Guidonia
Avezzano
Sulmona
ABRUZZI
M. Amaro 2795
Vasto
PIANOSA
ISOLE DI TREMITI



ITALY

Size of symbol indicates relative size of town ○ ⊙

Elevations in metres

MAP INDEX

Political subdivisions

Abruzzi	42·20n	13·45e
Basilicata	40·30n	16·10e
Calabria	39·00n	16·30e
Campania	41·00n	14·30e
Emilia Romagna	44·35n	11·00e
Friuli-Venezia-Giulia	46·00n	13·00e
Lazio	42·00n	12·30e
Liguria	44·30n	8·50e
Lombardia	45·40n	9·45e
Marche	43·30n	13·15e
Molise	41·35n	14·30e
Piemonte	45·00n	8·00e
Puglia	41·15n	16·15e
Sardegna	40·00n	9·00e
Sicilia	37·30n	14·00e
Toscana	43·25n	11·00e
Trentino-Alto-Adige	46·08n	11·07e
Umbria	43·00n	12·30e
Valle d'Aosta	45·44n	7·20e
Veneto	45·30n	12·00e

Cities and towns

Abbiategrasso	45·24n	8·54e
Acireale	37·37n	15·10e
Acqui	44·41n	8·28e
Adrano	37·40n	14·50e
Adria	45·03n	12·03e
Agnone	41·48n	14·22e
Agrigento	37·18n	13·35e
Alba	44·42n	8·02e
Albano Laziale	41·44n	12·39e
Albenga	44·03n	8·13e
Alcamo	37·59n	12·58e
Alessandria	44·54n	8·37e
Alghero	40·34n	8·19e
Altamura	40·50n	16·33e
Amalfi	40·38n	14·36e
Ancona	43·38n	13·30e
Andria	41·13n	16·18e
Anzio	41·27n	12·37e
Aosta	45·44n	7·20e
Aprilia	41·36n	12·39e
Arborea	39·47n	8·34e
Arezzo	43·25n	11·53e
Ariano Irpino	41·09n	15·05e
Ascoli Piceno	42·51n	13·34e
Assisi	43·04n	12·37e
Asti	44·54n	8·12e
Augusta	37·13n	15·13e
Avellino	40·54n	14·47e
Aversa	40·58n	14·12e
Avezzano	42·02n	13·25e
Avigliano	40·44n	15·44e
Avola	36·54n	15·09e
Bagheria	38·05n	13·30e
Bagnara Calabra	38·18n	15·49e
Barcellona Pozzo di Gotto	38·09n	15·13e
Bardonecchia	45·05n	6·42e
Bari	41·07n	16·52e
Barletta	41·19n	16·17e
Bassano	45·46n	11·44e
Belluno	46·09n	12·13e
Benevento	41·08n	14·45e
Bergamo	45·41n	9·43e
Biella	45·34n	8·03e
Bitonto	41·06n	16·42e
Bologna	44·29n	11·20e
Bolzano (Bozen)	46·31n	11·22e
Bonorva	40·25n	8·45e
Borgomanero	45·42n	8·28e
Borgo Val di Taro	44·29n	9·46e
Bosa	40·18n	8·30e
Bovino	41·15n	15·20e
Bozen, see Bolzano		
Bra	44·42n	7·51e
Brennero	47·00n	11·30e
Brescia	45·33n	13·15e
Bressanone (Brixen)	46·43n	11·39e
Brindisi	40·38n	17·56e
Busto Arsizio	45·37n	8·51e
Cagli	43·33n	12·39e
Cagliari	39·13n	9·06e
Caltagirone	37·14n	14·31e
Caltanissetta	37·29n	14·04e
Campobasso	41·34n	14·39e
Canicattí	37·21n	13·51e
Canosa di Puglia	41·13n	16·04e
Cantù	45·44n	9·08e
Capannori	43·50n	10·34e
Capua	41·06n	14·12e
Carini	38·08n	13·11e
Carloforte	39·08n	8·18e
Carmagnola	44·51n	7·43e
Carpi	44·47n	10·53e
Carrara	44·05n	10·06e
Casale	45·08n	8·27e
Casalmaggiore	44·59n	10·26e
Caserta	41·04n	14·20e
Cassino	41·30n	13·49e
Castellammare di Stabia	40·42n	14·29e
Castelvetrano	37·41n	12·47e
Castrovillari	39·49n	16·13e
Catania	37·30n	15·06e
Catanzaro	38·54n	16·36e
Caulonia	38·23n	16·25e
Cavarzere	45·08n	12·05e
Cefalù	38·02n	14·01e
Ceglie Messapico	40·39n	17·31e
Cerignola	41·16n	15·54e
Cesena	44·08n	12·15e
Chiari	45·32n	9·56e
Chiavari	44·19n	9·19e
Chieri	45·01n	7·49e
Chieti	42·21n	14·10e
Chioggia	45·13n	12·17e
Chivasso	45·11n	7·53e
Cittadella	45·39n	11·47e
Città di Castello	43·27n	12·14e
Cividale del Friuli	46·06n	13·25e
Civitavecchia	42·06n	11·48e
Codogno	45·09n	9·42e
Comacchio	44·42n	12·11e
Como	45·47n	9·05e
Conegliano	45·53n	12·18e
Copparo	44·54n	11·49e
Corato	41·09n	16·25e
Corigliano Calabro	39·36n	16·31e
Corleone	37·49n	13·18e
Corneto, see Tarquinia		
Cortona	43·16n	11·59e
Cosenza	39·17n	16·15e
Courmayeur	45·47n	6·58e
Crema	45·22n	9·41e
Cremona	45·07n	10·02e
Crotone	39·05n	17·07e
Cuglieri	40·11n	8·34e
Cuneo	44·23n	7·32e
Dorgali	40·17n	9·35e
Eboli	40·37n	15·04e
Empoli	43·43n	10·57e
Este	45·14n	11·39e
Fabriano	43·20n	12·54e
Faenza	44·17n	11·53e
Fano	43·50n	13·01e
Fasano	40·50n	17·22e
Favara	37·19n	13·40e
Feltre	46·01n	11·54e
Ferentino	41·42n	13·15e
Fermo	43·09n	13·43e
Ferrara	44·50n	11·35e
Firenzuola	44·07n	11·23e
Fiumicino	41·46n	12·14e
Florence (Firenze)	43·46n	11·15e
Foggia	41·27n	15·34e
Foligno	42·57n	12·42e
Fondi	41·21n	13·25e
Forlì	44·13n	12·03e
Fossano	44·33n	7·43e
Fossombrone	43·41n	12·48e
Francavilla Fontana	40·31n	17·35e
Frascati	41·48n	12·41e
Frosinone	41·38n	13·19e
Gaeta	41·12n	13·35e
Galatina	40·10n	18·10e
Gallarate	45·40n	8·47e
Gallipoli	40·03n	17·58e
Gangi	37·49n	14·13e
Gela	37·03n	14·15e
Genoa (Genova)	44·25n	8·57e
Ginosa	40·34n	16·46e
Grammichele	37·13n	14·38e
Gravina in Puglia	40·49n	16·25e
Grosseto	42·46n	11·08e
Grottaglie	40·32n	17·26e
Guastalla	44·55n	10·39e
Gubbio	43·21n	12·35e
Guidonia	42·01n	12·45e
Iesi	43·31n	13·14e
Iglesias	39·19n	8·32e
Imola	44·21n	11·42e
Imperia	43·53n	8·03e
Isernia	41·36n	14·14e
Ivrea	45·28n	7·52e
Lanusei	39·52n	9·34e
L'Aquila	42·22n	13·22e
Larino	41·48n	14·54e
La Spezia	44·07n	9·50e
Lauria	40·02n	15·50e
Lavello	41·03n	15·48e
Lecce	40·23n	18·11e
Lecco	45·51n	9·23e
Leghorn, see Livorno		
Legnano	45·36n	8·54e
Leonforte	37·39n	14·24e
Lercara Friddi	37·45n	13·36e
Lipari	38·28n	14·57e
Livorno (Leghorn)	43·33n	10·19e
Lodi	45·19n	9·30e
Lucca	43·50n	10·29e
Lucera	41·30n	15·20e
Lugo	44·25n	11·54e
Macerata	43·18n	13·27e
Magenta	45·28n	8·53e
Maglie	40·07n	18·19e
Manduria	40·24n	17·38e
Manfredonia	41·38n	15·55e
Mantova	45·09n	10·48e
Marsala	37·48n	12·26e
Martina Franca	40·42n	17·21e
Massa	44·01n	10·09e
Massafra	40·35n	17·07e
Massa Marittima	43·03n	10·53e
Matera	40·40n	16·37e
Mazara del Vallo	37·39n	12·36e
Mazzarino	37·18n	14·13e
Merano (Meran)	46·40n	11·09e
Mesagne	40·33n	17·49e
Messina	38·11n	15·33e
Mestre	45·29n	12·15e
Milan (Milano)	45·28n	9·12e
Milazzo	38·14n	15·15e
Minervino Murge	41·05n	16·05e
Minturno	41·15n	13·45e
Mistretta	37·56n	14·22e
Modena	44·40n	10·55e
Modica	36·51n	14·47e
Molfetta	41·12n	16·36e
Moliterno	40·14n	15·52e
Mondovì	44·23n	7·49e
Monfalcone	45·49n	13·32e
Monopoli	40·57n	17·19e
Monreale	38·05n	13·17e
Montepulciano	43·05n	11·47e
Monte Sant' Angelo	41·42n	15·57e
Monza	45·35n	9·16e
Mortara	45·15n	8·44e
Naples (Napoli)	40·51n	14·17e
Nardò	40·11n	18·02e
Nicastro	38·59n	16·20e
Nola	40·55n	14·33e
Novara	45·28n	8·38e
Novi Ligure	44·46n	8·47e
Nuoro	40·19n	9·20e
Olbia	40·55n	9·29e
Orbetello	42·27n	11·13e
Oristano	39·54n	8·35e
Orosei	40·23n	9·42e
Ortona	42·21n	14·24e
Orvieto	42·43n	12·07e
Ostuni	40·44n	17·35e
Otranto	40·09n	18·30e
Ozieri	40·35n	9·00e
Padova (Padua)	45·25n	11·53e
Palermo	38·07n	13·21e
Palmi	38·21n	15·51e
Pantelleria	36·49n	11·57e
Parma	44·48n	10·20e
Pàrtinico	38·03n	13·07e
Paternò	37·34n	14·54e
Pavia	45·10n	9·10e
Penne	42·27n	13·55e
Perugia	43·08n	12·22e
Pesaro	43·54n	12·55e
Pescara	42·28n	14·13e
Pescia	43·54n	10·41e
Piacenza	45·01n	9·40e
Piazza Armerina	37·23n	14·22e
Pieve di Cadore	46·26n	12·22e
Pinerolo	44·53n	7·21e
Piombino	42·55n	10·32e
Pisa	43·43n	10·23e
Pisticci	40·23n	16·34e
Pistoia	43·55n	10·54e
Poggibonsi	43·28n	11·09e
Polistena	38·25n	16·05e
Pontedera	43·40n	10·38e
Pontremoli	44·22n	9·53e
Pordenone	45·57n	12·39e
Portoferraio	42·49n	10·19e
Portogruaro	45·47n	12·50e
Porto Torres	40·50n	8·23e
Potenza	40·38n	15·49e
Pozzuoli	40·49n	14·07e
Prato	43·53n	11·06e
Quartu Sant' Elena	39·14n	9·11e
Ragusa	36·55n	14·44e
Rapallo	44·21n	9·14e
Ravenna	44·25n	12·12e
Recanati	43·24n	13·32e
Reggio di Calabria	38·07n	15·39e
Reggio nell' Emilia	44·43n	10·36e
Rieti	42·24n	12·51e
Rimini	44·04n	12·34e
Rionero in Vulture	40·55n	15·40e
Rome (Roma)	41·54n	12·29e
Rossano	39·35n	16·39e
Rovato	45·34n	10·00e
Rovereto	45·53n	11·02e
Rovigo	45·04n	11·47e
Ruvo di Puglia	41·07n	16·29e
Sabaudia	41·18n	13·01e
Sala Consilina	40·24n	15·36e
Salemi	37·49n	12·49e
Salerno	40·41n	14·47e
Salò	45·36n	10·31e
Saluzzo	44·39n	7·29e
San Bartolomeo in Galdo	41·24n	15·01e
San Benedetto del Tronto	42·57n	13·53e
San Donà di Piave	45·38n	12·34e
San Giovanni in Fiore	39·16n	16·42e
San Marco in Lamis	41·43n	15·38e
San Marino	43·55n	12·28e
San Remo	43·49n	7·46e
San Severo	41·41n	15·23e
Santa Maria Capua Vetere	41·05n	14·15e
San Vito al Tagliamento	45·54n	12·52e
Sassari	40·44n	8·33e
Savigliano	44·38n	7·40e
Savona	44·17n	8·30e
Schio	45·43n	11·21e
Schluderns, see Sluderno		
Sciacca	37·30n	13·06e
Senigallia	43·43n	13·13e
Serravalle Scrivia	44·44n	8·51e
Sestri Levante	44·16n	9·24e
Sezze	41·30n	13·03e
Siderno Marina	38·16n	16·18e
Siena	43·19n	11·21e
Siracusa (Syracuse)	37·04n	15·17e
Sluderno (Schluderns)	46·40n	10·35e
Sondrio	46·10n	9·52e
Sora	41·43n	13·37e
Soresina	45·17n	9·51e
Sorrento	40·37n	14·22e
Spinazzola	40·58n	16·06e
Spoleto	42·44n	12·44e
Sulmona	42·03n	13·55e
Susa	45·08n	7·03e
Syracuse, see Siracusa		
Taormina	37·52n	15·17e
Taranto	40·28n	17·15e
Tarquinia, (Corneto)	42·15n	11·45e
Tarvisio	46·30n	13·35e
Tempio Pausania	40·54n	9·07e
Teramo	42·39n	13·42e
Termini Imerese	37·59n	13·42e
Termoli	42·00n	15·00e
Terni	42·34n	12·37e
Terracina	41·17n	13·15e
Tirano	46·13n	10·10e
Tivoli	41·58n	12·48e
Tolmezzo	46·24n	13·01e
Torre del Greco	40·47n	14·22e
Torremaggiore	41·41n	15·17e
Tortona	44·54n	8·52e
Trani	41·17n	16·26e
Trapani	38·01n	12·31e
Trento	46·04n	11·08e
Treviglio	45·31n	9·35e
Treviso	45·40n	12·15e
Trieste	45·40n	13·46e
Trino	45·12n	8·18e
Turin (Torino)	45·03n	7·40e
Udine	46·03n	13·14e
Urbino	43·43n	12·38e
Varazze	44·22n	8·34e
Varese	45·48n	8·48e
Vasto	42·07n	14·42e
Vatican City	41·54n	12·27e
Velletri	41·41n	12·47e
Venice (Venezia)	45·27n	12·21e
Ventimiglia	43·47n	7·36e
Vercelli	45·19n	8·25e
Verona	45·27n	11·00e
Viadana	44·56n	10·31e
Viareggio	43·52n	10·14e
Vibo Valentia	38·40n	16·06e
Vicenza	45·33n	11·33e
Vieste	41·53n	16·10e
Vigevano	45·19n	8·51e
Villacidro	39·27n	8·44e
Villafranca di Verona	45·21n	10·50e
Viterbo	42·25n	12·06e
Vittoria	36·57n	14·32e
Voghera	44·59n	9·01e
Volterra	43·24n	10·51e
Voltri	44·26n	8·45e

Physical features and points of interest

Adamello, Gruppo dell', *mountains*	46·10n	10·35e
Adda, *river*	45·08n	9·53e

MAP INDEX (continued)

Adige (Etsch), *river*	45·10n	12·20e
Adriatic Sea	42·30n	16·00e
Agri, *river*	40·13n	16·44e
Alice, Punta dell', *point*	39·24n	17·10e
Alicudi, *island*	38·33n	14·21e
Alps, *mountains*	46·30n	10·00e
Alto, Giogo (Germen Hochjoch), *pass*	46·47n	10·47e
Amaro, Monte, *mountain*	42·05n	14·05e
Ampezzo, Valle d', *valley*	46·30n	12·10e
Apennines (Appennino), *mountains*	43·00n	13·00e
Apuan Alps, *mountains*	44·09n	10·15e
Arno, *river*	43·41n	10·17e
Asinara, *island*	41·05n	8·18e
Asinara, Golfo dell', *gulf*	41·00n	8·30e
Bernina, Piz, *peak*	46·21n	9·51e
Biferno, *river*	41·59n	15·02e
Blanc, Mont, *mountain*	45·50n	6·52e
Bolsena, Lago di, *lake*	42·36n	11·56e
Bonifacio, Strait of	41·20n	9·15e
Bracciano, Lago di, *lake*	42·07n	12·14e
Bradano, *river*	40·23n	16·51e
Brenner Pass	47·00n	11·30e
Brenta, *river*	45·11n	12·18e
Brenta, Cima, *mountain*	46·11n	10·54e
Cagliari, Golfo di, *gulf*	39·05n	9·10e
Campidano, *plain*	39·30n	8·47e
Capraia, *island*	43·02n	9·49e
Caprara Punta, *point*	41·07n	8·19e
Caprera, *island*	42·08n	15·31e
Capri, Isola di, *island*	40·33n	14·13e
Carbonara, Capo, *cape*	39·06n	9·31e
Carmo, Monte, *mountain*	44·11n	8·11e
Carnic Alps, *mountains*	46·40n	13·00e
Catania, Golfo di, *gulf*	37·25n	15·15e
Catania, Piana di, *plain*	37·25n	14·51e
Cervino, Monte, see Matterhorn		
Chiana, Val di, *valley*	43·15n	11·50e
Chiante, Monti del, *mountains*	43·32n	11·25e
Coghinas, *river*	40·56n	8·48e
Colline Metallifere (Ore Mountains)	43·12n	11·00e
Colonne, Capo di, *cape*	39·02n	17·11e
Comacchio, Valli di, *lagoon*	44·38n	12·06e
Comino, Capo, *cape*	40·31n	9·50e
Como, Lago di, *lake*	46·00n	9·20e
Conero, Monte, *mountain*	43·33n	13·36e
Corno, Monte, *mountain*	42·28n	13·34e
Cottian Alps, *mountains*	44·45n	7·00e
Dolomitiche, Alpi (Dolomite Alps), *mountains*	46·25n	11·50e
Dora Baltea, *river*	45·11n	8·05e
Egadi, Isole, *islands*	37·56n	12·16e
Elba, Isola d', *island*	42·46n	10·17e
Eolie, Isole, *islands*	38·30n	15·00e
Etna, Monte, *volcano*	37·46n	15·00e
Etsch, see Adige		
Filicudi, *island*	38·35n	14·34e
Flumendosa, *river*	39·26n	9·38e
Fréjus, *pass*	45·08n	6·41e
Gaeta, Golfo di, *gulf*	41·06n	13·30e
Garda, Lago di, *lake*	45·40n	10·41e
Gargano, Testa del, *point*	41·49n	16·12e
Gela, Golfo di, *gulf*	37·00n	14·10e
Gennargentu, Monti del, *mountains*	40·01n	9·19e
Genova, Golfo di (Gulf of Genoa)	44·10n	8·55e
Giannutri, Isola di, *island*	42·15n	11·06e
Giglio, Isola del, *island*	42·21n	10·54e
Giulie, Alpi, see Julian Alps		
Gorgona, *island*	43·26n	9·54e
Graian Alps, *mountains*	45·30n	7·10e
Gran Paradiso, *mountain*	45·32n	7·16e
Gran Sasso d'Italia, *mountains*	42·27n	13·42e
Great St. Bernard Pass	45·50n	7·10e
Ionian Sea (Ionio, Mar)	37·30n	17·00e
Irpinia, *historic region*	41·00n	15·00e
Ischia, Isola d', *island*	40·43n	13·54e
Iseo, Lago d', *lake*	45·43n	10·04e
Italian Riviera, see Levante, Riviera di		
Italian Riviera, see Ponente, Riviera di		
Julian Alps (Alpi Giulie), *mountains*	46·20n	13·45e
Lampedusa, *island*	35·31n	12·35e
Lampione, Isolotto, *islet*	35·33n	12·19e
Lepini, Monti, *mountains*	41·35n	13·00e
Lesina, Lago di, *lagoon*	41·53n	15·26e
Levanna, *mountains*	45·24n	7·12e
Levante, Riviera di (Italian Riviera), *coast*	44·15n	9·30e
Licosa, Punta, *point*	40·15n	14·54e
Ligurian Sea (Ligurienne, Mar)	43·30n	9·00e
Linaro, Capo, *cape*	42·02n	11·50e
Linosa, Isola di, *island*	35·52n	12·52e
Lipari, *island*	38·30n	14·57e
Liri, *river*	41·25n	13·52e
Little Saint Bernard Pass	45·41n	6·53e
Lucania, *historic region*	40·30n	16·00e
Lugano, Lago di, *lake*	45·58n	9·00e
Maggiore, Lago, *lake*	46·00n	8·40e
Malta Channel	36·20n	15·00e
Manfredonia, Golfo di, *gulf*	41·35n	16·05e
Mannu, *river*	40·50n	8·23e
Maritime Alps, *mountains*	44·15n	7·10e
Marmolada, Monte, *mountain*	46·26n	11·51e
Marmora, Punta la, *mountain*	39·59n	9·19e
Matterhorn (Monte Cervino) *mountain*	45·59n	7·43e
Mediterranean Sea	37·30n	11·00e
Messina, Stretto di, *strait*	38·15n	15·35e
Monferrato, *historic region*	44·55n	8·05e
Mont Blanc Tunnel	45·50n	6·53e
Montecristo, Isola di, *island*	42·20n	10·19e
Montone, *river*	44·24n	12·14e
Murge Salentine, *hills*	40·02n	18·13e
Napoli, Golfo di, *gulf*	40·43n	14·10e
Nera, *river*	42·26n	12·24e
Noto, Golfo di, *gulf*	36·50n	15·15e
Ofanto, *river*	41·22n	16·13e
Oglio, *river*	45·02n	10·39e
Ombrone, *river*	42·39n	11·00e
Ore Mountains, see Collina Metallifere		
Oristano, Golfo di, *gulf*	39·50n	8·29e
Orobie, Alpi, *mountains*	46·00n	10·00e
Orosei, Golfo di, *gulf*	40·10n	9·50e
Otranto, Capo d', *cape*	40·00n	19·00e
Padano-Venetian-Emilian Plain	45·00n	12·00e
Panarea, *island*	38·38n	15·05e
Panaro, *river*	44·55n	11·25e
Pantelleria, *island*	36·47n	12·00e
Passero, Capo, *cape*	36·40n	15·09e
Pelagie, Isole, *islands*	35·40n	12·40e
Pescara, *river*	42·28n	14·13e
Pianosa, *island*	42·13n	15·45e
Pianosa, *island*	42·35n	10·04e
Piave, *river*	45·32n	12·44e
Piccole Alpi Dolomitiche, *mountains*	46·10n	12·30e
Platani, *river*	37·23n	13·16e
Po, *river*	44·57n	12·04e
Po, Foci del, *river mouth*	44·52n	12·30e
Policastro, Golfo di, *gulf*	40·00n	15·30e
Ponente, Riviera di (Italian Riviera), *coast*	44·10n	8·20e
Pontine, Isole, *islands*	40·54n	12·57e
Potenza, *river*	43·25n	13·40e
Reno, *river*	44·37n	12·17e
Riviera, see Levante, Riviera di		
Riviera, see Ponente, Riviera di		
Rizzuto, Capo, *cape*	38·54n	17·06e
Rosa, Monte, *mountain*	45·55n	7·53e
Salentina, Penisola, *peninsula*	40·30n	18·00e
Salerno, Golfo di, *gulf*	40·32n	14·42e
Salina, *island*	38·34n	14·51e
San Antioco, Isola di, *island*	39·00n	8·25e
San Croce, Capo, *cape*	37·14n	15·15e
Sangro, *river*	42·14n	14·32e
Sannio, *historic region*	41·20n	14·30e
San Pietro, Isola di, *island*	39·08n	8·18e
Santa Maria di Leuca, Capo, *cape*	39·47n	18·22e
Sant' Eufemia, Golfo di, *gulf*	38·50n	16·00e
Sardinia (Isola di Sardegna), *island*	40·00n	9·00e
Secchia, *river*	45·04n	11·00e
Serchio, *river*	43·47n	10·16e
Sesia, *river*	45·05n	8·37e
Sicily (Isola di Sicilia), *island*	37·30n	14·00e
Sicily, Strait of	37·20n	11·20e
Sieve, *river*	43·46n	11·26e
Sile, *river*	45·30n	12·25e
Simplon Tunnel	46·15n	8·10e
Sinni, *river*	40·09n	16·42e
Spartivento, Capo, *cape*	38·53n	8·50e
Squillace, Golfo di, *gulf*	38·50n	16·50e
Strómboli, *volcano*	38·48n	15·13e
Tagliamento, *river*	45·38n	13·06e
Tanaro, *river*	45·01n	8·47e
Taranto, Golfo di, *gulf*	40·10n	17·20e
Taro, *river*	45·00n	10·15e
Tiber (Tevere), *river*	41·44n	12·14e
Trasimeno, Lago, *lake*	43·08n	12·06e
Tremiti, Isole di, *islands*	42·07n	15·30e
Trieste, Gulf of	45·40n	13·35e
Tirso, *river*	39·53n	8·32e
Tyrrhenian Sea	40·00n	12·00e
Ustica, Isola di, *island*	38·42n	13·10e
Varano, Lago di, *lagoon*	41·53n	15·45e
Vaticano, Capo, *cape*	38·38n	15·50e
Venezia, Golfo di, *gulf*	45·15n	13·00e
Vesuvio, Monte (Mount Vesuvius), *volcano*	40·49n	14·26e
Viso, Monte, *mountain*	44·40n	7·07e
Volturno, *river*	41·01n	13·55e
Vulcano, *island*	38·27n	14·58e

Today, the north is heavily populated, with numerous industrial cities and intensive agriculture, attracting steady migration from the south. The centre, focussed on Florence and Rome, traditionally an area of agriculture and local crafts, is becoming increasingly industrialized. The south, with the two ports of Bari and Naples and some recently developed industry, still preserves much of the traditional ways of life. The two islands, Sicily and Sardinia, are extensively cultivated, with citrus fruit and vineyards, pasture for sheep, fisheries, and a decreasing sulfur- and zinc-mining industry. The south and the islands are changing a great deal and gradually becoming more modernized. Within these four main divisions, the variety of the much smaller traditional regions is very great and depends on history as well as topography and economic conditions. Examples of different areas include Brianza, in Lombardy, a hilly region, highly industrialized; Monferrato, a group of hills in the Piedmont, given over to the production of wine; Mugello, the large, hilly basin of the Sieve River in Tuscany, strictly an agricultural area; Chianti, a hilly area of Tuscany famous for its vineyards and wines; and the Tavoliere di Puglia, as the tableland of Foggia is called, a dry and backward region where very ancient agricultural methods are still practiced. For a more thorough discussion of these regions, see elsewhere in this article.

SETTLEMENT PATTERNS

Rural areas. The majority of the population of Italy live in cities and villages; only a fraction live in hamlets, in very small clusters of houses, or in isolated houses.

Hydroelectric works and the Alpine landscape

In the long Alpine valleys the economy was always both agricultural and commercial, and there are many towns, such as Aosta and Bolzano, at the outlets of the lateral valleys. In settlements higher up or on the slopes of hills, an agricultural economy has remained predominant. On spurs of hillocks at the heads of valleys there are often old castles, originally built there for defense. The perpetual subdivision of landholdings makes a purely agricultural economy precarious in this region except in the upper Adige, where the Germanic system of primogeniture survived, producing the *masi,* family holdings that are passed on to the eldest son intact. Cattle raising remains profitable, but woodlands yield less return. Since the 1920s, hydroelectric works have been a feature of the Alpine rural scene, based on natural or artificially created lakes. These rural areas now also include an increasing number of tourist centres, such as Courmayeur and Valle d'Ampezzo. Although these developments have reduced both the seasonal and the permanent migration away from this area, rural living is, nevertheless, declining sharply here. In the band of Alpine and Apennine foothills, the villages, often situated on the knolls and flanks of the hills, are linked by roads that hold to the heights, away from the humid valley floors. Each village is usually grouped round a church, a castle, or a nobleman's palace, with its fields on the slopes around it and woodlands lower down. There are innumerable plum and cherry orchards and, above all, vineyards; their wines (Conegliano, Veronese, and Monferrato) are famous. Businesses are usually small or of only moderate size. Lombardy is the only area in which the ancient rural way of life has been displaced by the development of heavy industry. The population of its rural districts has been increased by migration from the neighbouring mountains and from the south. The Padano–Venetian–Emilian plain is the most important agricultural and stockbreeding region of Italy. The upland plain has now been virtually overrun by the great industrial centres such as Turin, Milan, and Busto Arsizio, but the lowland plain remains socially as well as economically rural. Wheat and corn are the most common crops, though each district has its specialty, such as sugar beet, grapes, and fruit.

Apennine settlement patterns

Villages high in the Apennines are less prosperous than those of similar altitude in the Alps. They are still isolated, the ground is infertile, and land is rarely owned by those who work it. These hopeless conditions have caused the more enterprising residents to emigrate to the north, leaving the villages in an even more desperate situation. Tourism and the expansion of cottage craft industries, such as the porcelain making at Gúbbio, near Perugia, have helped these towns survive. The lower hills and plains of Italy are covered with agricultural villages in which a wide variety of crops and vegetables are grown. The fields are heavily cultivated, but their yield is low. Specialized cultivation is more profitable, such as that in the south of hard-grain wheat, olives, almonds, figs, carobs, and hazelnuts. In Puglia and Basilicata large farms are staffed by labourers who live in urban centres, such as Cerignola and Altamura, and travel to work in the countryside. Some fertile and well-watered plains, such as the Neapolitan *campagna,* have a high level of productivity, especially of market vegetables. Here there is direct ownership of land and fairly dense settlement. In Sicily, settlement is minimal and scattered. Wheat is extensively cultivated. Especially on the coasts, pastureland is extensive though not very profitable; there is efficient cultivation of grapes, olives, citrus fruits, and vegetables, all of which bring the island some revenue, because they can be marketed as early produce. In Sardinia the settlement is also sparse and mainly inland, because of the need, in historical times, to avoid the dangers of malaria in low-lying areas and also of the risk of attack by pirates. Although islanders, the Sardinians have never wanted to work on the sea, and most of their fishing industry is carried on by men from the mainland. There are extensive meadows and forests.

Urban centres. From classical times and earlier, Mediterranean peoples have had highly developed urban centres. For historical as well as geographical reasons, Italy has never been dominated by one city, each district tending to possess its own urban centre. Today, there are several cities with a population of more than 1,000,000; but many more cities have a population of more than 100,000. Of these, almost half, including Sassari and Pisa, are on or near the sea; a similar proportion are in the north, and the rest are in the centre, in the south, and on Sicily and Sardinia. This irregularity of urban settlement reflects the economic imbalance among different parts of the country. The distribution of Italian cities also reflects historical and geographical conditions. In the Po Valley, cities such as Milan, Pavia, and Cremona are well placed for commerce, being situated at the confluence of roads or rivers. Another group of cities are those on the coast, built at the mouths of rivers, or on lagoons protected by sandbars; these include Savona, Genoa, Naples, Messina, Palermo, Ancona, and Venice. These cities, which originally grew up so close to each other, have, with increased population and industrialization, merged into enormous metropolitan complexes, sometimes characterized as mega-cities, such as that surrounding Milan. There are now several metropolitan areas in Italy, including Milan, Naples, Rome, Turin, Genoa, Florence, Palermo, and Bologna. (G.Na.)

The people

GROUPS HISTORICALLY ASSOCIATED WITH CONTEMPORARY ITALY

Non-Italian speakers

Ethnic and linguistic groups. Linguistically, modern Italy is fairly homogeneous. Non-Italian-speaking groups, a small minority, live mostly in the north of the country, where linguistic borders do not always coincide with political ones. The most important minority is the German-speaking population of the Adige River's upper reaches, in the province of Bolzano. It comprises most of the province's total population, while most of the rest speak Italian, and a few speak Ladine, a Neo-Latin language, a variation of which is also spoken in Switzerland. In 1921 there were about 195,600 German-speaking inhabitants; after an emigration during World War II, the German community started to expand again, with a relatively high birth rate. There were also once German-speaking minorities in Piedmont, Lombardy, and Venetia municipalities.

Slovene is spoken by a minority in the province of Trieste. Linguistic minorities exist in many municipalities scattered throughout the country. In some cases these groups have been rapidly diminishing. Among the Greeks, Albanians, and others, often only the older people speak the language. The Greek, Albanian, and Catalan minorities are historically interesting, because they are descended mostly from immigrants of the 14th and 15th centuries, and their language has retained archaisms while also being influenced by Italian.

Religions and races. *The religious background.* The overwhelming majority of Italians are Roman Catholics. Membership of other religious groups is marginal, and includes Evangelicals, Jews, and Greek Orthodox. The Jewish population was significantly reduced by Nazi and Fascist persecution, and since World War II their numbers have increased only through immigration.

Anthropological differences. The Italians vary anthropologically from area to area. The results of a survey based on about 300,000 conscripts born between 1859 and 1863 have been confirmed by subsequent research. They provide interesting conclusions about Italians from different regions in the late 19th century, though internal migration has made great subsequent changes.

Skull types

The cephalic index, a measure of the proportions of the skull, decreases from north to south; that is, brachycephalic (squat-headed) people are found in the Po Valley and around the Alps, extending into central Italy as far as the province of Rieti, though less marked. A similar area, even narrower, includes part of the Abruzzi, Campania, and Basilicata. Conversely, all the south is generally dolichocephalic (long-headed), particularly Calabria, southern Puglia, and Sicily, and most of all Sardinia. The north has some areas of pronounced dolichocephaly in east Liguria, southern Piedmont, and particularly northwest Tuscany.

In stature the inhabitants are generally smaller in the south. Three areas have relatively tall inhabitants: the largest area covers part of Venetia, the second is halfway between Tuscany and Emilia, and the third is in northeast Lombardy. The areas of small stature stretch from southern Marche to the south, becoming gradually more pronounced, especially in Sardinia. Although the greatest dolichocephaly and the smallest stature seem to coincide, there is no correlation; the small stature is largely the product of socioeconomic environment.

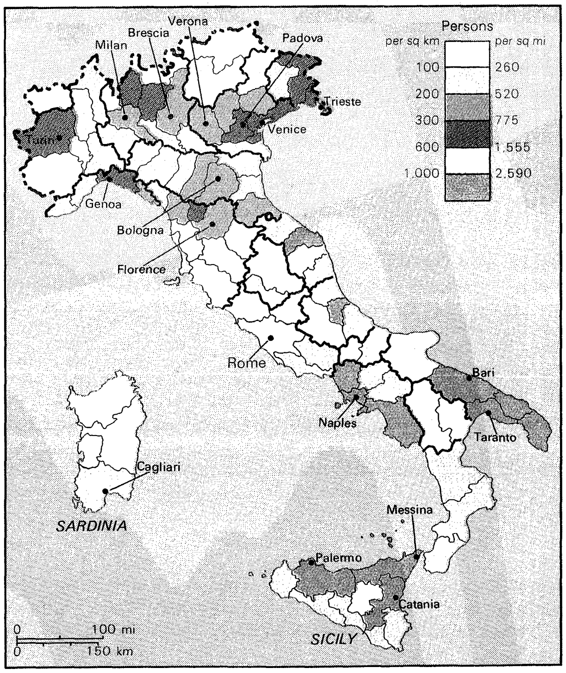
Italians are predominantly dark, though fair types are found in northern Italy, relatives of the fair natives of Savoy, Switzerland, and Austria. Throughout the Po Valley the people are notably darker than in Tuscany, Umbria, and Marche, in central Italy. Another area with fair types is found slightly farther south, in Sannio and Irpinia. Sardinia, again, has the highest frequency of dark types.

Despite modifications through internal migration, this analysis of anthropological characteristics is still reliable for studying the origins of Italians.

DEMOGRAPHY

Throughout the centuries Italy's population has shared many changes with other European countries. The mid-14th-century plague reduced its population considerably, and a long period of population growth ended at the beginning of the 17th century. From the early 18th century until the unification (in 1861), a slight, steady growth prevailed, though it was interrupted during the Napoleonic Wars.

From the latter half of the 19th century to the latter half of the 20th century, the population density more than doubled. The population increase is less than that for many other European countries. In the mid-20th century the rate of increase reached a level comparable to those in Great Britain, France, and West Germany. Population density is high for a country with territory that is one-third (35.2 percent) mountainous; only about one-half of the land is fit for arable farming.



Population density of Italy.

Checks to the population spiral

Birth and death rates. The decline of the birth rate during the 1890s, reflecting a growing awareness of family-planning techniques, occurred a decade or two after the death rate started to fall. The decline of each continued until the 1950s, levelling off thereafter. Migration drained the population until the eve of World War I; the loss was reduced between the wars, increasing again after World War II. Over the entire period, the net loss of the population has been several million people. The rate of natural population increase was particularly high in the decades when emigration was highest, although there is no obvious causal relationship between the two. The actual rate of increase of the population has remained approximately constant throughout the century. Fertility and mortality follow more advanced Western countries, although with a delay of one or more decades. The halt in the decline of Italy's birth rate lagged behind other countries because of its lower social and economic level and its slower development up to World War II. Peculiar to Italy are regional differences, particularly between north and south.

Regional differences. During the mid- and late 1900s most of the population in much of the north and the centre (particularly in Liguria, Piedmont, Friuli-Venezia Giulia, Tuscany, and Emilia) was not reproducing itself; that is to say, the average number of children per couple was under two. The south and the islands, on the other hand, had a comparatively high rate.

Factors tending to eradicate regional differences include high internal mobility, the policy of social and economic development for the south, the spread of mass communications, and the abolition of laws forbidding the advertisement of contraceptives and birth-control information.

Life expectancy. The mortality development follows a more uniform pattern. The death rate has declined significantly since 1880. But the death-rate level is heavily affected by the changing age structure and, since the 19th century, by the aging of the population.

The doubling of life expectancy since the late 19th century is well in line with European trends, reflecting higher nutritional and sanitation standards, improved medical care, and advances in medicine and pharmacology. Regional contrasts of mortality are less marked than those of fertility. Differences nonetheless arise, as a result of incidence of infectious diseases and deaths induced by "exogenous" factors, or peculiar environmental features. Malaria, typhus, and cholera once ravaged the south, while pellagra was a widespread cause of death in Lombardy and Venetia. Such incidences have been very much reduced by the advances of medicine, but mortality—particularly infant mortality—is still very much influenced by inadequate living conditions.

The decline in mortality rates is approximately the same for the four areas. Their relative situation has remained unchanged, with a much higher infant-mortality rate in the south and in the islands than in the rest of the country. The gap between Italy and other large western European countries is still considerable, though it is gradually narrowing. The poorer regions of Italy seem to have some advantages over the centre and north in their lower incidence of cancer and cardiovascular diseases.

Expectancy of life at birth is a good yardstick of the general sanitary conditions of the population. The south and the islands have nearly caught up with the rest of the country.

Migration. Italy is traditionally a country of emigration. Almost always Italians go in quest of work abroad, sometimes for a few years, sometimes for life. A little more than one-third of the total emigrated between the late 19th and late 20th centuries.

Changing character of emigration

The course of transatlantic emigration. During the 1860s and 1870s emigration emanated from the north, which long had contact with foreign countries. Emigration was largely associated with particular professions and skills: woodcutters and bricklayers went to neighbouring Alpine countries; farmers, vine growers, artists, and peddlers to North America. During the 1880s, however, emigration became a mass phenomenon, gradually involving the poor rural populations of the south. The flow rapidly increased as a result of improved internal and transatlantic transport facilities; the south's economic crisis, which hit both agriculture and industry, particularly in the late 1880s and 1890s; and the growing demand for labour in North and South America. Although emigration to Switzerland, Austria, and France was considerable, the destination was primarily transatlantic from the late 1880s to the early 1920s; about three-quarters of all the transatlantic emigrants since the late 19th century travelled during these four decades. Of the total of 12,000,000 who travelled in 100 years, about half went to the U.S., one quarter to Argentina, 1,300,000 to Brazil, and almost 1,000,000 to

Canada and Australia. During 1900–14, Italian emigration reached its peak, with 523,000 in 1906 and 565,000 in 1913, but the U.S.'s immigration acts of 1921 and 1924 practically closed the most important foreign labour market. Emigration to Brazil also declined considerably early in the 20th century because of the critical economic conditions of that country, and emigration to Argentina at the end of the 1920s slowed to a trickle. Subsequently, the economic depression, Fascism—which was hostile to emigration—and World War II practically eliminated emigration. After the war the very high unemployment revived emigration to South America, particularly Argentina and Venezuela. Traditional overseas emigration, however, had ended; with the economy's expansion, the decline of unemployment, and the assimilation of Italian minorities abroad, emigration for life fell to less than one-fifth of the total emigration. Departures to the U.S. and Canada in this period averaged 30,000 a year and to Australia about 15,000. An extraordinary feature in the early 1980s was that the number of returnees exceeded the number of departures.

Migration of the European labour market

European migration. During the mid-1900s, when there was still a considerable manpower surplus in the south, the islands, and parts of the north and centre, many of the unemployed took advantage of the tempting opportunities in other European countries, with rapidly expanding economies and pressing demands for labour. Thus, emigration to West Germany and Switzerland increased, while that to France (a traditional destination for Italian emigrants) and Belgium started to dwindle. This new wave was very different from the transatlantic emigration. Now the proportion of males was very high and entire families rare, because some countries refused entry to workers' relatives and some suffered from a housing shortage. The length of stay was very short, and repetitive expatriation for further periods of work abroad became common. As a result, the communities of Italian workers in Switzerland and West Germany became highly unstable, with a rapid exchange of individuals; many social and family problems were created both in the areas of origin and in the host countries. In the 1960s, migration between continental countries began to decrease, and Italian labour more often found worthwhile positions at home. Restrictive regulations, too, reduced the number of Italian workers in Switzerland.

Population movement within Italy. *Interregional flow.* Internal migration is a very important factor in the regional redistribution of the Italian population. In recent years the populations of the south and areas such as Venetia have had a much higher natural growth rate than the rest of the country. Nevertheless, the population's regional distribution has remained stable, with internal compensatory exchanges of population.

Since the unification of Italy, internal movements have followed the same direction, south to north and east to west: from the regions of the south and from the islands (especially Sicily) to centre regions—*e.g.,* Lazio–Rome and Tuscany—to the northwest—*e.g.,* Lombardy, Liguria, and Piedmont; and from the northeast (*e.g.,* Venetia) to the northwest. Movements from Emilia, Marche, and Umbria to other regions of the northwest and the centre have also been considerable. Mobility has grown exceptionally swiftly since World War II, corresponding to economic expansion, particularly in the industrial triangle of Lombardy–Piedmont–Liguria. Allied factors have been the reduction of unemployment in the north and a persistent labour surplus in the south. Italy is one of the few countries having a registration system for changes of residence from one commune to another.

Urbanization. Another aspect of the population redistribution in Italy, the vast urbanization, was especially rapid during the mid-20th century. The transformation of the economy that time—from mainly agricultural to mainly industrial—has strongly affected the distribution between rural and urban areas. The percentage of the labour force engaged in the primary sector has dwindled in the last several decades. A drop in the proportion of people living in the smaller *comuni* has been accompanied by a shift toward the large urban areas. The proportion of residents in the largest urban areas has tripled since the late 1800s. (M.L.-B.)

The economy

ITALY IN THE WORLD CONTEXT

Italy's economic growth after World War II was spectacular. At the end of the war, the country's economy lay in ruins, although much of the industrial machinery in the north had been saved from destruction. By 1951, prewar levels of production had been regained, and the next 20 years saw almost uninterrupted growth. Italy's economic growth since has been more sporadic, however, as it has faced rising labour costs and lower productivity, increasing petroleum prices, and worldwide recession. Italy cannot yet be classified as a mature industrial economy, although the transition period from a primarily agricultural society to a predominantly industrial one is nearing completion. The growing importance of Italy as an economic power can be gauged from the fact that, while the population of Italy has fallen, the proportion of Italy's national income to the world total has risen. Italy's greatest success in postwar years has been the rapid growth in industrial production, matched only by West Germany and Japan. The expansion of foreign trade has been a major factor in Italy's economic growth.

RESOURCES

Mineral resources. In terms of mineral resources Italy is one of the poorest countries in Europe. Both metalliferous and nonmetalliferous minerals are generally nonexistent or in short supply. The extractive industry has grown at a slower rate than other sectors because of both the poor quality of some minerals, such as coal, and the increasing inefficiency of the mines. Many of the sulfur mines in central Sicily and the coal mines in southwest Sardinia have been closed since World War II, and the decline continues. Italy has sufficient reserves of only a few minerals, such as mercury, sulfur, rock salt, and marble, while deposits of iron ore, coal, oil, and natural gas are meagre. Further reserves of hydrocarbons are being sought, especially in offshore areas. Italy also produces small quantities of zinc, lead, bauxite, pyrites, and a number of other minerals. Coal production now accounts for only a fraction of Italy's needs, and millions of tons a year are imported, mainly from the U.S., the EEC, eastern Europe, and Japan. Italy is a leading supplier of mercury, located mainly in the central part of the country, which accounts for a significant portion of the world output.

Deposits of both oil and natural gas exist in Italy, but the reserves of the former are rapidly being exhausted. After the discoveries of oil made in the 1950s at Ragusa and Gela, in Sicily, production rose and then gradually declined. The discovery of natural gas in the Po Valley and in various zones of southern and central Italy, including Sicily, raised hopes that Italy might be able to produce enough for domestic needs. Further deposits were discovered in the offshore Adriatic in the late 1960s, estimated at 2,120,000,000,000 cubic feet. Despite these discoveries, Italy initiated massive imports from diversified sources, both to meet the increased demand and to conserve total known deposits. Imported supplies came from Libya, the Soviet Union, and The Netherlands. Exploration, especially in the offshore areas, was intensified, above all by the State Hydrocarbons Corporation, the Ente Nazionale Idrocarburi (ENI).

Oil production accounts for only a minute part of domestic consumption. Imports of crude oil come mainly from the Middle East, Libya, and the Persian Gulf.

Italy's output of iron ore accounts for a small part of domestic needs. Imports come mainly from Liberia, Canada, Venezuela, and Brazil. Production of other minerals, such as pyrites, lead, zinc, and bauxite—deposits of which are scattered throughout the country—has been more or less static. Discoveries of rock salt, above all in Sicily, have boosted production. The quarrying industry—especially for marble and travertine (a hot-spring limestone deposit) from the world-famous quarries at Massa and Carrara (where Michelangelo, among other artists, found his raw material)—has enjoyed a modest but steady growth in postwar years, mainly as a result of the renewed popularity of the materials in the construction industry. Mining accounts for a very small percentage of the GNP.

Biological resources. Geologically, Italy is still at a very unstable stage of development. Earthquakes and tremors are quite frequent all over the mainland. Italy is characterized by a wide variety of climatic differences, topography, and soil types. For the most part, the terrain is rugged, arid, and unsuitable for intensive cultivation, except in the Po Valley, the plains of Puglia, and Campania. As a result of low rainfall, vegetation is sparse in parts of the country, especially in sectors of the Apennine Mountains. It has been mainly the pressure of population, however, leading to the cultivation of land generally unsuited for such use, that has caused widespread soil erosion and endangered the hydrogeological equilibrium of the land. The movement from the land of more than 6,000,000 peasants since 1945 has increased the dangers of neglect of property that was formerly cultivated or tended. Despite the exodus to the towns, about half of the productive land is still cultivated. As a result, Italy is deficient in timber and forest products, although the government is pursuing an active policy of reforestation. The balance of Italy's timber needs consists of imports of sawn wood, most of which comes from Austria and Sweden.

Timber production

Broad-leaved trees make up most of Italy's forest area, and conifers make up about one-fifth. Broad-leaved forests are fairly well spread over the country, with the exception of the Puglia, Sicilian, and Sardinian regions. Conifers are for the most part concentrated in the Alpine foothills, especially in the Alto-Adige region adjacent to the Austrian border. Because Italy is not well endowed with rivers, with the exception of the Po and the various smaller rivers that flow into the sea at or near the Po Delta on the northeast coast, intensive agriculture is concentrated in the overcrowded coastal plains, the mountain foothills, and above all in the Po Valley, the most fertile part of Italy. The need for land suitable for agriculture has led to many programs of land reclamation and improvement, especially in the south.

Hydroelectric resources. The three main lakes of northern Italy (Como, Maggiore, and Garda) are the source of six small rivers that flow southward into the Po River. These, together with the Bronta, La Sila, Tagliamento, and Adige rivers, which have their sources in the Alps, are the origin of most of Italy's hydroelectric power. As a result, more than half of the electricity produced in the northern regions of Piedmont, Lombardy, Alto-Adige, Venetia, and Friuli-Venezia Giulia is hydroelectric. Other regions of Italy with a similar proportion of hydroelectric power are Umbria, Abruzzo, and Calabria—all hilly or mountainous zones. Hydroelectric resources have been exploited almost to the maximum and account for more than one-fifth of the total energy output.

Power potential of the lakes

AGRICULTURE, FORESTRY, AND FISHERIES

Agriculture in Italy remains an important sector of the economy in terms of both employment and its contribution to the gross national product. Despite a number of agricultural reforms and EEC membership, agriculture is still backward by western European standards. The typical Italian farm is small, relatively unproductive, and geared more to subsistence than to modern market-oriented farming. Since 1945 there has also been a limited redistribution of land to small farmers, mainly under the 1950 agrarian reform acts. This has not stopped a massive flight from the land.

In Italy the complex land-tenure system is gradually being simplified by new legislation. The majority of farms are run by the owner and his family, without the employment of wage labourers; the remainder are run mainly under a sharecropping or rent system or, in the case of larger farms, with the employment of day labourers or permanent staff. The aim of agricultural policy since World War II has been to modernize agriculture by the development of efficient and well-organized farms, thus improving living standards for those working the land. This policy has been only partially successful. Objectives have been pursued largely through Green Plans, concentrating on mortgages and loans, land, irrigation, crop improvement, increasing profitable forms of production and productivity, and raising incomes.

The Green Plans

The most important areas for growing hard-grain wheat are in southern Italy, in the Sicilian and Puglian regions. Together they account for more than half of national production. The balance needed to meet domestic requirements is met by imports, mainly from the U.S. and Argentina. Domestic production of other cereals, notably rye, barley, and oats, has also been insufficient. The land area devoted to corn (maize) growing has been steadily increasing, but imports are still necessary. Italy is Europe's leading rice producer, most of it being grown in Piedmont and Lombardy; a large portion of its crop is exported to other European countries.

Wine and olive oil. Italy's most important agricultural products are wine and olive oil. Italy vies with France as the world's leading wine producer. A large quantity of the wine produced is not considered of high quality, but in a number of regions there are wines that compare favourably with the best. The government is active in upgrading Italian wines, and the system of *appellation contrôlée* (registered trade names) is gradually being extended to a wide range of Italian wines. Olives are grown principally in southern Italy; Puglia and Calabria are the main producing regions. Olive-oil production has been erratic. The grape and olive crops are of considerable economic importance to Italian agriculture, as most are processed on the farms themselves. Wine is exported in large quantities, although competition is particularly strong from other European countries, especially France and Spain.

Tomatoes and fruit growing. This crop has been one of the mainstays of agriculture in the south, particularly in Sicily and Campania, and canned tomatoes and tomato paste are exported throughout the world. Competition from Greece, Portugal, and Spain, however, has seriously eroded some traditional markets. Italy produces about one-half of all fruit grown in the original EEC area. Peaches, pears, and apples are very important and are grown primarily in the northeastern regions of the country. Citrus-fruit growing is also a major agricultural activity in southern Italy, especially in Sicily. Citrus fruits account for almost one-third of the value of Sicily's entire agricultural output. Because of high costs, the difficulties of mechanization in dense fruit groves, and an absence of marketing cooperatives, Sicily has sometimes been unable to sell all its crop, and consequently much has been withdrawn from the market.

Citrus fruits

Livestock. Production of livestock is inadequate, since with rising living standards the demand for meat has risen. The dairy industry also fails to satisfy domestic needs. Although a wide variety of cheese is produced—from goat, buffalo, and sheep milk, as well as from that of cows—it is done so mainly as a craft industry. As a result, imports of butter, cheese, and other dairy products are extensive.

Fisheries. Italy's waters are not well stocked with fish. Anchovies, sardines, and tuna account for a sizeable part of the total catch, while an important part of the remainder is made up of mackerels, mollusks, and shellfish. The gap between supply and demand is filled by large imports of fresh, frozen, dried, and salted fish coming mainly from Norway, Denmark, Japan, and Spain. The fishing industry still remains primarily an individual enterprise, although cooperatives and a few large fleets continue to expand. Some funds have been allocated for the renewal of fleets, but radical structural changes in the industry appear unlikely. Some fishing banks in or near Italian coastal waters are already exhausted, and many fishermen have left the industry in search of more profitable occupations.

Individualism in the fishing industry

INDUSTRY

Mining and quarrying. Mining is not an important sector of the Italian economy, and employment in it is rapidly declining. Coal and lignite production is declining gradually, as is that of sulfur, one of the few mineral resources to be found in any quantity in Italy. In any case, coal is gradually being replaced by oil, electricity, and natural gas. Likewise, production of iron ore, in which Italy is equally deficient, is destined to decline. The only sectors in which an increase is expected are in the mining of mercuric ores, rock and potassic salts (from deposits in Sicily), zinc, and asbestos, and in the extraction of natural gas. Quarrying activity remains high, however, especially because of the demand for marble, gravel, and other materials for the construction and road-building industries.

Manufacturing. The most remarkable feature of postwar economic development has been the spectacular increase in manufacturing. Since 1948, relative political stability has increased business confidence and led to annual industrial investment of more than 20 percent of produced income. Other important factors were Marshall aid (a special U.S.-financed aid program after World War II), safeguarding Italy during the crucial recovery period from 1946 to 1951 from balance-of-payments problems; an orthodox but sound monetary policy; and the liberalization of trade. The last factor has been especially important, since Italy's economic growth has been primarily based on exports. Another fundamental reason for Italy's rapid growth of manufacturing output has been the abundance of manpower. Italy's geographical position—near the oil fields of North Africa and the Middle East—was also an important stimulus to industry as a whole and to the development of the refining and petrochemical industries.

The steel industry. Despite a lack of mineral resources, a major expansion of the steel industry was begun in the early 1950s—one of the most farsighted and courageous decisions taken by the government. The steel industry has been the backbone of Italy's industrialization. Steel production has dramatically increased, as has crude-steel production. There have also been important qualitative changes in the industry.

Automobiles and electrical appliances. Parallel with the development of the steel industry has been that of the engineering and allied sectors. The automobile industry has grown spectacularly. Fiat is the main manufacturer, and the state-owned Alfa Romeo produces more expensive cars. The electrical appliance industry has played a major role in postwar industrial development, creating a European-wide market for refrigerators, cookers, washing machines, and dishwashers by selling at highly competitive prices and by using the latest production techniques.

The textiles industry. The textiles industry remains one of Italy's most important manufacturing industries. Silk production is centred in Lombardy, Piedmont, and Venetia but is also found throughout the country. Other important products include artificial and synthetic fibres and cotton, wool, and jute yarn. A large number of part-time piece-rate workers are employed.

Widening of the industrial base

Industrial diversification. An important element of Italy's manufacturing growth has been the widening of the base of industry. There has been a shift of emphasis from food and textiles to chemicals, steel, and engineering products. Large companies tend to dominate these new industries, but a vast number of small- and medium-sized companies, especially in the engineering and metalworking industries, are still important. Gradually, however, a process of consolidation is taking place, especially in those industries in which large-scale production is necessary for competitiveness. With Alfa Romeo, Fiat has run the automobile industry since it took over minor competitors such as Lancia, OM, and Autobianchi. Fiat and a rubber-and-cables company, Pirelli, have been instrumental in European mergers, Fiat taking an interest in the French companies Citroën and Michelin and Pirelli integrating its operations with that of the English company Dunlop. Other important companies in their respective sectors are Olivetti, making typewriters, calculating machines, and computer terminals; SNIA Viscosa, which leads the field in artificial fibres; and Montecatini Edison, which assumed control of SNIA in 1972, in the chemicals and pharmaceuticals sectors. The bulk of manufacturing output, however, comes from the small- and medium-sized companies, which are mainly concentrated in the north and centre.

ENERGY

By European standards, per capita power consumption is relatively low. Italy, mainly as a result of its proximity to the North African and Middle Eastern oil fields, has come to rely massively on oil for its energy requirements. Much of the crude oil is destined for re-export as refined petroleum products. Imports of oil are made by all the international companies as well as by Italy's State Hydrocarbons Corporation (Ente Nazionale Idrocarburi; ENI). Intensive exploration efforts by the ENI have resulted in a number of finds in Tunisia, Iran, Libya, and elsewhere. The ENI has also been searching for hydrocarbons in Italy. Discoveries in the Adriatic Sea have boosted reserves of gas, mainly in the Po Valley, Sicily, and parts of southern Italy. To conserve these domestic deposits as long as possible and to meet rising demand, the ENI has embarked on a large-scale import program of natural gas from diversified sources, including Algeria, Libya, The Netherlands, and the Soviet Union. Italy has also been developing its hydroelectric, nuclear, and coal-fired capabilities.

The State Hydrocarbons Corporation (ENI)

FINANCE

Italy's financial and banking system has a number of unique features, although its framework is similar to that of other European countries. The Bank of Italy is the central bank and the sole bank of issue. Power in monetary policy is vested in the Interministerial Committee for Credit and Savings, headed by the minister of the Treasury. In practice, the Bank of Italy enjoys wide discretionary powers and plays an important role in economic policy-making. Its primary functions include the control of credit and the formation and execution of monetary policy. There are three main types of banking and credit institutes. First, there are the commercial banks, which include three national banks, several chartered banks, the popular cooperative banks, whose activities do not extend beyond the provincial level, and ordinary private banks. Second, there is a special category of savings banks organized on a provincial or regional basis. Finally, there are the investment institutes, which collect medium and long-term funds by issuing bonds and supply medium- and long-term credit for industry, public works, and agriculture. The three national banks (Banca Commerciale Italiana, Banco di Roma, and Credito Italiano) are all owned by the largest state holding group, the Istituto per la Ricostruzione Industriale (Industrial Reconstruction Institute, or IRI). Many other important banks are also regulated by public law. Both these and the national banks are therefore subject to special government control and influence, although they act as normal profit-making banks in every respect. The savings banks are in a similar position, as are some of the major medium- and long-term-credit institutes, including the Istituto Mobiliaro Italiano (IMI), which is directly owned by the government and provides a considerable part of long- and medium-term funds for industry.

There are many institutes of various kinds supplying medium- and long-term credit. These special credit institutes have as their prime aim the increase of the flow and the reduction of the cost of development finance, either to preferential areas or to priority sectors (for example, agriculture or research) or to medium- and small-sized business. In addition to this network of special credit institutes, there is a subsystem of credit under which the government shoulders part of the interest burden.

The bond market in Italy is well developed. Mainly as a result of the special structure of government-sponsored institutes for development finance and subsidized interest rates, the growth of the capital market and stock exchanges is far less important than in other non-Communist industrialized countries. Consequently, the issuing of industrial debentures, too, except by the state holding corporation (IRI) and a few private and public companies, has been relatively meagre. There has been a high degree of self-financing by companies. The development of the stock exchange in Italy has been hampered by the archaic structure and rules of the markets and by tax problems connected with the registration of shares. Italy has also lagged in the development of mutual funds or unit trusts.

TRADE

Foreign trade makes a large contribution to the gross national product and economic growth. In absolute terms, however, Italy began to develop its foreign trade at a low point, and its rapid growth since the 1950s should be seen as a recovery from a low base rather than as a phenomenal growth. The main factors enabling Italy to catch up with its west European neighbours were relatively low labour costs, a much greater rise in productivity than in wage increases, and above all an abundance of manpower. The liberalization of international trade and the rise in demand in other countries were prerequisites to Italy's foreign-trade boom.

Trade deficit

While Italy has consistently had a trade deficit since 1957, the gap between imports and exports in most years has been narrow. At the same time, Italy has always had a considerable surplus of invisible exports, notably tourism and emigrants' remittances, which led to a structural surplus in the 1960s. This situation changed in 1970, however, when imports so outstripped exports that the invisible receipts were unable to bridge the gap. Italy in recent years has suffered from severe inflation, high unemployment, and chronic trade deficits.

In view of Italy's meagre natural resources, exports consist almost entirely of manufactured goods and agricultural produce. Engineering products, automobiles, and household appliances have been in the vanguard of Italy's export development, supplementing such traditional exports as fresh and processed fruit and vegetables, textiles, leather, and marble goods. Other products exported by Italy on a large scale include petroleum products and chemicals. The main impetus toward the growth in exports comes from three main sectors—transport equipment, machinery of all types, and clothing and textiles.

Imports, on the other hand, consist, to a large extent, of raw materials for industry, such as oil, iron ore, coal, cotton, and copper; foodstuffs, especially meat and cereals; and investment and consumer goods. The demand for foodstuffs has been particularly high; meat and cereal imports increased especially rapidly as living standards rose. Similarly, the propensity to import machinery and equipment with a high technological content has grown along with the process of industrialization in Italy. Another feature of Italy's trading pattern has been the rapid growth of trade with its Common Market partners. Italy has also been a pioneer of East–West trade, trading with the members of Comecon (Council for Mutual Economic Assistance, the east European Common Market).

ADMINISTRATION OF THE ECONOMY

The private sector. Italy's economy is characterized by a small number of large private and publicly owned industrial groups and a large number of medium-sized and small companies. The vast majority of small firms are of the artisan type; there are also many medium-sized industrial firms. At the other end of the scale there are the large private groups such as Fiat, Pirelli, Olivetti, and the labour-intensive household-appliance and textile companies. For the most part, private industry is located in the triangle formed by the cities of Milan, Turin, and Genoa, but other regions, in particular the northeastern ones of Venetia and Emilia-Romagna, developed an extensive network of industries in the mid-1900s. Central Italy has become increasingly industrialized. In the south and islands, industry tends to be concentrated in certain highly populated areas, such as Bari, Naples, and eastern Sicily, although incentives to encourage the further industrialization of the relatively poor south have been in force since the 1960s. Until the mid-1960s the burden of industrialization in the south had fallen almost entirely on the government and the public sector, while private industry invested mainly in capital-intensive sectors such as petrochemicals, which did little to boost employment. By fiscal and other incentives the government sought to encourage the growth of medium-sized and small industries in depressed areas. Foreign investment—mainly from the U.S.—played a significant role in the postwar development of Italy. The oil industry was the principal beneficiary, followed by the chemical, pharmaceutical, and engineering sectors.

The public sector and the role of government. By comparison with the other economies of western Europe, the public sector plays an exceptionally large role in Italian economic life. In addition, Italy has a unique formula for operating the majority of the state-owned corporations under the government's control. This formula applies in the case of the three major state holding corporations, the Istituto per la Ricostruzione Industriale (IRI), the Ente Nazionale Idrocarburi (ENI), and the Ente Finanziaria per L'Industria Meccanica (EFIM), all of which are responsible to the Ministry of State Participations, created in 1956. The IRI is organized on a pyramid basis, with the holding company at the top, a middle layer of financial holding companies divided according to the sector of activity, and below them a mass of operating companies, many of which are partly owned by private shareholders and quoted on the stock exchange. The IRI's activities are extremely widespread, including a number of major banks, Alitalia (the national airline), RAI-TV (the state radio and TV network), telephone and cable companies, manufacturing companies (such as Alfa Romeo, making automobiles and commercial vehicles), Italsider (the largest steel manufacturer), engineering companies, shipyards and shipping companies, and expressway construction. The ENI, the second largest state holding company, operates in the oil and natural-gas fields but also has interests in textiles, nuclear energy, and chemicals. Apart from this special involvement of the state in public utilities, manufacturing, and services, the government also controls the bulk of electricity generation and distribution through the nationalized electricity corporation (ENEL), almost the entire railway and road network, and the monopoly of tobacco and cigarette manufacture and salt. The IRI also has an important role in the Italian economy. Its function has been to act as a propellant to the industrialization of Italy and in some cases to act as a protective umbrella for industries in declining sectors; even more important, it also bypasses the archaic and generally inefficient bureaucracy. The IRI built the bulk of Italy's extensive expressway system in record time and is now increasingly called in to carry out public works of all kinds that the state administration had proved incapable of executing. A prime example was the low-cost public-housing program that the government began implementing in the 1970s. The IRI and the ENI between them have made a vital contribution to the growth of the Italian economy since the 1950s. In particular, the two groups have been the principal investors in the underdeveloped southern part of Italy. Although responsible to the Ministry of State Participations, both the IRI and the ENI enjoy considerable financial and operative autonomy. This autonomy is reinforced by the fact that these state corporations do not obtain their funds from the Treasury. The lack of unified control over state corporations and the absence of centralized planning have reinforced their independence.

Main public corporations

Economic policy is established by the Interdepartmental Committee of Economic Planning and carried out under the supervision of the Minister for State-Controlled Enterprises and the Court of Accounts. Many public corporations carry on economic activity in competition with similar private companies, but the most notable form of state intervention in industry and business concerns is through state-run holding companies such as IRI and ENI. The state usually but not necessarily has a controlling interest in the companies in which it holds stock. Public corporations are also responsible for most medium- and long-term loans and about half of all life insurance.

The government has also intervened on a massive scale in the development of the southern part of Italy and the islands, known as the Mezzogiorno. The southern Italy development fund, called the Cassa per il Mezzogiorno, was set up in 1950, to stimulate investment in agriculture and industry in the south between 1951 and 1970. This institution represented the first step by the Italian government toward national economic planning. An attempt at more widespread planning came in 1954, but it was not until 1967 that the First Five-Year Plan was approved by Parliament. The plan, which was of the flexible type, updated annually, was binding only on the public sector. In the 1980s an economic reform package was initiated in order to cut consumption and reduce public spending. The national economic planning body (Comitato Inter-

The five-year plans

ministeriale per la Programmazione Economica, or CIPE) was assigned overall responsibility for development of the south, with the Cassa per il Mezzogiorno acting as the executive body, as well as supervising overall planning in Italy.

Taxation. One important feature of the Italian fiscal system is the low percentage of receipts accounted for by income tax levied on persons and companies. The main source of revenue comes from various taxes on goods and services, in particular a general tax (IGE) normally levied on all transactions. Indirect taxes, on the consumer, monopolies, and business transactions, make up the rest.

In the 1950s and 1960s both the complications of the tax system and the ease with which many direct taxes could be evaded reduced the government's possibilities of using fiscal policies as an instrument of economic management. Furthermore, Italy traditionally relied on monetary rather than fiscal policy to direct the economy. While tax evasion in effect constituted a huge subsidy to the self-employed and the rich, it also acted, more significantly, as a powerful stimulus for growth, since firms benefitted from the weakness of the tax system too. By 1970, however, it was generally agreed that a better tax base to implement social reforms—housing, hospitals, schools, and urban transport—was badly needed. The first answer to this problem was the tax-reform bill of 1971, which introduced a single progressive tax on personal income. A major reform of indirect taxation was introduced in 1973, when the value-added tax replaced the turnover tax and a number of other indirect taxes.

Principal trade unions

Trade unions and employer associations. The Italian union movement that emerged after World War II and 20 years of suppression under the Fascist regime had, not surprisingly, strong political orientations. In 1947 breakaways from the Communist- and Socialist-dominated General Confederation of Labour led to the formation of two rival union confederations—the Confederazione Italiana Sindacati Lavoratori (Italian Confederation of Workers' Trade Unions, or CISL), which was dominated by Catholics and adherents of the Christian Democrat Party, and the Unione Italiana del Lavoro (Italian Union of Labour, or UIL), controlled by a troika of Socialists, Social Democrats, and Republicans. The biggest of the three union movements remained the Communist-dominated General Confederation of Labour (CGIL), which until the late 1960s was considered a voice of Communist Party policy. Although relatively weak in numbers and lacking in funds to sustain the cost of prolonged strikes, the unions showed a new aggressiveness toward the end of the 1960s and won major wage increases. At the same time, as a result of the weakness of the government and the political parties, the unions took on a mantle of authority. A series of 24-hour general strikes took place, calling for reforms including pensions, more and cheaper housing, and a new approach to the problems of the depressed south. While the unions acquired new status and authority, production was badly affected at a time when industry was suffering from sharply rising costs. In the 1970s the unions extended their intervention beyond wages and political and social reforms to the organization of productive methods inside factories. The principal problem for the unions is to reforge the unity of the three main confederations.

General Confederation of Industry

The principal employers' association in Italy is the Confindustria (General Confederation of Industry). Under its wing are a vast number of smaller employers' associations organized either on territorial or on industrial lines. Confindustria has exercised considerable political and economic influence. When it lost a determined battle against the nationalization of electricity in 1962, however, its influence waned for several years. Its loss of influence was accentuated when its traditional political ally, the conservatively oriented Liberal Party, was replaced in the government by the Socialists in the same period. Confindustria, however, regained some of its old vigour, partly as a result of a thorough reorganization.

Current economic policies. After the reconstruction following World War II, a main objective of economic policy was to ensure rapid growth while maintaining a strong currency and external-payments position. To a large extent this goal was achieved with the aid of an adroit monetary policy, the development of a highly successful export trade, large-scale imports of raw materials, and the liberalization of trade. Less successful were the efforts to reduce unemployment and underemployment, to develop southern Italy and the islands, and to reduce the disparity in incomes between industry and agriculture. In the space of only 20 years, however, Italy achieved a wealth that few could have predicted in 1945. Subsequent economic policies were directed at maintaining an annual growth rate while transferring resources on a far larger scale from private to public consumption. Much of the success of economic policies depends on political stability and modernization of the machinery of state.

In 1970 there was the beginning of a break in the almost uninterrupted growth since 1950. A large increase in costs, reducing the competitiveness of Italian exports, plus stagnating investments, forced the government to intervene with inflationary measures. Growing inflation accompanied by high unemployment, however, subsequently prompted the government to pursue restrictive measures, including reductions in public spending.

(E.I.U.)

TRANSPORTATION

Movement patterns and means of transportation. As a source of revenue, transportation has accounted for 10 percent of Italy's gross national product since World War II, while net investments in it have averaged around 6 percent of the national total. The proportion of the national wealth actually tied up in transportation is tending to diminish because of rapid obsolescence of transportation equipment and structures and the slowness of appropriate renewal projects. The Italian roads are approaching saturation, while costs of maintenance and modernization are rising. Attempts to alleviate road traffic include modernizing the railways and building subway networks. A law effective since 1969 provided that metropolitan systems should be financially supported by each interested municipality.

The capital investment background

Traffic is most intense in northwest Italy, where Turin, Genoa, and Milan form the industrial triangle; but a dense road network there distributes traffic fairly evenly. The traffic magnet of northeast Italy is paradoxically the canal-riddled city of Venice, the country's third largest port. The most popular route is Venice–Padua–Verona–Milan, which, with Genoa, forms Milan's second traffic outlet. In summer the traffic density more than doubles, with tourists travelling to Venice and the coast.

In central Italy, traffic is massed on the north–south Bologna–Florence–Rome route, and, except on the Florence–Livorno tourist route, there is little east–west traffic. This concentration has isolated certain zones, such as southern Tuscany and northern Lazio (Latium). Umbria is more developed, with important steelworks at Terni and industrialization progressing rapidly along the Terni–Perugia route. Traffic in the Abruzzi is insignificant, but the new Rome–l'Aquila–Pescara highway should open the depressed region to development.

Rome is the traffic hub of central Italy, and, while traffic with Lazio is rather low, the level is high toward the south (Latina) and the southeast (Frosinone), where industry is developed. The density remains constant along the whole route from Rome to Naples, declining sharply farther south, where conditions are more backward. The small flow of traffic along the national coastal route reflects Mezzogiorno's economic depression. The Adriatic coastal route, the Naples–Bari route, and the Taranto–Cosenza route are little used; traffic increases along the Adriatic, between Puglia (Foggia, Bari, and Taranto) and the areas of the Po Valley, and becomes heavier, in fact, than between Naples and Rome.

Sicily and Sardinia have little traffic. On Sicily the Messina–Palermo, Agrigento–Palermo, and Messina–Catania–Syracuse routes are pre-eminent, with the heaviest traffic on the stretch between Catania and Syracuse. On Sardinia the only significant route is Cagliari–Oristano–Sassari.

Components of Italian transportation. *The road net-*

Categories of roads

work. The Italian road network is subdivided into four administrative categories—express highways (*autostrade*) and national, provincial, and municipal roads (*strade statali, strade provinciali, strade comunali,* respectively), to which urban streets can be added.

The mid-1900s witnessed extensive construction of express highways. The principal branches are the Po Valley, joining Turin, Milan, Austria, and Yugoslavia, with important branches to ports and mountain passes; the peninsular route, joining Milan to Naples via Bologna, Florence, and Rome (the Autostrada del Sole); the Tyrrhenian coastal route from the French border to Reggio Calabria in the far south, passing through Genoa, Rome, and Naples; the Adriatic coastal route from Bologna, joining the Tyrrhenian route in Calabria; and its Sicilian extension, the Messina–Palermo highway.

The two coastal routes are linked by two transverse highways, Rome–Pescara and Naples–Bari. Construction of the network was impeded by mountainous stretches, which were traversed by tunnels and viaducts, some of them splendid examples of modern architecture, such as the elevated roadway joining the Tyrrhenian coastal route to the Genoa road. Most highways have four lanes (the Milan–Turin has six) and are 50 feet wide, with shoulders of 10 feet and a central reservation. The construction and administration have been entrusted both to state-run and to private companies. The chief concession holder is the Autostrade agency of the national Istituto per la Ricostruzione Industriale (IRI). There are tolls, calculated on distance travelled and legal horsepower, to cover construction costs.

In the north, highway traffic is mainly commercial, the coastal routes concentrate on tourist areas, and in the south they open up economically backward areas. Serious saturation occurs on certain stretches due to increased traffic flow.

The completion of the southern network is envisaged to redress the balance between north and south. The 13-kilometre Mount Frejus highway tunnel was opened in 1980 and links Italy to France through the Alps.

The *strade statali* have been relieved of much of the long-distance traffic by the *autostrade.* These national roads now carry local traffic, which is dense in places. The principal national highways are the consular roads, so-called because they follow the roads built by ancient Rome's consuls. They radiate from Rome to various parts of Italy, the most famous being the Via Aurelia, linking Rome and Genoa, the Via Cassia, linking Rome and Genoa via Florence, and the Via Appia, linking Rome and Brindisi. Other important national highways are the Padana, joining Turin to Venice via Milan; the Emilia, joining Milan with Bologna and the Adriatic; and the Brenner Highway, from Pisa to the Brenner Pass.

Strade statali are maintained by a state agency, the Azienda Nazionale Autonoma della Strada (the ANAS, or National Road Board). A flow of more than a million new vehicles a year makes widening a continuing necessity.

Railways. Italian railway lines are administered by the Azienda Autonoma delle Ferrovie dello Stato (the FS, or Italian State Railway Board). The FS also administers maritime lines and bus services.

The three types of railway network

The railway network is of three types: national, regional, and urban (metropolitan). Over long distances the railways have advantages over road transportation, and the national lines are therefore the most heavily used. The network of principal national lines corresponds to that of the express highways: parallel to each express highway is a railway line, electrified and with a double track.

The most important line is the peninsula route Milan–Bologna–Florence–Rome–Naples, of which the most heavily used stretches are Milan–Bologna and Rome–Naples; the first section of the high-speed direct Florence–Rome railway was opened in 1977. The Turin–Milan–Venice and Milan–Genoa lines link the industrial belt of the country with Genoa and Venice for freight transportation. Other principal lines are the Genoa–Rome–Reggio Calabria–Palermo and Bologna–Ancona–Bari coastal routes. The principal mountain routes linking the Italian railways with the rest of Europe are Turin–Frejus (to France), Milan–Simplon Tunnel (to Switzerland), Verona–Brenner (to Austria and Germany), and Venice–Tarvisio (to eastern Europe). Railway traffic, which had levelled off with the completion of the numerous express highways, began increasing again in the late 1960s with attractive new rolling stock and improved tracks.

Whereas the long-distance railway network is satisfactory, the regional lines are not. Competition from road transportation, the antiquated lines (mostly single-track and not electrified), and the state of the track have brought these lines into disuse, and many have been abandoned for more economical buses. The increasing saturation of the highways and the commuters' need for rapid and economical service have, however, made reopening the abandoned lines seem attractive. There have been efforts to reorganize them on a metropolitan basis and to introduce second tracks and electrification for key sections in order to reduce travelling time.

Efforts have also been made to coordinate the regional trains with urban traffic networks. The former usually run with the great national lines in the vicinity of the cities. Because the national trains have the right of way there are slowdowns and reductions in frequency. These might be prevented by integrating regional with metropolitan lines.

Railways, with stiff competition from automobiles and a tariff policy that tends to emphasize the social uses of trains (*i.e.,* that provide for low fares), are increasing their unprofitability. The FS has been striving for competitiveness by renovating and modernizing.

Water transportation. The principal dry-cargo Italian ports are Venice, Cagliari, Civitavecchia, and Piombino, while those handling chiefly petroleum products are Genoa, Augusta, Trieste, Bari, and Savona. Those handling both sorts are Naples and Livorno.

Most port traffic consists of imports, land transportation being preferred for exports.

Many ports are inadequate. The more congested larger ports, such as Genoa, Savona, Venice, and Livorno, have imposed primage (a small addition or percentage added to the freight rates) at the time of entry and sometimes at departure, which increases production costs.

In the north of Italy there are only a few large ports, such as Savona, Genoa, and Venice. Nearby, there are industrial ports, such as La Spezia and Ravenna, ports of local importance, such as Oneglia and Monfalcone, fishing ports, such as Chioggia, and tourist ports, such as the numerous ones of the Ligurian Riviera. Half of the commercial port traffic is concentrated on only one-tenth of the coastline. The industries of Lombardy and Piedmont make heavy demands on the maritime outlets, particularly at Genoa, which, although the most extensive and important Italian port, has great difficulty in expanding because of the mountains surrounding it.

In central and southern Italy there are several harbours, but most are modest and function locally. There are a few large ports in these regions, some born of recent industrial development, such as Augusta (about 15 miles north of Syracuse in Sicily), a petroleum port that serves an immense refinery. Prospects for the central southern ports depend on industrialization in the hinterland.

The coastal navigation on the Genoa–Gela–Taranto–Venice route is very important, as are the lesser ports of La Spezia, Civitavecchia, Naples, Ravenna, Trieste, and Cagliari.

Air transportation. Alitalia, the air-transport company, is part of the Istituto per la Ricostruzione Industriale group, in which the government participates. Rome, Milan, Naples, Genoa, Venice, Palermo, Catania, and Turin all have major airports, foremost being the airports of Rome (Fiumicino, Ciampino, and Urbe) and the Milanese airports (Linate and Malpensa). Italian airport traffic has a high rate of increase, calling for improvement of airports, services, and instruments for controlling air traffic. Aircraft enterprises, though few, have reached high technical levels, participating in international construction projects and producing aircraft on foreign license. Electronic and air-space plants have been planned for the south as part of its development.

(M. Del V.)

Administrative and social conditions

GOVERNMENT

Constitutional framework. The Italian state grew out of the Kingdom of Piedmont and Sardinia, where, in 1848, King Charles Albert introduced a constitution that remained the basic formal law of his kingdom and, later, of Italy for nearly 100 years. It provided for a bicameral Parliament with a Cabinet appointed by the king. With time, the power of the crown gradually diminished, and ministers became responsible to Parliament rather than to the king. Although the constitution remained formally in force after the Fascists seized power in 1922, it was devoid of all substantial value. On June 2, 1946, after the end of World War II, the Italians voted in a referendum to replace the monarchy with a republic. A Constituent Assembly worked out a new constitution, which came into force on January 1, 1948.

The constitution has built-in guarantees against easy amendment, in order to make it virtually impossible to substitute for it a dictatorial regime. It is upheld and watched over by the Constitutional Court. The republican form of government cannot be changed. The constitution contains some perceptive principles, applicable from the moment it comes into force, and some programmatic principles, which can be realized only by further precise legislation.

The constitution is preceded by the statement of certain basic principles, including the definition of Italy as a democratic republic based on work, in which sovereignty belongs to the people. Other principles concern the inviolable rights of man, the equality of all citizens before the law, and the obligation of the state to abolish social and economic obstacles that limit the freedom and equality of citizens and hinder the full development of individuals. The constitution guarantees many forms of personal freedom: the privacy of correspondence; the right to travel at home and abroad; the right of association for all purposes that are legal, except in secret or paramilitary societies; and the right to hold public meetings, if these are consistent with security and public safety. There is no press censorship, and freedom of speech and writing is limited only by standards of public morality. The constitution stresses the equality of the spouses in marriage and of their children, although the old Civil Law Code regards the husband as head of the household. Family law has seen many reforms, including the introduction of divorce.

Personal freedom

One special article in the constitution concerns the protection of linguistic minorities. Religious liberty is conditioned by the Lateran Treaty made with the Vatican in 1929, which asserts that both state and Roman Catholic Church are independent and sovereign in their own spheres. The constitution establishes the liberty of all religions before the law but adds that churches other than the Roman Catholic Church are free to act according to their charters only so long as they do not conflict with the general law. The Catholic Church has retained considerable privileges, particularly in tax exemptions and in jurisdiction over church marriages in cases of nullity. The position of the Catholic Church is the cause of considerable friction in Italian political life. Despite the liberal tendencies of the Second Vatican Council, the church defends its privileges tenaciously and continues to fight any lay attempts to reform, for example, certain aspects of family law and divorce law.

The constitution is upheld by the Constitutional Court, which comprises 15 judges, of whom five are nominated by the president of the republic, five by Parliament, and five by various legal bodies. Members must have certain legal qualifications and experience. The term of office is nine years, but Constitutional Court judges are not eligible for re-appointment. The court performs four major functions. First it judges the constitutionality of state and regional laws and of acts having the force of law. Secondly, the court resolves conflicts of competence of jurisdiction between ministries of administrative offices of the central government or between the state and a particular region or between two regions. Thirdly, it judges indictments instituted by Parliament. When acting as a court of indictment, the 15 Constitutional Court judges are joined by 15 additional lay judges chosen by Parliament. Fourthly, the court determines whether or not it is permissible to hold referendums on particular topics. The constitution specifically excludes from the field of referendums financial decisions, the granting of amnesties and pardons, and the ratification of treaties.

The legislature. Parliament is bicameral and comprises the Chamber of Deputies and the Senate, both elected by popular vote and with equal powers. In theory, the Senate should represent the regions and in this way differ from the lower chamber, but in practice the only real difference between them lies in the minimum age required for the electorate and the candidates: 18 and 25, respectively, for deputies and 25 and 40 for senators. Deputies and senators alike are elected for a term of five years, which can be extended only in case of war. Parliamentarians cannot be penalized for opinions expressed or votes cast, and constituents cannot oblige their deputy to vote according to their wishes. Deputies and senators enjoy immunity from arrest, criminal trial, and search. Their salary is established by law, and they qualify for a pension.

Both houses are officially organized into parliamentary parties. Each house is also organized into standing committees, which reflect the proportions of the parliamentary groups. Besides studying bills, these committees act as legislative bodies. The new parliamentary rules have followed the United States' pattern and have given the standing committees extensive powers of control over the government and administration. Parliament also sets up special joint investigatory committees.

Special majorities are required for constitutional legislation and for the election of the president of the republic, Constitutional Court judges, and members of the Superior Council of the Judiciary. An unusual feature of Italian parliamentary procedure is use of a secret ballot. Votes of confidence are necessarily made openly, while voting in presidential elections is by secret ballot; in normal divisions, voting can be either open or secret, though the secret ballot is generally preferred. While granting parliamentarians greater independence, it enables them to vote contrary to party instructions. Moreover, it effectively prevents any control of the representatives by the electorate.

Secret voting in Parliament

The two houses meet jointly to elect and swear in the president of the republic and to elect one-third of the members of the Superior Council of the Judiciary and one-third of the judges of the Constitutional Court. They may also do so to indict the president of the republic, the president of the Council of Ministers, or any individual ministers.

Each year, the budget and the account of expenditure for the past financial year are presented to Parliament for approval. The budget, however, does not cover all public expenditure, nor does it include details of the budgets of many public bodies, over which, therefore, Parliament has no adequate control. International treaties are ratified by means of special laws.

The most important function of Parliament is ordinary legislation. Bills may be presented in Parliament by the government, by individual members, or by other bodies, such as the National Council of Economics and Labour, various regional councils, or communes. Bills are passed either by the standing committees or by Parliament as a whole. In either case, the basic procedure is the same. First, there is a general debate followed by a vote; secondly, each separate article of the bill is discussed and voted on; finally, a last vote is taken on the entire bill. All bills must be approved by both houses before they become law, so, whenever one house introduces an amendment to the draft approved by the other house, the latter must approve the amended draft. The law is then promulgated by the president of the republic unless he considers it unconstitutional or inappropriate; in that case, he remands the bill to Parliament for reconsideration. If the bill is, nevertheless, passed a second time, the president is obliged to promulgate it. The law comes into force when published in the *Gazzeta Ufficiale.*

The presidential office. The president of the republic is irremovable, and his seven years of office cannot be shortened. He is elected by a college comprising both chambers of Parliament, together with three representatives from every region. The two-thirds majority required guarantees that the president is acceptable to a sufficient proportion of the populace and of those in public life. The minimum age for presidential candidates is 50. If the president is temporarily unable to carry out his functions, the president of the Senate acts as his deputy. If the impediment is permanent or if it is a case of death or resignation, a presidential election must be held within 15 days.

Special powers and responsibilities are vested in the president of the republic. In certain cases, his powers exceed those of the government, which must, however, always countersign his acts. He can be indicted for high treason or failure to uphold the constitution. He has the power to call special sessions of Parliament, to promulgate laws and delay legislation, to authorize the presentation of government bills in Parliament, to promulgate executive orders, and, with Parliamentary authorization, to ratify treaties and declare war. He commands the armed forces and presides over the Supreme Council of Defense and the Superior Council of the Judiciary. He has the power to dissolve Parliament either on his own initiative or at the request of the president of the Council of Ministers. He may appoint five life members of the Senate and appoints five of the 15 Constitutional Court judges. He also appoints the president of the Council of Ministers, the equivalent of a prime minister. It is his duty, whenever a government is defeated, after consulting eminent politicians and party leaders, to appoint the person most likely to win the confidence of Parliament; the large number of political parties gives him a very real choice. The president of the republic grants amnesties and pardons on the advice of Parliament.

The government. The government comprises the president of the Council of Ministers and the various other ministers responsible for particular departments, whom he has nominated. They are appointed to office by the president of the republic. Each new government must receive a vote of confidence in both houses of Parliament within 10 days of its appointment. If at any time the government fails to maintain the confidence of either house, it must resign. Splits in the coalition of two or more parties that had united to form a government have sometimes caused resignations of governments. The president of the Council of Ministers is not merely the first among ministers of equal merit but is solely responsible for directing government policy and coordinating administrative policy and activity. Ministers are responsible jointly for the acts of the council, such as the emanation of decree laws and, severally, for the acts of their ministries. The government is the summit of executive power. In times of emergency, it can issue decree laws signed by the president of the republic, provided such laws are presented to Parliament for authorization the day they are issued and receive its approval within 60 days. Without such approval they automatically lapse. The government and, in certain cases, individual ministers issue administrative regulations and provisions, which are promulgated by presidential decree.

President of the Council of Ministers

Regional and local government. The republic is divided into regions, provinces, and communes. There are 15 ordinary regions and an additional five (Sicilia [Sicily], Sardegna [Sardinia], Trentino-Alto Adige, Friuli-Venezia Giulia, and Valle d'Aosta) to which special autonomy has been granted. The regions with ordinary powers are Piemonte (Piedmont), Lombardia (Lombardy), Veneto, Liguria, Emilia-Romagna, Toscana (Tuscany), Umbria, the Marche, Lazio (Latium), Abruzzi, Molise, Campagnia, Puglia (Apulia), Basilicata, and Calabria. Italy can thus be considered a regional state. The modern regions correspond to the traditional territorial divisions. The powers of the five special regions derive from special statutes adopted through constitutional laws. The organs of regional government are the Regional Council, a popularly elected deliberative body with power to pass laws and issue administrative regulations, the Giunta Regionale, an executive body elected by the council from among its own members, and the president of the Giunta Regionale. The Giunta Regionale and its president are required to resign if they fail to retain the confidence of the council. Voting in the regional councils is rarely by secret ballot.

Participation in national government is a principal function of the regions: regional councils may inaugurate parliamentary legislation, propose referendums, and appoint three delegates to assist in presidential elections. In regional legislation the five special regions have exclusive competence in certain fields, while the ordinary regions have competence within the limits of fundamental principals established by state laws and including areas such as agriculture, forestry, and town planning. The legislative powers of both special and ordinary regions are subject to certain constitutional limitations, the most important of which is that regional acts may not conflict with national interests. The regions can also enact legislation necessary for the enforcement of state laws when the latter contain the necessary provisions. The regions have administrative competence in all fields where they have legislative competence. Additional administrative functions can be delegated by state laws. The provision that normally regional administration is to be carried out at provincial and commune level aims at avoiding excessive bureaucracy. The regions are financially autonomous; they have the right to acquire property and the right to collect certain revenues and taxes.

The state has powers of control over the regions. The validity of regional laws that are claimed to be illegal can be tested in the Constitutional Court, while those considered inexpedient can be challenged in Parliament. State supervisory committees presided over by government-appointed commissioners exercise control over administrative acts. The government has power to dissolve regional councils that have acted contrary to the constitution or have violated the law. In such an event, fresh elections must be held within three months.

Communes

The organs of the commune, the smallest local government unit, are the popularly elected common council, the Giunta Comunale, or executive body, and the mayor. Both the Giunta Comunale and the mayor are elected by the council from among its own members. The communes have the power to establish and collect local taxes; they have their own police; they issue ordinances and run certain public health services; and they are responsible for such services as public transport, garbage collection, and street lighting. Control over the activity of the communes, until recently vested in the state and exercised by the prefects, has been transferred to the regions. Common councils may be dissolved for reasons of public order or for continued neglect of their duties.

The organization of the provinces, units midway in size between regions and communes, is analogous to that of the communes; they have councils, giunte, and presidents.

There are certain central-government officials whose duties lie in the sphere of local government. These include the government commissioner of the regions, who supervises the administrative functions performed by the state and coordinates them with those performed by the region; the prefect, resident in each province, who is responsible for enforcing the orders of central government and has powers of control over the state organs of the province and communes; and the *questore,* who is the provincial chief of the state-run police. Certain local-government officials also have central-government duties: among them are the president of the Giunta Regionale who, in directing the administrative functions that the state delegates to the region, performs a specific state duty; and the mayor of a commune who, in his capacity as an agent of the central government, registers births, deaths, marriages, and migrations, maintains public order (though, in practice, this is dealt with by the police), and can, in cases of emergency, issue ordinances concerning public health, town planning, and the local police.

The political process. *Elections.* In Italy there are parliamentary, regional, and local elections. Systems of proportional representation are used in the elections of the Chamber of Deputies and the Senate. Regional elections are governed by state laws and are also based on proportional representation. A system of limited vote is used in municipal elections by communes with less than 5,000 inhabitants, while the more highly populated communes use a list system of proportional representation. The system used in provincial elections is analogous to that used in senatorial elections.

Political parties. The constitution guarantees all citizens the right to associate freely in political parties in order to contribute through democratic procedure to the determination of national policy. The essential characteristic of a democracy is the existence of rival political parties. A plurality of parties is encouraged by systems of proportional representations and in Italy has led to the formation of several principal parties, as well as other smaller parties. Participation in the primary elections (to make up the lists of candidates) is strictly limited to party members and hence is free from outside control. Parties are not required to publish accounts or disclose the source of their income, which frequently comes from public bodies, pressure groups, and individuals.

The chief Italian political parties are the Democrazia Cristiana (DC; Christian Democratic Party); the Partito Comunista Italiano (PCI; Italian Communist Party); the Partito Socialista Italiano (PSI; Italian Socialist Party); the Partito Socialista Democratico Italiano (Italian Social Democrats); the Partito Liberale Italiano (PLI; Italian Liberal Party); the Movimento Sociale Italiano (MSI; Italian Social Movement [neo-Fascists]); the Partito Repubblicano Italiano (PRI; Italian Republican Party); the Partido Radicale (PR; Radical Party); and the Südtiroler Volkspartei, or the Partito Populare Sud Tirolese (SVP; South Tirolean People's Party).

The Christian Democratic Party, in practice supported by the church hierarchy, aims to unite all Italian Catholics in a single political grouping. For this reason it contains both highly conservative and strongly progressive elements. The consequent difficulty in forming coherent policies reduces its electoral impact. The Italian Communist Party is the largest European Communist party outside the Soviet bloc. As an opposition party it wins most of the floating protest votes of the electorate, besides those of its own members. The Socialist Party, a section of the Socialist International, was formed in 1967 by the amalgamation of the Italian Socialist Party, led by Pietro Nenni, with the Italian Social Democrats, led, until his election as president of the republic in 1964, by Giuseppe Saragat. The merger of these two parties was dissolved in 1969. The Socialist Party of Proletarian Unity, a left-wing group that had seceded from the Socialists in 1963, merged with the Communists in 1972.

The Liberal Party and Republican Party are the successors of two political movements that originally contributed to the unification of Italy. The Liberals stress the importance of personal initiative and freedom. The Republicans have attempted to show themselves as representative of the democratic left. A more particularist party was the Italian Democratic Party of Monarchist Unity, the vehicle of those seeking a return of the monarchy. It had some contact with the Italian Social Movement, a neo-Fascist party, appealing not only to the few who want a return of Fascism but to others who consider the government coalitions too left-wing. These two parties merged in 1972. The Radical Party campaigns on civil rights issues. Finally, the South Tirolean People's Party exists to unite the German-speaking population of that area and to gain greater political and administrative autonomy for the province of Bolzano within the region of Trentino-Alto Adige.

Trade unions. The constitution establishes the right to organize trade unions. The right to strike is guaranteed by the constitution and remains a very potent weapon in the hands of the trade unions. Unofficial and wildcat strikes also occur. Civil servants are covered by the general right to strike but the Constitutional Court has established that strikes by those engaged in fundamental public services are unlawful.

The participation of the citizen. The constitution seeks to establish the effective participation of all citizens in the political, economic, and social organization of the country. This, however, is a stated ideal rather than a binding obligation, and it has not and perhaps cannot be fully realized. In practice, except for the few who can become political commentators of some sort and so influence public opinion, the opportunities for participation are restricted to those connected with elections. All citizens of 21 years and over may vote in national, regional, and local elections. Partly because voting is considered a civic duty and partly because of spontaneous involvement, the turnout for elections in Italy is very high, reaching more than 90 percent of the electorate for parliamentary elections. Citizens may also subscribe to national referendums or petitions, the purpose of which is the abrogation of a law or an executive order; such a petition must be signed by 500,000 members of the electorate or sponsored by five regional councils. In certain circumstances national constitutional amendments are subject to a more ordinary form of referendum, in which the electorate vote in favour or against specific proposals. Abrogative referendums are provided for in relation to all regional legislation, and there is provision in some regions for holding ordinary referendums. The constitution also provides that 50,000 members of the electorate may jointly present to Parliament a draft bill.

Voting

On a more executive level, the right of workers to take part in the management of companies is guaranteed by the constitution but lacks enabling legislation. Participation in politics and public affairs through the medium of the press and radio and television is restricted to a relatively small number of individuals and to certain, mainly political, groups. The press is free but, because of high production costs, it is run either by public or private industrial groups or by political parties. The Radiotelevisione Italiana (RAI) is a state-run monopoly sanctioned by the Constitutional Court. It should provide equal facilities to all groups and individuals, but it is widely claimed that this impartiality has not been realized.

JUSTICE

The Italian judicial system consists of a series of courts and a body of career judges who are civil servants. Frequently, cases are heard by a collegial bench consisting of two or more judges, and the legal profession provides for interchangeability between the position of judge and of prosecuting attorney. The courts form either part of the regular court hierarchy or are special courts with a specific and limited competence. The judicial system is unified, and every court is part of the national network. The highest court in the regular hierarchy is the Court of Cassation; it has appellate jurisdiction and gives judgments only on points of law. The 1948 constitution prohibits special courts with the exception of administrative courts and courts martial, although a vast network of tax courts has survived from an earlier period. The administrative courts have two functions: the protection of *interessi legittimi,* individual interests strictly connected with public interests and protected only for that reason, and the supervision and control of public funds. Administrative courts are also provided by the judicial sections of the Council of State, the oldest juridical-administrative advisory organ of government. The Court of Accounts has both an administrative and a judicial function; the latter involves primarily fiscal affairs. The courts martial have criminal jurisdiction in cases involving military personnel on active service and even over reserve personnel on unlimited leave, with respect to certain military crimes. The Superior Council of the Judiciary, provided for by the constitution and intended to guarantee the independence of the judiciary, was only formed in 1958.

Italian law is codified and is fundamentally based on Roman law, in particular, as regards civil law. The codes of the Kingdom of Sardinia in civil and penal affairs, derived from the French Napoleonic model, were extended to the whole of Italy when unification was achieved in the mid-19th century. In the period between World Wars I and II, these codes were revised. The Constitutional Court has declared a number of articles unconstitutional. Besides the codes, there are innumerable statute laws that integrate the codes and regulate areas of law, such as public law, for which no codes exist.

The constitution stresses the principle that the judiciary should be independent of the legislature and the executive.

For this reason jurisdictional functions can be performed only by ordinary magistrates, and extraordinary tribunals may not be set up. Judges cannot be dismissed.

ARMED FORCES

The armed forces are commanded by the president of the republic, who also presides over the Supreme Council of Defense, comprising the president of the Council of Ministers, the ministers of Defense, Foreign Affairs, Industry, and the Treasury, and the chief of staff. Military service is obligatory. Italy's military expenditure is one of the highest in the world. Although the constitution specifies that the armed forces must embody the democratic spirit of the republic, their activity is free from any political control. Italy's adhesion in 1949 to the North Atlantic Treaty Organization transferred to the allied command a certain degree of control over the Italian forces.

Conscription

There are two police forces in Italy with general duties: the Pubblica Sicurezza, which is under the authority of the home secretary, and the Carabinieri, a corps of the armed forces that is, therefore, under the minister of Defense. The functions of the police are the prevention and suppression of crime; both functions are performed by both police forces. The administrative police to whom preventative duties are assigned see that the activities of individuals and of groups do not contravene the law. Their authority stems from a Fascist law of 1931 only partially modified by Constitutional Court decisions but reduced by a law of 1956 that requires that restraint be imposed on potential offenders by court order only. The administrative police are also responsible for the issue of passports and other permits. The constitution places the judicial police, who are engaged in suppressing crime, under the authority of the courts, but the actual subordination of the two forces to the two government ministries conflicts with their technical subordination to the judiciary. Besides these two police forces, there are also special police for customs, excise and revenue, and communal police and prison guards. There are also private police that operate in a limited field under the supervision of the regular police.

EDUCATION

The constitution guarantees the freedom of art, science, and teaching, the existence of private schools (mainly run by religious bodies) alongside the state schools, and the independence of the universities. It further states that the public schools are open to all and makes provision for scholarships and grants. Education is compulsory only from the ages of six to 14 years. The school system begins with kindergarten for the three- to six-year-olds. Elementary schools are attended between the ages of six and 11, at which stage most children go on to secondary schools for 11- to 14-year-olds, but those wishing to study music go directly to the conservatories. Postsecondary schooling is not compulsory and includes a wide range of technical and trade schools, art schools, teacher-training schools, and scientific and classical grammar schools. Pupils from these schools can then go on to university, where courses vary from four to six years.

Maintenance grants are few and inadequate, and the high cost of supporting children while they study, particularly at trade or grammar schools and at university level, when they could otherwise be earning, effectively limits higher education to a privileged elite.

HEALTH AND WELFARE

Health. The constitution guarantees the protection of health as an individual right and a community interest; the support of those who are unable to work and are indigent; and the right of workers to social-insurance benefits in the case of accident, illness, disablement, old age, or unemployment. A comprehensive national health service and national medical insurance were introduced in 1980. Special doctors in the provinces and communes are responsible for the public health of their respective areas. The Instituto Nationale della Previdenza (INPS) provides a system of social benefits, including unemployment, disability, retirement pensions, and family allowances. Industrial injuries protection is also state-provided.

Housing. A law was passed in 1971 to facilitate and speed up communal expropriation of sites suitable for subsidized housing and introduce advantageous terms for the leasehold purchase of the property. Until the early 1970s there was little low-cost housing, and purchase of property had been far beyond the means of the average family. Through the 1960s, many thousands were homeless and living in caves, cellars, shacks, and warehouses.

Zoning regulations for building are established by the communes under state supervision. Historic monuments and natural-beauty spots are protected by law and come under the control of the Ministry of Education; but, despite this supervision, public as well as private development has depredated cities, countryside, and coastline.

Social and economic divisions. The constitution asserts the right of every worker to an adequate wage and the right of equal pay and maternity benefits for women. However, Italy is a country of great social and economic differences. A very small elite enjoys great wealth, while the largest stratum of society lives in varying degrees of poverty. Inherited fortunes have survived, and industrial wealth is aggregated in economic empires as well as being concentrated geographically in the Milan–Turin–Genoa triangle. The economic underdevelopment of the south has been a chronic problem. The average annual family budgets give a very rough idea of the differences in spending capacity of the various categories of workers. The vast differences in the educational level of the population also testify to the cultural extremes prevalent in Italy and invite conjecture as to the underlying economic circumstances that may have caused them.

Social inequalities

Cultural life

CULTURAL MILIEU

The country now called Italy has resulted from the amalgamation of many small territories. Political unification took place in 1861, but unification is still incomplete as a cultural and social process. In the matter of language, however, Italian is the standard commonly used for official, formal, and literary purposes and taught in the schools to natives and foreigners. The varied dialects, including Tuscan—the language of Florence and its territory and used by Dante, Petrarch, and Boccaccio in their writings—are now usually spoken and only to a limited extent written. Regional differences in self-expression still persist, and Tuscan remains to some extent a linguistic invention in regions that are not Tuscany. The other dialects (as they are disparagingly called), including Gallo-Italian, Venetian, Corsican, Central Italian, and Southern Italian, embody ways of thought and speech that are not simply different modes of expression but represent many different ideological, psychological, and cultural worlds. All speech forms called Italian belong to the Romance languages.

Another sign of imperfect national unity is the regions' jealousy of one another. The division between the north and the south (the Mezzogiorno) is the sharpest of all. The prosperous, progressive, industrial society of the north looks down on the backward, impoverished, primitively agricultural Mezzogiorno regions south of Naples in Puglia, Basilicata, Lucania, and Calabria. To many northerners the increasing efforts of government to assist the economy of the south are so much money poured down the drain in a hopeless cause. For his part, the southern day labourer remains suspicious of the central government despite the aid it offers the south through the introduction of civil-works projects and of industry bringing alternative employment. The administration has been indifferent to his desperate needs too long and has tacitly left him to his master, the absentee proprietor of the vast farm, or *latifundium,* on which he and his companions labour. In western Sicily this resentment of central authority is carried so far as to secure general acquiescence in the activities of the Mafia, underlining the fact that the Mafia indeed was once a secret society pledged to wage a perpetual struggle against the oppression of foreign rulers.

North versus south

The problem of regionalism, seen most clearly in the language question and the division between the north and the south, is basic to the Italian scene. But with the country's industrial development and its liberation after World War II from the political stultification of Fascism, a ferment of change and uncertainty has transformed society, especially in the large urban centres. The capture of foreign markets by the Italian automobile and gasoline industries and by agricultural cooperatives has strengthened Italy's economic links with other countries. Great impetus has been given to this process by the country's membership in the Common Market.

At the social level the transformation has been encouraged by the growth of the communications media, especially television. The result is that the old provincialism, wherein each city seemed to have its own life-style, is disappearing. The position of the sexes, too, is changing with regard to each other. The position of woman, politically enfranchised in the 20th century, marks an advance toward freedom. Continuing as a wage or salary earner even after marriage, she is no longer destined from an early age to household matriarchy, exerting her influence on the family from within it in counterpoise to her husband's more explicit rights. Men and women are less oppositely polarized in their confrontation with each other: to a novel extent they have become rivals in the same sphere of action.

The authority of the Roman Catholic Church may have been discreetly ignored in the past, but it was seldom openly challenged, other than by a minority of professed anticlericals. Now a more general disregard of the church and its system is evident in the extent to which life has become secularized and in the decline in church attendance. The struggle with the church over divorce—previously unobtainable in Italy under the terms of the state's Lateran Treaty (1929)—was a momentous one, rousing the devout and the diehard to an impassioned stand against it, before the divorce law was introduced in 1970. In the following year, the government introduced measures offering new possibilities in the way of birth control.

The *dolce vita*

In Rome especially, a feature of the generally prosperous, if erratic, postwar era has been the *dolce vita* ("sweet life") attitude of internationalized café society. The *dolce vita* was a reaction of release from the conventional family code of Italian life. Its philosophy was one of experiment in an Existentialist moral incertitude, rather than simple hedonism. The drifting, day-to-day existence of the practitioners of this way of life—the *pappagalli* ("parrots"), press and street photographers and jacks-of-all-trades pestering film stars and pretty girls on Rome's Via Veneto for preference—was symbolic of the *dolce vita* and its ephemeral futility. The *vitelloni* ("big calves"), young loungers of no occupation, have been another symptom of this unsettled period.

Italian culture has felt the impact of avant-garde modes of feeling and expression but in a rarefied way among few people. Traditional culture, of "aristocratic" and "educated" derivation, is disdained, ridiculed, and opposed by forward-looking intellectual elements who are aware of transatlantic developments in the arts and who claim to represent the advanced contemporary art situation in Italy. Yet the populace prefers the traditional culture and ignores the products of the cultural elite. The resulting split between the collective psychology and intellectual culture continues. The people remain unmoved before the constructions of the laboratories of the intellectual "upper classes," which are an exorcism of past culture and a rupture of continuity with its traditions. At the opposite pole, the "upper classes," having lost confidence in themselves in isolation, have abandoned their dream of universal interpretation and instead attempt to contribute to culture by elaborating intrinsically empty constructions inspired by a preoccupation with technique; the technique itself is regarded as a method of seeking truth and interpreting reality but is no more than a stratagem for filling an unexpected vacuum in ideology.

CONTRIBUTIONS OF THE ARTS

Italy was in the forefront of the development of the arts in the Renaissance, that crucial and brilliant period of transition when European culture emerged from the Middle Ages and entered into the modern age. This rebirth received impetus from a reappraisal of the classical Greek and Roman world after centuries of supposed ecclesiastical obscurantism. Artists and scholars in Italy were especially well placed to take the lead in such a revival since they lived in what was the very heartland of the ancient Roman Empire, and the material remains of its civilization, whether as stone structures or as texts, lay beside them.

Literature. Through Dante, Petrarch, and Boccaccio, Italian literature blossomed to supreme greatness early; after Ariosto and Tasso in the later Renaissance it declined somewhat into formalism but renewed some of its fire through the Romantics, Vittorio Alfieri, Ugo Foscolo, Alessandro Manzoni, and Giacomo Leopardi; and then, in the 20th century, Gabriele D'Annunzio represented the last flowering of Romanticism. Since then, the writers who have made the most significant contribution to their country's literature have turned their backs on the rhetorical Romantic tradition. The plays of Luigi Pirandello (who died in 1936), inventive, psychologically disturbing, and impassive in their mood, and his terse, hard-edged short stories heralded a new attitude and a new technique; his masterpiece, the play *Henry IV,* with its thrillingly presented intersecting planes of normality and madness, comparable to the analysis made of pictorial reality by Picasso's Cubism, is as fresh as anything written since by the French Existentialists Camus and Sartre. This new mood, this dry tone, is realized also in the work of the novelists of Realism, or Verismo, who have been the dominant figures in modern Italian literature: Verismo was initiated by Giovanni Verga; after him Italo Svevo (Ettore Schmitz), disregarded by all except James Joyce and Eugenio Montale until the end of his life because the Italian literati were averse to his unadorned style, is now seen as one of the ironists of modern European literature.

The Realist novelists

Since Svevo's death in 1928, the novelists Ignazio Silone, Cesare Pavese, and Alberto Moravia and the poets Montale and Giuseppe Ungaretti have been outstanding figures. It is in the work of the Realist and Neo-Realist novelists that the attempt of Italian writing to come to terms with the modern world and its political and social pressures can be best appreciated. The works of Pavese and Moravia, the first with his acknowledged debt to the economical, crisp style of Ernest Hemingway and the other with his remorseless analysis from the standpoint of a European Existentialist intellectual, are far removed from the rhetoric and Latinist abstraction of the traditional plane of ideas of Italian letters.

The visual arts. The great names in Italian art and architecture through the centuries make a long catalog: those of Giotto, Donatello, Brunelleschi, Michelangelo, Leonardo da Vinci, Titian, Bernini, and Tiepolo call up a host of others. But continuous subjection to foreign powers had an enfeebling effect on Italy's artistic contribution, which sank into provincialism. Ties with European art were renewed about 1910 by the work of the painter Amedeo Modigliani and by the Futurist movement, which found its most characteristic expression of mechanistic dynamism in the work of its leader, the poet Filippo Marinetti (died 1944), and the painters Umberto Boccioni and Giacomo Balla. Futurism was succeeded by the "metaphysical painting" of Giorgio De Chirico, at one time associated with the Surrealists for his timeless dream landscapes until he turned his back on this early work to produce, from the 1950s on, canvases loaded with reminiscence of traditional styles. Giorgio Morandi's subtle, quietist paintings of endlessly varied arrangements of bottles, pans, and jars are a product of metaphysical painting and, since his death in 1964, have become perhaps more highly regarded than the work of any other contemporary Italian painter. Lucio Fontana's work exemplifies the modern artist's solitary quest for form: blank canvas opened by a knife slash; an arrangement of pebble grains stuck onto the unicoloured canvas; and a room swathed in nylon textile (at the Palazzo Grassi in Venice in 1960).

The Rational Architecture movement of 1927 has produced one of the outstanding Italian architect-engineers of the 20th century in Pier Luigi Nervi, architect of the Turin exhibition complex and, with Marcel Breuer and Bernard Zehrfuss, of the UNESCO headquarters in Paris. Innovative

Italian architect-engineers

educational architecture is represented in Milan's Istituto Marchiondi by Vittoriano Viganò. The work of Nervi and of such other architect-engineers as Giovanni Ponti (who worked with Nervi on the Pirelli skyscraper in Milan) and the builders of hydroelectric dams in Africa and elsewhere represents their country's most serious contribution to modern art in hauntingly beautiful constructional work of which Brunelleschi himself might have been proud. Other movements have come and gone, that of the Six in Turin, the Roman school, and the school of Milanese Expressionism, to which the sculptor Giacomo Manzù once belonged.

Music. Italian music has been one of the supreme expressions of that art in Europe: Gregorian chant, troubadour song, the madrigal, the work of Palestrina and Monteverdi and of composers such as Vivaldi, Alessandro and Domenico Scarlatti, and Cimarosa, followed by the 19th-century flowering of Italian opera in the hands of Rossini, Donizetti, Bellini, and, greatest of all, Giuseppe Verdi. Arrigo Boito and Giacomo Puccini garnered the Verdian heritage, and then Verismo, or Realism, made itself felt in operatic tradition as in literature in the work of Pietro Mascagni and others. Since World War II, in the post-Schoenberg world of serial music, two Italians have made significant contributions: Luigi Dallapiccola and Luigi Nono.

Theatre. Theatrical production in Italy in the latter half of the 20th century has employed all forms of the art of theatre, from grand opera to the puppet show. Opera productions, notably at La Scala opera house in Milan, as well as at other opera houses such as the San Carlo in Naples and at the Teatro la Fenice in Venice, are world famous; and an annual summer production of an opera in the Roman amphitheatre in Verona also attracts foreign visitors. Modern operas by Italian composers that have been staged include *Il convento veneziano* by Alfredo Casella and *Sette canzoni* by Gian Francesco Malipiero.

In the drama the Italian theatre has been active in producing outstanding contemporary European work and in staging important revivals. Not a great deal of major new work has been offered to it by native playwrights: nothing to rival the work of Luigi Pirandello earlier in the century. Outstanding productions have included those of the company of the Teatro Stabile della Città di Genova and of the Piccolo Teatro of Milan. Leading producers, working with various companies, have been Giorgio Albertazzi (who caused a stir in the cinema with his film of Moravia's novel *Il conformista* ["The Conformist"] in 1971), Gianfranco De Bosio, Giuseppe Patroni-Griffi, Giorgio di Lullo, Luigi Squarzina, and Giorgio Strehler. New plays of the period include Primo Levi's documentary dramatization of his experiences in a Nazi death camp, *If This Be a Man;* Griffi's Pirandellian comedy *Imagine, One Evening at Dinner;* Franco Brusati's *La pietà di novembre* (*The Other Face of November*); and Moravia's symbolic anti-Nazi drama *Il dio Kurt* ("God Kurt").

Italy exercises a notable influence throughout Europe in the field of ballet, and contemporary Italian ballets include *Balli plastici* by Fortunato Depero, *Coro di morti* and *La follia di Orlando* by Goffredo Petrassi, and *Marsia* by Luigi Dallapiccola.

Motion pictures. It is in the cinema that Italy has probably made its most significant contribution to contemporary art on an international scale. Before World War II the Italian film industry had produced epic films that were by no means negligible in quality. But just after World War II Italy caught the world's eye, first with the Neo-Realism of the films of Roberto Rossellini and Vittorio De Sica in particular, dealing in a matter-of-fact way with conditions in Italy at the end of and just after that war, and later with the more freely imaginative interpretation of Realism exemplified in the films of such directors as Michelangelo Antonioni, Federico Fellini, Cesare Zavattini, Pier Paolo Pasolini, and Luchino Visconti. The trenchantly laconic statement of many of these films and their affinity with the attitudes of Existentialist thinking marked a development in cinematic imagination that had a cross-fertilizing influence, in particular, on the young French filmmakers of the "New Wave."

Neo-Realism in films

CULTURAL INSTITUTIONS

Academies and societies. Academies and societies representative of almost every academic and social activity have proliferated in Italy as nowhere else. Literary art and academies flourish in the major Italian cities that are regional capitals. Indeed, academies of fine arts had their origins in Italy, the Accademia di Belle Arti of Florence (founded as the Accademia del Disegno) in 1563 and that of Perugia in 1573. Rome's Accademia di San Luca was a guild of painters, founded in 1577; today its collections are open to the public. Italy's most famous learned society is the Accademia Nazionale dei Lincei, of which Galileo was once a member. The most distinguished literary society is the Accademia della Crusca, founded in Florence in 1582, whose *Vocabolario della Crusca* stabilized the Italian literary language on the basis of Tuscan speech. There are likewise many historical and scientific societies including the Accademia del Cimento, established (1657) in Florence.

A feature of Italian academic life is the contribution made by the foreign schools maintained in Rome by the United States, France, West Germany, Great Britain, and others for the study of Italian architecture, art, and archaeology.

The foreign schools in Rome

Among cultural institutions, the society Italia Nostra, with membership open to the general public, holds a special place. Its importance lies in the work it does to call attention to the care of the country's architectural treasures and to the preservation of what is left of the beauty of its landscape, rural and urban, from the encroachments of industry and bad building and the effects of environmental pollution. Italians traditionally have been indifferent toward matters not immediately affecting them and are disinclined to bestir themselves in public causes. The efforts of Italia Nostra have made them aware of the penalty of such indifference. The society has encouraged coordinated study and planning by industry, local authority, and government to resolve the conflicting needs of conservation and of industry—and the provision of work is vital in a country where underemployment is chronic.

Galleries and museums. Italy is exceptionally rich in architectural monuments, in art galleries and museums, and in examples of architecture of the past still in use, as well as of ancient ruins; indeed, churches, palaces, villas, and other buildings that are works of art in their own right often also contain art treasures in the shape of frescoes on their walls, easel pictures and sculpture, and furniture and ornaments. The regionalism of Italy is typified by its art galleries and museums. The national galleries at Florence and Bologna are also municipal institutions, and indeed the great galleries of Italy are often concerned with their own regional heritage—the Capitoline Museum and the Borghese Gallery in Rome are mainly built up of the work of painters and sculptors working in Rome, even if they were not all strictly of the Roman school; the Pinacoteca di Brera in Milan has the most representative collection of north Italian painting of the Lombard school; the Accademia in Venice of Venetian painting; the Uffizi and the Palazzo Pitti galleries in Florence are supreme for Florentine painting; the Galleria Nazionale dell'Umbria in Perugia contains magnificent examples of the Umbrian school; the Pinacoteca Nazionale in Siena of the Sienese school; the Palazzo Bianco in Genoa of the Genoese school; and so on.

It must be understood, of course, that the galleries also contain masterpieces of other Italian schools and indeed from foreign countries, but their collections are built up around the body of regional work exhibited. Other major galleries and museums besides those mentioned above include the Vatican and Lateran museums and the Galleria Nazionale d'Arte Moderna in Rome; the Museo Nazionale del Bargello and the Museo dell'Opera del Duomo in Florence; the Galleria d'Arte Moderna, the Castello Sforzesco, and the Pinacoteca Ambrosiana in Milan; the Galleria e Museo Estense in Modena; the Museo Archelogico Nazionale, the Museo Civico Filangieri, the Museo Principe Diego Arangona Pignatelli Cortes, and the Museo e Gallerie Nazionale di Capodimonte in Naples; the Galleria Nazionale in Palermo; the Palazzo Doges and Museo Civico Correr in Venice.

ENTERTAINMENT

Festivals. Regional life in Italy is typified by diversity of costume and cuisine and by a great variety of festivals. The latter are, however, changing in character, and many civil and religious festivals are no longer forceful expressions of local idiosyncrasies that brought them into being; as in most other Western countries, an element of unreality enters the dressing up in clothes belonging to bygone days and only brought out once or twice a year. The appeal to the tourist industry helps as much as anything to keep festivals alive.

The Italians are a lively Mediterranean people, and the climate of their country is favourable to social exchange and display out of doors. The *passeggiata,* or "promenade," conducted with conversation and gossip up and down the main street or about the principal square at noon and evening, is still a feature of urban provincial life. In the same spirit of outward projection, Italians enjoy festivals and processions. Festivals in Italy are indeed manifold and can be divided into two main kinds, religious and secular, though the religious observations generally extend their impulse to cover a good deal of accompanying secular celebration. The secular festivals and sometimes the religious festivals contain strong elements of folklore.

Religious festivals. Many places hold special religious festivals and great processions in costume on such holy days as Easter, Corpus Christi, and the Feast of the Assumption. At Christmas the crib (*presepio*) is set up in churches, and children receive their presents at Epiphany, the feast of the three kings, from the fairy-witch Befana. The New Year is celebrated with fireworks and noise, supposedly to aid in driving away the devil. At Epiphany, in Rome, it is the custom to bring presents for children in the Piazza Navona, and at the Feast of St. Anthony there, on January 17, priests of the Church of St. Eusebio bless working animals and pets paraded by their owners.

In Palermo the Festival of St. Rosalia is held in July, with a procession and illuminations. In Naples the Miracle of St. Januarius, a procession of the vial containing that saint's dried blood, which is said to become liquefied again upon invocation, is held at special times. Venice holds two festivals, in memory of the cessation of the plagues of 1576 and 1630, respectively: the first festival is the Feast of the Redeemer, held on the third Sunday in July, when a bridge of boats is built across the wide channel of the Canale della Giudecca, a solemn high mass is held in the church of the Redentore, and a breathtaking fireworks display takes place at night; the second is held on November 21, at the church of Sta. Maria della Salute. At Bari the feast of the patron saint of the city and of sailors, St. Nicholas, is held on May 8 and is attended by many pilgrims who take part in an illuminated procession in boats in the harbour.

Secular festivals. Secular festivals take a number of forms, including modern arts and crafts festivals, and, as a general rule, the more local they are, the more vitality they have. At Bari the Levant Fair, which lasts for two weeks or more in September and in conjunction with which a motor show is held, is an important regional fair. Earlier, in May, the city holds a famous procession, the Vidua Vidua, and a folklore festival, while Foggia, not far away, conducts a flower show and fair with outdoor opera. Still in the same region, Brindisi holds a parade of horses and riders in medieval costume. In Naples the Festival of Piedigrotta commemorates the Battle of Velletri (1744), and people go to the Grotta Nuova to hear new songs sung for the event. At Cocullo, in the Abruzzi, on the first Thursday in May, St. Domenic's statue is carried in procession, live poisonous snakes writhing round it to be made harmless and later bought by apothecaries for the medicinal properties of their venom. One of the best known to foreigners is the Corso del Palio (Parade of the Banner), held in Siena on July 2 and August 16; it is a parade and horse race around the main square, which is said to go back to 1275, with the riders in medieval costume of the colours of their respective city districts.

The Siena Palio

Siena, Mantua, and Spoleto hold notable music festivals. Still in central Italy, Arezzo holds its Giostra del Saraceno, a tournament originating in the 13th century, in June. On the first Sunday in May and on June 24, Florence holds its Calcio, a kind of football match dating back to 1530, in costume. At Pescara, on the Adriatic, an international folklore festival is held. Varese, in Lombardy, holds a national festival of mountain singing in December. In Piedmont, Aosta holds a Battle of Queens in October. And the Ligurian riviera resorts enjoy celebrated Battle of Flowers festivals. Famous Venetian festivals include La Sensa (on Ascension Day), the Regata Storica (regatta) on the Grand Canal, and festivals of drama, music, and films.

Sports. Cycling, football, basketball, tennis, motoring, motorcycling, winter sports, and hunting are popular sports and recreations in Italy. The major sporting events are the professional cycle races, chief among them the Tour of Italy road race, which attracts the best foreign as well as Italian professional road-racing cyclists, and the major professional football matches. In football every sizable town supports its own professional team; teams from Milan and Turin play before capacity crowds in their stadiums and have provided much of the national talent that took Italy into the final of the World Cup in 1970.

Other major sporting events include the Italian Grand Prix motor race at Monza, the international Italian Tennis Championship at Rome, the Martini Fencing Trophy at Turin, and the Show Jumping Championships at Rome.

Tourism. For centuries foreigners have been attracted to Italy by its varied architectural monuments, scenery, and climate; Rome, the "Eternal City," has drawn visitors to it especially for its classical antiquities and as an early centre of Christianity and the seat of the head of the Roman Catholic Church. In the 18th century it became a custom for English gentlemen or for their sons in the company of a tutor to make the Grand Tour, an educative tour of western Europe in which the visit to Italy was the highlight. In the 19th century a number of English literary figures chose to live in Italy for a time, and their example led to the growth of little colonies of expatriates, principally in Florence and Rome, who were pleased to receive visiting fellow countrymen.

The Grand Tour

By the 20th century the pleasure of a trip to Italy ceased to be reserved for a well-to-do, cultivated minority when the rise of the tourist industry in the hands of experienced travel agents removed much of the expense and even peril of travel in Italy and made it possible for thousands to enjoy it. Since World War II, flies and mosquitos, formerly two of the greatest enemies of the traveller in Italy, have almost been eliminated. But from the 1960s onward the Italian tourist industry has also felt competition from such countries as Spain, Portugal, and Yugoslavia, where the cost to the tourist has been lower than in Italy.

Foreign visitors once sought the great cultural centres of Rome, Florence, Venice, and Naples, but many now spend time at coastal resorts and islands or among the Alpine hills and lakes of the north: the Ligurian and Amalfi rivieras; the northern Adriatic coast; the small islands in the Tyrrhenian Sea (Elba, Capri, Ischia, Ponza, Lipari, Stromboli); the "Emerald Coast" of Sardinia; Sicily, especially the resort of Taormina; the National Park of the Gran Paradiso and the Dolomites in the Western and Eastern Alps, respectively; the north Italian lakes (especially Maggiore, Como, and Garda); and the National Park of the Abruzzi, easily reached by a highway from Rome.

Most Italians take their holidays in their own country and after much the same pattern as that of foreign visitors, with the addition of much visiting of near relatives; but they make more use of the southern Adriatic beaches in Puglia, at Manfredonia, Siponto, and the little villages around the foot of the Promontorio del Gargano (Gargano Promontory) and down the same coast at Trani, Bari, and Brindisi. They have also been able to keep the Calabrian beaches much to themselves, at Crotone and Reggio Calabria, for example, and up the Tyrrhenian coast to Paestum and Salerno. But all these places are being increasingly frequented by foreign visitors.

PRESS AND BROADCASTING

Press. The Italian press is generally provincial and local in outlook, though the major daily newspapers carry foreign news and comment on it. The dailies that command most notice abroad have included the *Corriere della sera* and *Il Giorno* of Milan, *La Stampa* of Turin, and *Il Messaggero* and *Il Tempo* of Rome. Political parties publish or control newspapers, and daily newspapers are issued in the large regional capitals.

Broadcasting. Radiotelevisione Italiana (RAI) conducts all broadcasting and is a state enterprise. Programming includes current news, culture, sports, light music, debates, interviews with people prominent in contemporary Italian life, classical music, and jazz.

Television programming includes variety shows, film cycles, quiz programs, musical games, athletic events, news broadcasts (*telegiornale*), programs for children, political debates, investigations in depth into contemporary problems (*boomerang*), theatre, etc. Dramatized novels, foreign-language courses, and football matches are also broadcast. For statistical data, see the "Britannica World Data" section in the current *Britannica Book of the Year.*

(Ed.)

HISTORY

Italy in the early Middle Ages

THE BARBARIAN INVASIONS

Italy in the 5th century. Toward the middle of the 5th century, Italy was the only province of the Western Empire in which Germanic barbarian peoples had not established permanent occupation. Imperial dignity retained considerable prestige, and the new capital, Ravenna, chosen for its easily defensible position and its lines of communication by sea to the East, was enriching itself with splendid monuments.

Although no barbarian successor state had been established in Italy, groups and individuals of barbarian origin had acquired great importance in the political and social life of the peninsula. Barbarians made careers for themselves in the army, and some of them attained positions of great power, married into the imperial houses, and even deposed and created emperors. Such a barbarian was Ricimer, *magister militum* ("master of the soldiers") of the Western emperor Avitus. As a barbarian and an Arian, Ricimer could not be emperor, but his strength in Italy was such that he deposed three emperors in the mid-5th century. Ricimer was succeeded briefly by the Burgundian Gundobad, who in 473 placed Glycerius on the throne. Two years later, contingents of the barbarian tribes of the Scyri, Heruli, and Rugii chose one Odoacer as their leader. In 476 the last Roman emperor, young Romulus Augustulus, was deposed, bringing the Western Empire to an end. Odoacer ruled over Italy in the double capacity of king of the barbarians and of *patricius*—that is, as an unrecognized representative of the Eastern ruler.

Overthrow of the Western Empire

The accomplishment of these barbarians cannot be dismissed as a purely negative one. However crude and brutal they may have been, they made a place for themselves in the Roman world, which they defended to the best of their abilities. Ricimer led the long struggle against the Vandals, a Germanic tribe that swept through Spain and North Africa, led a seaborne expedition to Italy, and in 455 sacked Rome. Odoacer, in his turn, compelled the Vandal king Gaiseric to give back Sicily and succeeded in occupying Dalmatia. In short, these barbarian generals and their rough-and-ready soldiers, although they oppressed local populations, had a role to play as defenders of the empire, and it would be a mistake to blame them for the decay of Roman civilization. Indeed, the Romans in Italy had been accustomed for centuries to contact and cohabitation with the barbarians. Between the last half of the 4th and the first half of the 5th century, many barbarian prisoners were farmed out to the countrysides and cities of the north in order to repopulate them. These Goths, Huns, Alemanni, and others have left their traces in place-names throughout the region.

The reign of the Ostrogoths. In 488 Italy was invaded from the east by a new barbarian army, that of the Ostrogoths, which, after a succession of victories climaxed by the siege of Ravenna in 493, destroyed Odoacer's feeble regime. The invaders this time were not merely armed groups; they comprised an entire population (numbering perhaps some 300,000) that had left the Balkans with the firm intention of settling in Italy. The Eastern emperor Zeno had encouraged their migration because he was dissatisfied with Odoacer's rule in Italy and was also anxious to remove the Ostrogoths from the Byzantine frontiers. The Ostrogoth leader was the able Theodoric, who had lived for a long time as a hostage in the Eastern court, where he grew to appreciate Roman-Byzantine civilization.

Theodoric ruled over Italy both as king of his own people and as *magister militum* of the Eastern emperor. The main feature of Theodoric's policy was to keep Ostrogoths and Romans apart, allotting to the former the exercise of arms and to the latter posts in the civil government. This was, he thought, the only way for the Ostrogoths to keep the upper hand, since their numbers were few in comparison with those of the local population. This separatist policy was based on a difference of religion, since the Ostrogoths were Arians and the Romans were Catholics; the policy included the prohibition of *connubium,* or mixed marriages. But it was, fundamentally, contrary to historical reason, since it was inevitable that, in time, the two peoples should be drawn closer together.

There were signs of crisis in 519, when Theodoric suspected Zeno of plotting against him. He attacked the Roman leaders who had lent him their support. The senators Albinus, Boethius, and Symmachus were tried and condemned to death in 524–525; Pope John I was arrested and died in prison in 526.

When Theodoric died in 526, his Italian policy had to be considered a failure. But, in any event, Italy and Sicily (Sardinia and Corsica were still under Vandal rule) had enjoyed a period of tranquillity and well-being under his government. There had been a renaissance of Classical culture, as witnessed by the writers Boethius and Cassiodorus, who lived at Theodoric's court. And the capital cities—Pavia, Verona, and, above all, Ravenna—were adorned with splendid buildings and monuments.

Reconquest by Byzantium. Theodoric left the kingdom in the hands of his daughter Amalasuntha, who ruled as regent for her son Athalaric. When Athalaric died prematurely in 534, she had to share the throne with her cousin Theodahad, head of the Goth nationalist faction, which within a year found a way of getting rid of her. This assassination furnished a pretext for Justinian I, the Eastern emperor, to continue the reconquest of the lands along the western Mediterranean, already begun with general Belisarius' victorious expedition against the Vandals of Africa (533–534). Belisarius then launched a campaign in Italy, disembarking in Sicily in 535. The war went first one way and then the other; it was long lasting (until 553) and exhausting and destructive for both armies and civilian populations. In spite of their stubborn resistance, the successive Ostrogoth kings—Witigis, Totila, and Teias—could not prevent the occupation carried out by the troops of Belisarius or his successor Narses. The Eastern Empire retook Sicily, Sardinia, and Corsica. The Ostrogoths were killed, taken prisoner, or dispersed, and, of their stay in Italy, very few traces remain.

Extermination of the Ostrogoths

The whole of Italy returned to direct dependence on the empire, though under quite different conditions than before; it was no longer a centre of power and privilege but a mere outlying province. With the promulgation of the Pragmatic Sanction in 554, Justinian gave Italy a new ordering, creating a *praefectura Italiae* (with its capital at Ravenna), subdivided into 11 provinces in which there was a clear-cut separation between civilian and military authority. Sicily was governed directly from Constantinople, while Sardinia and Corsica were lumped into the Exarchate of Africa. This was the beginning of a long period of Byzantine influence in Italy, not only political but cultural and artistic as well. Roman law, revived and codified under Justinian, came into widespread use; the magnificent basilicas were built in Ravenna; and Byzantine architecture and mosaics adorned Rome and other places. But the Byzantine government, avid for taxes and heedless of its subjects' religious convictions, provided considerable opposition, of which the popes later took advantage in order to bolster their authority. The popes also could count on the support of increasingly numerous monasteries. Saint Benedict of Nursia (Norcia) had laid out a model of monastic organization and rules during the tragic epoch of the Greco-Gothic War, and the monasteries later became strongholds of the Catholic religion and the Latin tradition.

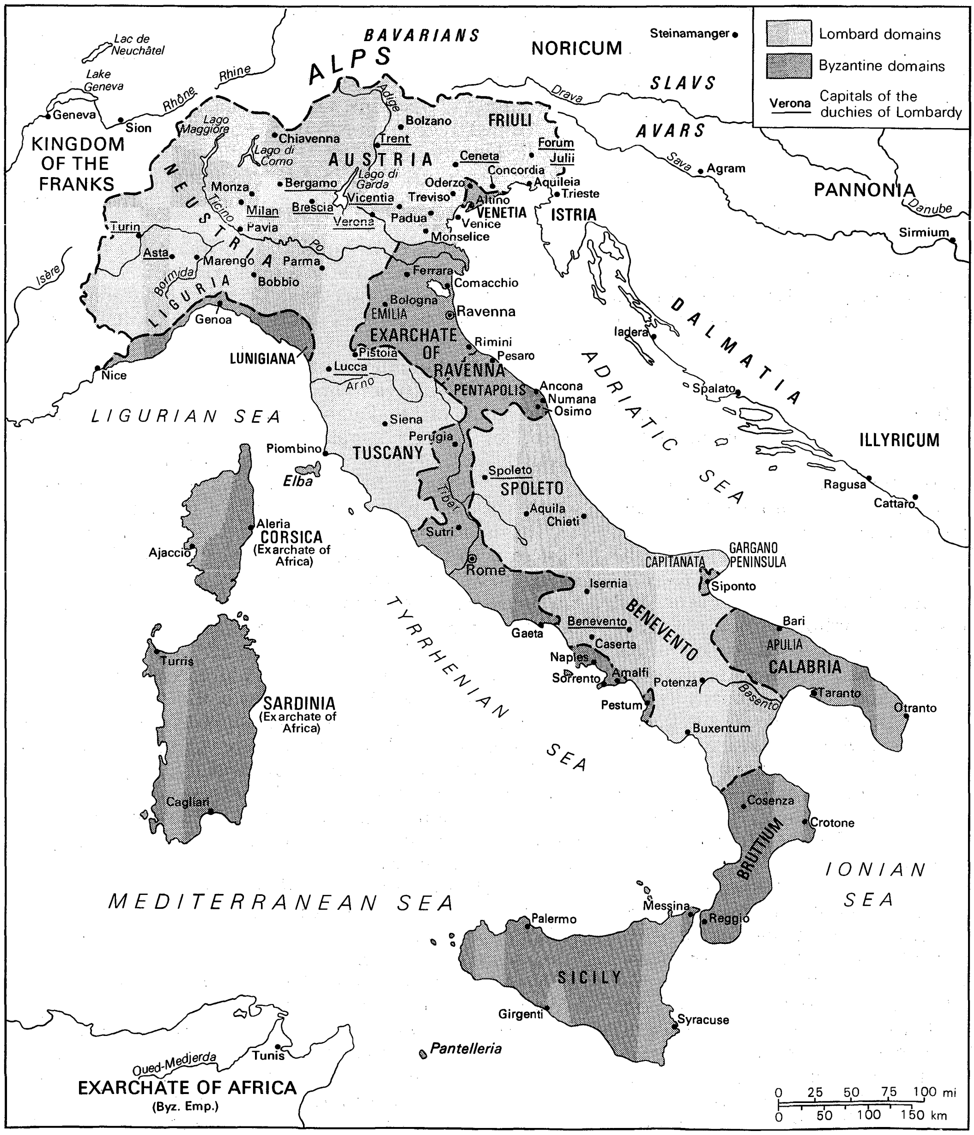
THE LOMBARDS

The conquest and the political structure. In 568 a new Germanic people, the Lombards, appeared at the eastern gateways of Italy, coming from Pannonia and Noricum. Their numbers were about equal to those of the Ostrogoths, but they were considerably cruder, had no links with the Byzantine Empire, and looked on Italy as a land to be conquered. The Byzantines, whose Italian forces were meagre, put up their first resistance at Pavia (Ticinum), which fell to the Lombard king Alboin after a three-year siege. The clergy and population of many places (Aquileia, Milan, and elsewhere) fled before the Lombard invasion to inaccessible coastal areas, where they could count on the protection of the Byzantines, who still controlled the seas. The Lombards soon occupied the whole interior of the peninsula as far south as Benevento. Their administrative units were called "duchies," a name reflecting the fact that their army units were led by *duces.* There were 35 such duchies, named after their capital cities. Among the most important of these were Forum Julii (the present-day Cividale del Friuli), Brescia, Pavia, Pistoia, Lucca, Spoleto, and Benevento. The occupied territory as a whole took the name of Longobardia or Langobardia (Lombardy). Later on, this appellation was restricted to the central and northern region, seat of the capital, just as it came to be the case with the Byzantine territories, collectively known as Romania, a name that subsequently came to apply only to the Romagna, with its capital at Ravenna.

The Lombard duchies

Within each duchy the land was divided for ownership or tax purposes into *farae,* groupings of related families that made up the social and military fabric of the Lombard people. About the administrative structure little information remains, most of it being from an account by the 9th-century historian Paul the Deacon. It is certain, however, that the class of the Roman *possessores* was broken up and practically destroyed. This was true particularly during the Interregnum (574–584), when the initial unity of the Lombards was attenuated and the "dukes" ruled independently, some of them even passing over to the service of the empire. The dukes, however, soon realized the danger of such a state of anarchy in the face of the hostile

Adapted from *Enciclopedia Italiana di Scienze, Lettere ed Arti,* vol. 19



Italy under the Lombards and the Byzantine Empire, *c.* 603.

Byzantines and also of the Franks, who were pressing at the northern borders, and they restored the kingdom with Authari as their king (584–590). The king was transformed from a military leader into a regular monarch, as half of the occupied lands were incorporated into a royal domain.

The Lombards introduced radical change in the peninsula. Whereas the Goths and other barbarians had respected the Roman political and administrative framework, the Lombards did away with it completely in favour of their own customs. The only institutions to be saved were the churches. The king promulgated laws and pronounced judgments together with the assembly of the *arimanni* or *exercitales,* composed of all the freemen able to bear arms. Such, at least, was the theory, but, since the Lombard *farae* were widely dispersed, in practice the assembly of the *arimanni*—the *gairethinx*—was soon replaced by the *gasindi,* a group of councillors close to the king and of such powerful nobles (*adalingi*) as were able to make prolonged stays at the court. The court, housed in the royal palace of Pavia, grew in importance with the concentration of government power and the differentiation of offices; within it there was even a remarkable school of law. The conquered territory was ruled locally by dukes and *gastaldi,* the former as heads of family lines, the latter as royal officials. The *gastaldi* were originally in charge of crown possessions, but their authority grew, to the detriment of that of the dukes. In some cases, *gastaldi* took the place of rebellious dukes; in others they ruled a provincial city from the start. In the reign of Rothari (636–652), it seems that there were, in the cities, *gastaldi* appointed by the king to watch over the dukes, and there were rural military–judicial units where *gastaldi* ruled. At the time of King Liudprand (712–744), when the government was further centralized, newly conquered lands were made into *gastaldatus,* directly dependent upon the capital at Pavia. The duchies of Spoleto and Benevento, on the other hand, had a development of their own, and the *gastaldi* were appointed by the dukes.

The reign of King Agilulf

Authari tried, by both arms and treaties, to hold back the Franks and the Byzantines; in order to combat the former, he allied himself to Garibald, duke of Bavaria, by marrying his daughter Theodelinda. She was a Catholic, and it was on her initiative, with that of Pope Gregory the Great, that there was a first move to convert the Lombards, who were Arians or pagans. After the death of Authari (590), Theodelinda married Agilulf, duke of Turin, who held the throne from 591 to 615 and completed the occupation of the hinterland of Venetia. Together with the dukes of Spoleto and Benevento, Agilulf led expeditions to central and southern Italy and intended to occupy Rome until the strong-minded Pope Gregory dissuaded him. Aware of the spiritual power of the church, Agilulf adopted a pro-Catholic policy; he allowed Adaloald, his son by Theodelinda, to receive a Catholic baptism and favoured the Irish monk Columban, who in 612 founded the monastery of Bobbio, near the lines of communication between the Po Valley and the Byzantine territories of Lunigiana and Liguria.

Rothari, an Arian elected in the wave of a reaction to Agilulf's indulgence toward Catholicism, extended the Lombard kingdom to its greatest territorial extent, with the conquest of the Ligurian coast and of Oderzo in Venetia. He is famous for his Edict of 643, the first codification of Lombard customs. The edict, written in Latin, shows some influence of the secular life of the Romans and of the church. There is an effort to contain the *faida*—that is, personal revenge—by means of the *guidrigild,* an objective and closely calculated compensation for damage done, which reflects the social rank of the damaged party and, by extension, the whole structure of this barbarian society, divided into *arimanni, aldii* (semi-freed men), and slaves.

After Rothari there was a long period of contested successions, of rebellions, and of struggles among the dukes until the election in 712 of Liudprand, the greatest of the Lombard kings. In this period, beneath the violence of political events, a silent and deep transformation was taking place. Conversion to Catholicism was becoming widespread. With conversion came a new closeness, indeed the beginning of a fusion, between the Lombards and the native Romans. The two peoples were drawn gradually together by everyday life, by participation in the same liturgy, by common hostility toward Byzantium and mistrust of Rome, by the use of a common Latin language, and by devotion to the monarchy. Although certain social and juridical distinctions were slow to disappear, they increasingly lost their original ethnic imprint. The consequences of this transformation can be seen in the structure of the state and its political aims. Liudprand founded and protected churches and monasteries, based his laws on religious principles, contributed to the struggle against the Arabs of Provence (in southern France), and gave the bishops an important share in public life and in the administrative and judiciary branches of the government. He sought also to enlarge the kingdom. In 728, taking advantage of the wave of discontent and rebellion aroused in the Byzantine parts of Italy by Emperor Leo III the Isaurian's decree condemning the cult of images (the beginning of the famous Iconoclastic Controversy; 725), he invaded the Exarchate of Ravenna and the Pentapolis (territory south of Ravenna), threatening Ravenna and advancing toward Rome. He claimed to be a defender of orthodoxy and the Pope, but the latter, Gregory II, was alarmed by his approach and sought the support of the dukes of Spoleto and Benevento. Liudprand, sensitive to the Pope's admonitions, withdrew after turning over to him the castle of Sutri (728), though he still had to face the rebel dukes. His later expeditions into the exarchate and the Roman duchy met with only ephemeral success, as was the case with the first occupation of Ravenna (739). His expansionist policy failed not only because of the resistance he encountered but also because his Catholic faith stayed him from attacking the Pope with armed force.

The papal alliance with the Franks

After Liudprand, Aistulf (749–756) continued the same expansionist policy. Aistulf occupied Ferrara, Comacchio, and Ravenna and acquired control of the duchy of Spoleto. The papacy, alarmed by the Lombard advance, turned for support to the powerful Frankish kingdom in northern Europe. In 754 Pope Stephen II went to France, where he anointed and crowned King Pepin the Short, thereby consecrating the legitimacy of the Carolingian dynasty. In return, Pepin and his major dignitaries promised a Frankish intervention in Italy and restitution to the Patrimony of St. Peter of the territories to which the popes made claim. This alliance between the Holy See and the Frankish monarchy had a decisive influence upon Western history; it marked the beginning of the collapse of the Lombard kingdom, the formation of a papal state, and the renewal of the idea of a universal empire. Pepin came twice down to Italy, defeating Aistulf and forcing him to give to the Church of Rome the territory he had wrested from the Byzantines. The new Lombard king, Desiderius (757–774), tried at first to allay the Frankish threat by alliance: he gave his two daughters in marriage to Pepin's sons, Charlemagne and Carloman. But he failed in his intent. When he made a new break with Pope Adrian I in 772, Charlemagne, then occupying the Frankish throne (and having repudiated his Lombard wife), descended upon Italy (773–774), defeated Desiderius and his son Adelchis, and put an end to the Lombard kingdom. The duchy of Spoleto came under his rule, although keeping a character of its own. Benevento, however, remained independent.

The Lombards left a lasting imprint upon Italian history, even if it is not always possible to establish how many customs of the period from the 6th to the 8th century were theirs and how many should be attributed to the crude conditions into which Roman society, in an economic and cultural regression, had fallen. The Regnum Langobardorum (Regnum Italiae) survived for centuries within the new medieval empire, mostly as an ideal but retaining, nevertheless, some of its political and juridical structures (the coronation of the king at Pavia, Milan, or Monza, the archbishop of Cologne's custody of the record office, the institution of counts and judges). The Lombards left noteworthy traces in the fields of law (particularly penal law) and art (sculpture and jewelry). Excavations of the necropolises of Cividale del Friuli, Bolsena, Nocera Umbra, Benevento, and other places have brought to light swords, buckles, gold pectoral crosses, coins, and other objects linked by a definitive common style. Noteworthy are the stylization of the animals, the use of precious stones, and the taste for braided ornamentation. The Lombards were quick to learn Latin and were definitely bilingual toward the end of their rule. Traces of the Lombard language are to be found as late as the 10th century; even more important was the incorporation into Latin of certain words that subsequently passed into Italian.

The Frankish conquest lowered the conquered Lombards to the level of the Romans, thereby hastening the process of integration, already well under way by the end of the 7th century. This integration was chiefly of the masses, since aristocratic families and certain permanent groupings maintained, because of their privileges, a separated position, although they, too, gradually fused with the new society. From the integration of Romans and Lombards, a new people—the Italians—was born.

The Exarchate of Ravenna

Byzantine territories in Italy. At the beginning of the 7th century, after the first wave of the Lombard conquests, the Byzantines retained scattered areas along the coast. Of the ancient province Venetia et Istria there remained not only Istria but also the coast of the northern Adriatic, with a wedge-like extension pointing in the direction of Oderzo. This last separated Friuli from Treviso, for which reason Rothari later took it over. The inhabitants of Aquileia, Concordia (now Concordia Sagittaria), Altino, Treviso, Padua (Padova), Monselice, and Oderzo, cities attacked or destroyed during the successive conquests, sought refuge on the marshy coast or on the islands of the Lagoon (Laguna Veneta), where formerly there were only a few fishermen's villages. Farther south lay the Exarchate of Ravenna, corresponding to the present-day Romagna, with the addition of Bologna and part of Emilia. Ravenna was the headquarters of the exarch, the empire's chief representative, who had jurisdiction over all of Italy except Sicily (Sardinia, Corsica, and the Balearic Islands belonged to the Exarchate of Africa). Ravenna was also the seat of an important archbishopric, which the emperors favoured in every possible way in order to set it against Rome. As early as the end of the 7th century, it obtained the privilege of autonomy. The territory of the Exarchate was contiguous with that of the Pentapolis, the five cities of which were probably those of Rimini, Pesaro, Ancona, Numana, and Osimo, but the full extent of which was approximately the same as that of the present-day region of the Marches. A line of strongholds, some of them along the Via Flaminia and others on a lesser road to the west, connected the Pentapolis with the duchy of Perugia and this to the duchy of Rome (present-day Lazio). Territorial continuity was thus achieved between the lands on the Adriatic and those on the Tyrrhenian Sea, and a sort of diaphragm was set up between Tuscia (Tuscany) on the west and the duchy of Spoleto in the east. Lunigiana and maritime Liguria were isolated and could not long resist the invaders; in 640 they were conquered and occupied by Rothari. The duchy of Naples was better protected by virtue of its location and its access to communications by sea. At the southernmost end of the peninsula, divided by the territory of Benevento, were the two regions of Calabria (present-day Puglia) and Bruttium (present-day Calabria). The Byzantines also possessed Sicily and the other islands, which, between 650 and 750, had to be defended not so much from the Lombards as from the Arabs of North Africa.

Byzantine administration

An important change had come about at this time in the Byzantine administration. In the face of the Lombard invasion, the classic principle, reasserted by Justinian, of the separation between military functions and civil functions was abolished. Thus, the exarch acted as both military and administrative leader. His powers were much broader: he could make peace, contract alliances, adjudicate lawsuits, make official appointments, intervene in the affairs of the church, and confirm the election of the pope. At a lower level the same thing held true: the dux combined civil and military powers, administered justice, saw to the collection of taxes, and named the executives of his *officium.*

Byzantine rule in Italy was anything but tranquil, quite aside from the struggle against the Lombards. The remoteness of the emperor and his preoccupation with resistance to the Arabs made for a certain independence on the part of his subordinates, but, at the same time, it deprived them of the military and financial support that they needed for administrative purposes. The imperial officials had to provide for themselves with whatever means they could find on the spot. Moreover, their efforts were hampered by the emperor's intervention in religious affairs. Such intervention was often motivated by the necessity of resolving differences in the eastern provinces and creating a greater spiritual unity with which to resist external enemies. But, when the emperors pronounced themselves on doctrinal matters, they sometimes offended the religious conscience of the Italic peoples and aroused their enmity, which was often fostered by the bishops and especially by the pope. At times, the imperial officials found themselves in the embarrassing situation of carrying out orders directed against the popes, orders that encountered resistance on the part not only of the people but also of the militiamen who were supposed to execute them. For, since Byzantine soldiers were few in number, defensive and policing tasks were carried out by a locally recruited militia, the *scholae.*

There resulted considerable political and administrative confusion. Exarchs, dukes, and tribunes often carried out policies of their own, in opposition to one another and to imperial orders; local aristocrats joined the game to further their own advantage. When Justinian II ordered the arrest of Pope Sergius I (687–701), the militias of Ravenna and the Pentapolis marched on Rome (*c.* 694) in order to protect him. Thus, there was no repetition of the case of Pope Martin I, who was arrested by the emperor Constans II and died in exile in 655. Similar rebellions took place in the first years of the 8th century in Rome, in 711–712 in Ravenna, and in 717–718 in Sicily, premonitory of the major crisis at the time of the Iconoclastic Controversy. From that year (725) up to 751, the date of the Lombard occupation of Ravenna and the end of the exarchate, the impulse of the Italic people toward autonomy was accelerated. The Greek dukes were driven out and replaced by local rulers; the pope in Rome and the archbishop in Ravenna acquired greater political power, while Venice, Gaeta, Naples, Sorrento, and Amalfi found in independence an incitement to seeking their fortune in sea trade. In short, the Byzantine possessions in Italy, already geographically scattered and progressively diminishing in size, finally lost political cohesion; the emperor's sovereignty over the cities and over Sardinia became purely nominal. Exceptions were Sicily (soon to be conquered by the Arabs), Puglia, and Calabria, which remained under the direct rule of Byzantium. Indeed, with the new Macedonian dynasty begun by Basil I (867–886), Byzantium resumed an active role in southern Italy, determined by the military prowess of the strategist Nicephorus Phocas. After the Arabs had been expelled from Taranto and from the Ionian coast and after the princes of Benevento had been driven to the north, the Puglian region, together with the Capitanata and the Gargano peninsula, came to form the theme of Longobardia, while Calabria, a base essential to what was left of Sicily, was liberated from Arab occupation and made into the theme of Calabria. The division into thema fitted into the new political and military organization imposed upon the Eastern Empire by Leo VI the Philosopher (886–912). The conquest and re-organization of southern Italy were accompanied by new Eastern monastic institutions, notably those of the Basilian monks, that had great importance not only from a political but also from a religious and cultural point of view. Byzantine rule in the south lasted until 1071, when Bari, the seat of the emperor's representative, fell to the Normans.

The duchies of Spoleto and Benevento. The duchy of Spoleto was formed at the beginning of the Lombard invasion. Its first duke, Faroald (*c.* 571–591), and his successor, Ariulf (591–601), staked out boundaries that lasted until the Carolingian age. The duchy did not precisely coincide with any of the natural geographical regions of Italy. It occupied the eastern part of Umbria (the western part, with the fortresses or cities of Amelia, Narni, Terni, Perugia, and Gubbio, made up the Byzantine corridor between Rome and Ravenna), a southern area of the Marches (with Fermo and Camerino), and a strip in the northern section of the Abruzzi. Politically, it had an autonomous development, even if there were, in certain periods, close ties with the Lombard kingdom.

The autonomy of Spoleto

There were many reasons for this autonomy, first among them being its geographical isolation in an inaccessible and impregnable area of the central Apennines, which was further cut off from the Lombard kingdom by the Rome–Ravenna corridor. The dukes of Spoleto could pursue policies of their own, and they knew quite well how to assert themselves in the complex diplomatic and military interplay of their immediate neighbours—the popes, the exarchs, and the dukes of Benevento—none of whom was powerful enough to crush them. The Lombard kings might have had such power, but they exercised it in a reluctant and intermittent fashion.

Later, Charlemagne was able to impose his rule upon the duchy, but even he did not deprive it of its political and administrative individuality. A succession of dukes of Frankish origin supported the Carolingians in their expansionist designs upon Benevento. Meanwhile (in the first years of the 9th century), it seems that Spoleto acquired Chieti and Ortona, while Fermo and Camerino became independent and jointly formed first a "county" (*comitatus*) and then a "march" (*marca*). During the period of confusion and struggle after the end of the Carolingian dynasty (the deposition of Charles the Fat), a duke of Spoleto, Guy, managed to have himself crowned king of Italy at Pavia (889) and emperor at Rome (891). His son Lambert was emperor after him, but, by then, the title had no real meaning.

In the 10th century, the history of the duchy of Spoleto became increasingly confused and obscure. Dukes followed one after the other in rapid succession, but no one of them was able to found a dynasty. Later, with the houses of Franconia and Swabia, it passed into the hands of the great German vassals until, through the efforts of Innocent III and Gregory IX, it became a province of the Papal States.

The duchy of Benevento, too, goes back to the early years of the Lombard invasion, and its origins are bound up with the names of its first two dukes, Zotto (died 591) and Arichis (died *c.* 641), who led its extensive conquests. The distant Lombard kings had neither lands nor officials in the duchy. Indeed, Grimoald, duke of Benevento, entered the contest for the Lombard throne, was elected king, and reigned from 663 to 671, leaving Benevento in the hands of his son Romuald, who left it, in turn, to his descendants. A hereditary right was clearly established.

The borders of the duchy varied, but, except for some temporary conquests (Bari, Taranto, and Brindisi), they enclosed the southern part of the Abruzzi; Molise; the interior of Campania (with Capua and Salerno on the Tyrrhenian Sea); Lucania, with a piece of coastline on the Ionian Sea; and strips of northern Puglia (with Lucera) and Calabria. The territory was divided into some 30 administrative units ruled by ducal *gastaldi,* who enjoyed full civil and military powers. Like Spoleto, Benevento was a landlocked state, with no interest in sea trade and an agricultural economy that could not prosper because of the mountainous terrain and the lack of communications. A military regime, continuous wars, and social oppression made it a backward and conservative zone as compared to the Lombards' other Italian possessions.

Benevento's resistance to the Carolingians

With Arichis II (758–788), son-in-law of King Desiderius, the duchy of Benevento became the last seat of resistance (at times open, at others undeclared) to the Carolingians: as such it was changed from a duchy into a principality, and the capital was moved to Salerno. Lombard traditions lingered, and, even after the 9th century, there were expressions of "national" consciousness. A form of writing, the "Beneventan hand," remained independent of the "Caroline hand." The cultural and religious centre of the whole region was Montecassino, where the influence of the Cluniac reform was felt in the 10th century. Grimoald III, son of Arichis II, continued to combat the Carolingians, but, after his death in *c.* 806, a series of internal struggles led to a division into two principalities: Benevento and Salerno. A third territory gradually broke away from these two and became the independent county (later principality) of Capua. All three principalities fell into the hands of the Normans. Benevento was captured by Robert Guiscard in 1081 and immediately turned over to the Papal States, to which it belonged, with few interruptions, until 1860. Salerno and Capua were absorbed by the new Norman state.

CAROLINGIAN AND FEUDAL ITALY

The kingdom of Italy. In April 774, after sweeping away weak Lombard resistance, Charlemagne arrived in Rome, placed a new act of donation to the Roman Church (confirming Pepin's of 18 years before) on the tomb of the Apostle Peter, and acquired the title of *rex Francorum et Langobardorum* ("king of the Franks and Lombards"). A few pockets of rebellion remained (Benevento, Trento, Friuli), which obliged him to return to Italy in 776 and in 780–781. On this last occasion, his son Pepin was crowned in Rome by Pope Adrian I as "king of Italy." In a fourth descent upon Italy, in 787, Charlemagne defeated Arichis of Benevento and also Adelchis, son of Desiderius, who had landed in Calabria with Byzantine re-enforcements. The duchy of Spoleto and the march of Fermo were added to the kingdom.

In the following years, Charlemagne campaigned against the Saxons, the Arabs in Spain, the Bavarians, and the Avars—making himself master of a large part of western Europe and winning immense prestige. The logical consequence was the famous *renovatio imperii,* or second edition of the empire, in the 9th century, the attempt to give Charlemagne's conquests a political unity and an ideological basis. The reconstructed empire drew inspiration both from the unforgotten traditions of ancient Rome and from a religious ideal. At the same time, it reflected political and social realities in the process of transformation, which had been deeply influenced by the law, customs, and national characteristics of the Germanic peoples. But the medieval empire, at the time of Charlemagne and in its successive restorations, was never a true state but rather a symbol of the community of peoples of Christian Europe. Even contemporaries referred to it as an *imperium plurimarum nationum,* an empire of many nations.

The coronation of Charlemagne

The coronation of Charlemagne as emperor on Christmas Day, 800, in Rome, aroused the opposition of the Byzantines. But the differences between them were soon overcome, and in 812 they made a peace agreement. From the death of Pepin (810) until 887, Italy was governed by Carolingian princes, and, in spite of the dismemberments to which the empire was subjected, it retained the political and territorial setup of the time of the Carolingian conquest. Bernard, son of Pepin, eventually rebelled, but he was captured and blinded and died in 818. His successor, Lothair I, son of Louis I the Pious, took on the title of emperor. Of considerable importance is the Constitutio Romana, put out by him in 824, in which he prescribed that the pope, after he had been elected by the clergy and people of Rome, must swear allegiance to the emperor. By virtue of the territorial division of 843, Lothair retained the title of emperor (implying the possession of Italy) and also a long, narrow strip of territory between France and Germany that took from him its name of Lorraine. This division took no account of ethnic and linguistic differences but affirmed, rather, the principle that the possessor of the imperial crown must necessarily possess the capitals of the empire, Rome and Aix-la-Chapelle (Aachen).

The son and successor of Lothair, Ludwig II (855–875), bore the double title of king and emperor but reigned over Italy only. There ensued a renewal of Carolingian power and polity in the peninsula and a new wave of expansion toward the south. Ludwig made numerous expeditions into southern Italy, where political fragmentation and continuous differences among the local rulers invited raids on the part of the Arabs, who were just completing their conquest of Sicily. But political, military, and climatic difficulties thwarted the ambitious project of driving out the Arabs and achieving pacification.

The last Carolingians

The last Carolingians to govern Italy—in spite of the fact that their main interests lay elsewhere—were Charles II the Bald (875–877), Carloman (877–879), and Charles III the Fat (880–887).

Politically and administratively, Charlemagne's conquest did not cause the Lombard kingdom to be annexed to the Frankish state; on the contrary, it kept a juridical character all its own. There was, at first, a sort of "personal" union, inasmuch as, for a while, Charlemagne had two crowns; later, the kingdom of Italy had its own sovereigns, chosen from the Carolingian dynasty. Most often, however, these were emperors who enjoyed also the title of king of Italy. Noteworthy is the fact, remarked above, that the appellation Regnum Italiae took the place of Regnum Langobardorum. As the conquered Lombards' memory of their glorious past faded, a historical and geographical term that reflected the peninsula's allegiance to the empire came into use.

Under the Carolingians, no important changes were made in the central administration. Pavia remained the capital and seat of the court, government offices, and the assembly. The assembly was made up of the dignitaries of the kingdom, including numerous ecclesiastics who in the Carolingian age played a greatly increased role in the state; the assembly's legislative powers, however, were less than they had been in the Lombard period. Laws decreed by the emperor for the whole empire applied also to Italy, and in the palace of Pavia a special authority was exercised by the *comes sacri palatii,* or count palatine, who headed the royal tribunal.

Carolingian administration

More drastic modifications were imposed upon the outlying administration. The Lombard duchies were eliminated and were replaced by new districts, or counties, ruled by counts; that is, royal representatives with military and judicial functions, drawn from the Frankish aristocracy. The county organization in Italy, however, was not as widespread and important as was formerly supposed; certainly, it cannot be compared to that of the Frankish kingdom. Counts were not installed everywhere; in some places the old Lombard *gastaldi* remained after making an act of submission. In others, Lombard *gastaldi* and judges stayed on, alongside the counts, and exercised a judiciary function rivalling theirs. This was the case mostly in the cities, traditionally seats of government and the administration of justice. Because the greater part of the cities were also diocesan headquarters, the Frankish counts found other powerful and prestigious rivals among the bishops. Another limitation of the counts' power derived from the institution of the *missi dominici,* pairs of itinerant controllers, one a cleric and the other a layman, who made periodical visits to the counties and presided over their judiciary assemblies (*placita*). Thus, in Italy the Frankish counts found no way of taking root in a background so different from their own. Some of them moved into country estates and castles, from which they could exercise no more than a limited jurisdiction; others returned home. There were, of course, exceptions. The count-duke of Spoleto and the count-marquis of Friuli ruled over frontier territories where the powers of the *gastaldi* and bishops had been limited from the start and where it was easy to establish dynasties.

Aside from the greater or lesser ability of the Carolingian counts to settle on their Italian lands, they did, very definitely, introduce a new mentality and new systems into governmental practice. Like other functionaries, they were chosen among the *fideles* or *vassi* of the sovereign and bound to him by personal as well as bureaucratic dependency. It often happened that the holder of a public *honor* or *officium* held also, as a vassal of the king, lands *in benefice* far from his own county. And often these lands, as those belonging to churches and monasteries, had immunities or special privileges, consisting of relief from subjection to the royal courts and exemption from the payment of tribute. Because of the progressive weakening of the central power between the mid-9th and mid-10th centuries, this period is rightly called one of feudal anarchy. Central and northern Italy witnessed continuous and ferocious struggles among the noble families. Marquises of Friuli, Ivrea, and Tuscany and dukes of Spoleto disputed the titles of king and emperor, calling, at intervals, for aid from the powerful lords of Carinthia, Provence, and Burgundy. Holders of the royal crown between 888 and 961 were Berengar I of Friuli, Guy of Spoleto and his son Lambert, Arnulf of Carinthia, Ludwig III of Provence, Rudolph II of Burgundy, Hugh of Provence and his son Lothar II, Berengar II of Ivrea and his son Adalbert. A few of them held the imperial crown as well; the last was Berengar I, who reigned until 924.

Decline of central government

In view of the inadequacy of the central power and the general political insecurity, to which must be added repeated incursions of Arabs and Hungarians, there was considerable local individuality in the feudal forms prevalent all over Europe. The exemptions and privileges connected with feudal lands and ecclesiastical properties were extended until, to include administration of petty justice, the control exercised by the *missi dominici* was reduced, and, indeed, the *missaticum* was conceded to bishops and feudatories, thus increasing their power and autonomy; *honores* and *officia* became feudal benefices, so that the recipients could hold them in perpetuity and eventually transmit them to their children; bishops obtained increasing administrative power because, in the rulers' view, they could hold lay feudal elements in check. As a result of these developments, the state was no longer run by a central government (except for the remnants surviving at Pavia) but by a hierarchy of noblemen who were bound by a relationship of personal dependency and closely linked to individual localities and to the land.

It would be naïve to give an absolute value to the term feudal anarchy and to think that feudalism represented a pure and simple degeneration of the preceding political system. The fact is that the preceding system had broken down, and another one—feudalism—had to be found to take its place. Feudalism, in spite of its great disadvantages, was an orderly construction, and, as such, it served as the framework of European society for a number of centuries. Although feudalistic fragmentation favoured disagreement among the powerful, it had in some ways a positive function; for example, the efficient organization of local defense against Hungarian incursions. In the course of the 10th century, the plains of northern Italy were covered with *castra;* that is, fortified areas where the people could store food and take refuge in case of danger. At the same time, city walls were restored and fortified by order of the bishops. These local enterprises were important in speeding the demographic and economic recovery that was already under way.

Frankish feudalism, whether because of the influence of Roman law or because of local conditions, did not long keep its original character. The accordance of a "Lombard" fief (*jure Langobardorum*) demanded only an oath of loyalty on the part of the vassal, without any ceremony of homage. The next step was the hereditary conception of the fief (with division, real or theoretic only, among all the sons) and the fragmentation of feudal rights. This explains the unrest of the first decades of the 11th century, culminating in the famous Constitutio de Feudis of 1037, with which Emperor Conrad II conceded the inheritability of minor fiefs.

The origin of the Papal States. Between the middle of the 9th and the middle of the 10th centuries, an event portentous for Italian and world history took place: the formation of the temporal states of the church. It was the end result of a long and complex historical development, in which varied elements—religious, political, military, and cultural—played a part. The first to be considered is the progressive detachment, mentioned above, of the peoples of Byzantine Italy from Byzantium itself. The popes often had to stand up against the interference of the *basileus* in religious affairs and the claims to superiority of the patriarch of Constantinople and to protect the population from Byzantine officials and tax collectors. For this reason the popes came to be considered the leaders of what might be called a national movement. Moreover, the popes took an active part in aid to the poor, especially in Rome. For this purpose they drew upon the church's considerable financial resources and did what the negligent Byzantine government left undone. Meanwhile, all over the West, the authority and prestige of the popes were growing. Everywhere, Rome was looked to as the centre of spiritual and civil union of all the peoples of the West.

The decisive development in the creation of the Papal

Benefits of the papal and dynastic alliance

States was the alliance between the papacy and the new Carolingian dynasty. The alliance benefitted both parties: for the papacy, the Franks were strong enough to hold both the Lombards and the Byzantines in check, but far enough away that, unlike the Lombards, they represented no imminent threat; for the Carolingians, the popes' spiritual prestige ensured the legitimacy of their dynasty. In 756, on his second Italian campaign, Pepin ceded to the papacy the former Byzantine territories of the Exarchate of Ravenna, the Pentapolis, and the duchies of Rome and Perugia. This donation was confirmed and amplified by Charlemagne but not recognized in Constantinople. Thus, the temporal rule of the Roman see was founded on an insecure legal basis. It was no longer a personal and private matter, such as the previous *patrimonium beati Petri,* but neither was it public, *pleno jure*—that is, with an explicit recognition of sovereignty. The religious significance of the donation, however, with its sacred and intangible character, together with the effective exercise of power that followed, eventually lent sanction to the popes' temporal rule. It is equally true that the Frankish king, as a *patricius Romanorum* (the title conferred by Pope Stephen II upon Pepin in 754), had taken on the right and the duty of protecting Rome and the pope against all enemies; hence, the emperor's function as *advocatus ecclesiae,* which came to be an integral part of the ideology of the medieval empire. Although the concept was generally accepted, its application raised serious problems and complicated the relationship between the empire and the papacy.

There has been and still is considerable discussion of the circumstances of the notorious Donation of Constantine, which is so closely linked to the beginnings of the church's temporal power. This was a forgery, a pretended document in which the Emperor, after narrating his miraculous recovery from leprosy and his subsequent conversion by Pope Sylvester, donated to this pope the Lateran palace, Rome, Italy and its islands, and, indeed, the entire western part of the empire. The whole thing is probably the work of a cleric attached to the Roman Curia, between the pontificates of Stephen II and Adrian I. The size and vagueness of the donation make it a statement of principle rather than a legal proof; this is confirmed by the importance that is given in the document to the concession to the pope of the diadem, purple garments, and other symbols of empire and also the equality established between the papal dignitaries and those of the imperial court. It seems as if the forger's primary intent was to establish the pope's claim to a *dignitas* equal to that of the emperor and as if the territorial donation was merely a corollary of that dignity. The use made of the document in the Roman Curia confirms this interpretation. In any case, the Donation of Constantine is an extremely important document for understanding the development of the political ideology of the papacy.

During the reign of Charlemagne, the popes had little chance to make a political place for themselves in the framework of the empire. Pope Leo III had, indeed, crowned the emperor in 800, but there was no doubt that Charlemagne was the dominant partner in the alliance. A few decades later, however, there was a radical reversal in the situation: the various partitions of imperial territories, the wars among Charlemagne's successors, and the Norman invasions had brought the empire to a state of crisis. Energetic popes, such as Nicholas I (858–867) and John VIII (872–882), took the political initiative and tried to save what was salvageable of the empire's unity. For this purpose they had to emphasize the principle of their authority in temporal affairs. John VIII, for instance, declared that it was up to the pope to choose the future emperor (a declaration that was the beginning of the legend that Leo III was the creator of the Holy Roman Empire). The pope was no longer presented as the officiator of a religious ceremony; he appears, rather, as the bestower of the imperial crown.

Papal influence in temporal affairs

The pope's growing political authority and prestige of this period underwent a later eclipse. But the accumulation of past experience was not lost. It served as a testing ground at the period of the struggle over investitures in the 11th century, when the doctrines of the Curia began to take more definite shape. For the moment, the popes were involved in the general process of political fragmentation and feudal localization that characterized Italian and European history in the 10th century. Whereas, for a time, they had held first place on the political scene, after the death of John VIII they remained in the shadow of one or the other faction contending for power in Rome.

Political morality had sunk to a low level in the 9th and 10th centuries; political murders were frequent, and many popes were imposed by force. Notorious is the posthumous trial of Pope Formosus (897), whose corpse was dug up, judged before a council, and then thrown into the Tiber. Yet, in spite of everything, there were examples of civic conscience and even of moral stature. The *vestararius* Theophylactus, a papal official who founded the fortune of his family, and his daughter Marozia, successively the wife of Alberic I of Spoleto, Guy of Tuscany, and Hugh of Provence and, hence, involved in the major political currents of the time, stand out as examples of boundless energy and genuine constructive ability. Marozia's son, Alberic II, assumed the title of *princeps atque omnium Romanorum senator* and ruled over Rome from 932 to 954. Alongside episodes of violence, one finds remembrance of the traditions of Classical antiquity and a desire to restore order to both church and state. Alberic, for example, led an expedition against the monastery of Farfa in Sabina, where religious life was in decay, and turned the monastery over to Cluniac reformers.

But it is in the struggle against the Muslims that one finds the brightest sporadic examples of faith, patriotism, and courage on the part of the popes, the aristocracy, and the people of Rome between the 9th and 10th centuries. The Arab threat was very real. In 846 the Arabs had sailed up the Tiber and sacked the basilicas of St. Peter and St. Paul. Pope Leo IV and the Romans replied by raising walls around St. Peter's (the "Leonine city") and stretching a chain across the river. In 849 a Muslim fleet was defeated at the mouth of the Tiber by ships belonging to a league made up of Rome, Gaeta, Naples, and Amalfi. John VIII tried his best to maintain a united front of southern princes against the Arabs, though without success. Indeed, the Muslims built a military base at the mouth of the Garigliano, which became the departure point for destructive expeditions against the towns and rich monasteries of the interior. "Redacta est terra in solitudine" ("The land was given up to wilderness"), wrote one chronicler. Finally, Pope John X, a creature of Theophylactus, managed to rebuild the league, which now included a Byzantine fleet; in 915 he personally led a successful attack on the base on the Garigliano. At this period of Roman history, traditionally known as the Iron Age of the papacy, there existed a vital impulse, an exceptionally powerful charge of energy, and an organizing ability that were not confined to the pursuit of personal ambitions but were often inspired by idealistic political purposes.

The Arab threat

Venice and the cities of Campania. Venice is one of the few Italian cities stemming from the Middle Ages and the only one to have a demographic, economic, urban, cultural, and political character of an individual and exceptionally original kind. During the barbarian invasions of the 5th and 6th centuries, the clergy and many of the people of the interior of Venetia sought refuge on the coast and on the coastal islands between Grado and Chioggia. There, they soon developed an intensely associative life, witnessed by the transfer of the patriarchate of Aquileia to Grado and by the creation of the new bishoprics of Caorle, Eraclea, Iesolo, Torcello, Malamocco, and Olivolo (Venice), which, one by one, inherited the ecclesiastical structures of the hinterland dioceses. The increase of the population and the consequent necessity of procuring supplies and transportation stimulated economic development. Not all the refugees were poor; among them there was a majority of *possessores* who had brought with them their transportable wealth and hence were able to put up capital for commercial enterprises. At the same time, the *possessores* kept—or later recovered—at least part of their landholdings in the back country. In the 8th and 9th centuries, there arose an aristocratic and business-minded class that exploited its capital in trade, transportation, salt mines, and moneylending, a class that was fated to play more and more of a role in politics.

Politically, the various centres in the process of formation along the coast and on the Lagoon joined together toward the end of the 7th century in a duchy ruled by a Byzantine duke. The seat of the duchy was first at Eraclea, then, after about 740, at Malamocco, and finally at Rialto, the first nucleus of Venice, from the beginning of the 9th century, a period of decisive importance for that city. The Carolingians, who at first tried to conquer Venice, gave up this idea and, in 812, made peace with Byzantium through the Treaty of Aix-la-Chapelle. Venice found itself in the best possible position for economical and political development, as an indispensable intermediary for the relationship between East and West.

The economic growth of Venice

The principal reason for Venice's extraordinary economic growth lay in the change undergone by the economy of the whole Mediterranean area. With the Arab conquest of Syria and Egypt, the Eastern Empire had lost two major markets and also an important source of its food supplies. No longer was there sea traffic between France and the East passing through the western Mediterranean. Under these circumstances the Po Valley assumed an important role, because of its agricultural production and also because the navigability of the river as far as Pavia assured communications with western Europe. Venice found itself at the crossroads of East and West, which needed each other economically even if, politically, they were divided. Thanks to its independent status and the enterprising spirit of its citizens, it succeeded in putting them in touch and becoming the great trading place for goods (silk, spices, and luxury objects from the East; wheat, oil, and salt from the West) and moneys (Byzantine, Lombard, and even Arab gold coins and Carolingian silver). Later Venice took advantage of its privileged position to make contact with Egypt, Sicily, and the Arab world in general, with which it traded in pelts, wood, arms, and slaves.

As in the other Byzantine territories in Italy, Venice gradually acquired autonomy in the 8th century. *Tribuni, magistri militum,* and *duces* were less and less often chosen by the emperor or his representatives and more frequently elected on the spot. In these choices, quite naturally, the interests of the local aristocracy prevailed. The post of tribune, which was both civil and military, soon became hereditary and so, later on, did that of the duke (subsequently called doge), though with alternating families and with factional strife. Important among these families or dynasties were those founded by Agnello Parteciaco (or Partecipazio; 810–827) and Pietro Candiano (887), which alternately held rule until 976. In that year the first Orseolo came to the fore and founded a new house, the greatest member of which was Pietro II (991–1008). After obtaining recognition of his authority from Byzantium, he carried out an intelligent policy of territorial and commercial expansion. In 1000, to free the Adriatic from pirates' raids, he conquered the coast of Dalmatia as far south as Kotor and assumed the title of *dux Veneticorum et Dalmaticorum.* Then, in 1002, he gave aid to the Byzantines when they were defending Bari against the Saracens. But his dynasty lasted no later than 1032.

The role of the doge

At the time of the Orseoli, the doge was similar to a king. He commanded the armed forces, presided over the court of appeals, appointed government officials, and discharged public funds. He also controlled the church, at least in its material possessions, appointments, and benefices. In all his administrative functions, particularly those concerned with the administration of justice, he was assisted by a *curia ducis,* or ducal council, composed of high government officials, judges, the patriarch, the bishops, and the abbots of the major monasteries. Representatives of the people, called *boni homines* or *fideles,* had seats there as well, but it is likely that they, like the other members, were chosen by the doge. There existed also a *concio civium,* an assembly of all the freemen of the duchy that met on important and solemn occasions. It is probable that this assembly complied with the will of the wealthy landowners and merchants established at Rialto, who, toward the middle of the 12th century, gave the Commune Veneciarum its oligarchic character.

There are interesting analogies but also striking differences between Venice and the maritime cities of Campania: Gaeta, Naples, Sorrento, and Amalfi. They, too, were in the Byzantine sphere and won increasing independence and also importance in sea trade; but the stuff of their history was far more brittle, and they met a very different fate. The essential difference is this: the Campanian cities, surrounded and threatened by hostile or potentially dangerous powers (the Byzantines, the Saracens, the Lombards of Capua, Benevento, and Salerno) never enjoyed a security such as that of Venice. On the contrary, they were constantly forced to remain on guard and on the defensive, with diplomacy and arms, until they were definitively absorbed into the Norman kingdom. Because of the continuous political tension, the aristocracy of these cities could not or would not transform itself into merchants, such as those of Venice, but remained land bound and warlike, while the merchants, drawn from the middle class, were socially and, above all, politically weak. Norman rule, when it came along, merely gave definite sanction to a pre-existent state of affairs. The small centres—Sorrento and Amalfi—once they were amalgamated with a vaster and centralized governmental apparatus, lost all autonomy and were cut back to the limitations of local power. The big city, Naples, did, indeed, maintain and increase its importance (especially later on, under the House of Anjou).

Sicily under the Arabs. From the second half of the 7th century, Sicily was the object of Arab attacks from Africa. The real Muslim conquest, however, took place in the 9th century. Opportunity was provided by the rebellion of the commander of the Byzantine fleet, Euphemius of Messina, who turned for help to the Emir of al-Qayrawān (in present-day Tunisia), a member of the Aghlabid dynasty. His appeal was answered, and an Arab army landed at Mazara in 827. The Arabs soon discarded Euphemius and continued the expedition on their own, conquering Mineo and then Palermo (831). The Byzantines put up a long and stubborn resistance; Syracuse held out until 878, when it was mercilessly plundered. The island became an Aghlabid province, passing after 910 into the hands of the Fāṭimids, the Shīʿite dynasty that in 972 moved its capital to Cairo. From the middle of the 10th century until 1040, Sicily was an emirate, to all effects and purposes independently ruled by the Kalbī family. After 1040 internal dissension divided Sicily into small local lordships until the Normans took advantage of such obvious political weakness and, after a 30-year war (1061–91), imposed their rule.

For approximately 200 years, Sicily was the chief base for Arab expansion on the seas and along the coasts of central and southern Italy. They were not always engaged in piracy and war against the Christians; there were truces and trade agreements and profitable exchanges at a cultural as well as a commercial level. Indeed, political agreements and alliances were sometimes set up between Christian and Muslim princes.

Sicilian prosperity under the Arabs

In these two centuries Sicily enjoyed economic prosperity and an intellectual flowering. The natives were treated with respect and allowed to keep their Christian faith, although this meant that they had to pay tribute money and were in a position of legal inferiority. By right of conquest, land became largely the property of the state or of individual Arabs. There was intense agrarian colonization and much breaking up of large holdings, accompanied by technical improvements, particularly in irrigation. Vineyards were destroyed, but new products—citrus fruits, sugarcane, date palms, and mulberry trees (for raising silkworms)—were introduced. The capital city of Palermo was thickly inhabited and prosperous, with a busy harbour and much handicraft activity. In Palermo and elsewhere, the Muslims built castles, palaces, mosques, and elaborate gardens. Poetry, law, the arts, and the study of the Qurʾān were held in honour, especially at the Kalbī court.

CITIES AND COUNTRYSIDE IN THE EARLY MIDDLE AGES

Urban crisis: the "villae." The period between AD 96 and 180, the Age of the Antonines, was perhaps the happiest in the history of Roman cities. The emperors' policy benefitted the senatorial class—that is, the landowners who held municipal magistracies and made up the strongest single support of the government. Hence, it promoted the development of cities as centres of power and control over vast territories. But this system had its weak points, above all its finances, which steadily took a turn for the worse. In the 3rd and 4th centuries the taxes imposed on the prosperous urban class and the enforced contributions and furnishings of supplies became ruinous. City administrators, harassed by the hostile policy of the military monarchy, gave up their posts and retired to their country estates. The cities, left in the hands of greedy officials, exposed to the violence of military garrisons and barbarian invaders, and ruined by inflation, began to decrease in population and to decay. This was the fate of almost all the cities of the West, Rome included.

The duties of the citizen

In spite of depopulation and economic crisis, Italian cities retained some of their ancient importance. In the 4th century there began a movement toward popular control within the cities. Although it is difficult to measure its exact extent and value, it is an index and demonstration of a collective identity bound up with an urban tradition that endured even through the dark period between Classical antiquity and the Middle Ages. There are some legislative texts from which it appears quite clearly that the Roman state tried to persuade the citizenry of the cities to assume greater responsibilities. An imperial decree of 384 made it obligatory for citizen groups to attend to the repair of walls and aqueducts; in 396 there was instituted a special tax upon all *possessores* for the maintenance of the walls; in 400–401 it was established that only *collegiati* and *corporati* citizens could rent communal lands in or near the city; in 440 city dwellers were obliged to serve, under arms, for the protection of walls and gates. This last decree had serious implications. Not only did it point up the impotence of the central government, but it also contradicted the fundamental Roman principle of separation between civil and military powers. And there was more. In 443 it was established that the whole population of a city should take part in decisions regarding the transfer of municipal properties. By a decree in 458, the entire citizenry was allowed to participate in the election of the *defensor civitatis*—that is, the official, formerly appointed by the central government, who supplemented or took the place of the *curiales.*

Just at the time when there was a tendency to entrust collective responsibilities to the city populations, these populations were increasingly drawn together by a common faith and participation in the same liturgical and sacramental life. For it was in the 4th and the 5th centuries that conversion to Christianity became practically universal in the cities (though not yet in the countryside). The result was a communitarian experience of a new kind that did away with many legal and social differentiations of the Classical age. All these facts point to a new evaluation of the period of the late Roman Empire. Although it is traditionally defined as one of decadence, corruption, and disorganization, such a definition is not all-embracing. The life of those days contained new elements, which held hope for the future, chief among them the development of a civic conscience. People once passive and divided were driven by necessity to become aware of themselves, to organize, to express in a different sort of community the ideals of their faith and the object of their aspirations.

Without doubt the decay of the cities had a causal relationship to the increase and enlargement of *villae* and *saltus,* the tilled or wooded lands that became, more and more, centres of production, social organization, and defense. Soon they obtained exemptions; they could shut the door on the tax collector and get their farmers relieved of military service. One can imagine the attraction of these privileges upon small landholders; by an act called *accomendatio,* they ceded ownership of their lands (conserving its use) to a *possessor,* in return for his protection. The *possessores* were now real local *domini,* or masters. They administered petty justice, collected rents and taxes, held absolute power over their slaves, were accompanied by armed followers (*comites* or *buccellarii*), and sometimes had private churches.

The labour force

As for internal organization, the *villae,* successors to the earlier latifundia, inherited at the start their workings, which were based on slave labour. Later, there came important changes. When the wars of conquest were over, the number of newly captured slaves diminished and that of the older ones diminished as well, because of a low birth rate and also because religion and custom favoured setting them free. Besides these, there were the *accomendati,* the free small landowners who had joined their farms to a large estate. Part of the landed unit, divided into *mansi,* was occupied by farmers or freed slaves, and the *mansi* formed a *massaricium;* a larger part, made up of fields and woods, formed the *dominicum* and was worked by slaves or servants, sometimes with the aid of the free farmers. If the estate had an area of no more than 100 or so acres (40 or so hectares), it was a *curtis,* but, if it was larger (some had an area of as much as 2,500 acres [1,000 hectares]), it was divided into several *curtes,* each with its administrative centre, its *dominicum,* and its *massaricium.* This was the sort of farm estate that antiquity handed down to the Middle Ages, but there were, as shall be seen, other sorts as well.

The villas were not the only form of agricultural organization. Alongside them there were public demesnes belonging to cities, country districts (*pagi*), villages (*vici*), and valley communities (*comunalia, conciliaricia, vicanalia, compascua*), allods (private and free property), and, finally, the great rural properties of ecclesiastical organizations. For lack of data, one cannot say in what proportion each of these types existed. But it is clear that, at least in Italy, they were all represented, even if not uniformly. In the Po Valley, for instance, it seems that there was a majority of small and middling landowners and that only in the 4th century did larger agglomerations come into being, but never, even then, as large as the African *saltus* and the latifundia of Sicily.

It is probable also that, in Italy, the Lombards substantially altered the property structure, especially that of the *villae.* Lack of documentation prevents any definite statement, but it is known that the Roman landowning class was practically wiped out and that the farmers were made tributaries (*tertiatores*). One may most logically presume that the great latifundia were broken up and fell into the hands of the native farmers and the Lombard families that settled in the same countryside. This process was abetted by the Germanic concept of *gewere,* which linked the enjoyment or possession of an object to a right of ownership. Farmers in Italy had, on the whole, an advantage over those of the countries (France, the Rhineland, the region of the Moselle) where there was direct continuity between the Roman villa and the early medieval lordship over the land. This is a fact that may contribute to understanding the development of the Italian agricultural class in the 9th and 10th centuries.

The breakup of the latifundia did not prevent a reintegration of land, under different conditions, through the great increase of ecclesiastical wealth at the end of the Lombards' rule and the beginning of that of the Carolingians. One must remember, also, the survival of the great *curtes regiae,* such as Sospiro, Corteolona, and Bene Vagienna, of which the last named had an area of more than 75,000 acres (30,000 hectares).

Bishops and cities. The link between the bishop and the city was first sanctioned at the Council of Sardica in 343, where it was established that only an important city could be a bishop's seat, "ne vilescat nomen episcopi et auctoritas" ("lest the name and authority of the bishop be taken lightly"). Pope Leo I, toward 446, repeated the same concept, which later served as the basis of a corollary in reverse: there must be a bishopric in every important city. Such statements testified to a long-standing state of affairs—that is, to the increase of episcopal authority from 313 on.

After the Edict of Milan in 313, which extended toleration to the Christians, bishops took an increasing part in urban life. The good works demanded of their religious office included the relief of suffering and the prevention of disaster. They sponsored many charities, such as the distribution of food and the construction of hospitals, protected the poor from the rich and sometimes from the tax collector, exercised an influence over the behaviour of magistrates, both local and national, made up for the deficiencies of public officials (*e.g.*, the construction of an aqueduct at Vercelli and of dikes on the Po River at Piacenza), represented the cities in their dealings with the barbarian soldiery, and either directed or participated in the defense of the walls. These episcopal activities answered a deep-seated need of the times and were well received not only by the local populations but also by the state, which, indeed, acknowledged and encouraged them.

The *episcopalis audientia*

In this connection one must recall the so-called *episcopalis audientia.* In 318, only five years after the Edict of Milan, Constantine decreed that, in a dispute between Christians, the decision pronounced by a bishop should be equivalent to an unappealable civil verdict and be so executed by the appropriate imperial government officials. Decrees dating from 333, 398, 408, and 452 confirmed this recognition of the *episcopalis audientia,* although placing certain limitations upon it. Evidently legislators saw advantages to the state in the bishops' justice. It was preferred by a large part of the population (*e.g.*, that converted to Christianity); it was less bound to juridical formalism and hence more in tune with the new social mentality (*i.e.*, more just); and, finally, it cost nothing. With Justinian the *episcopalis audientia* became more infrequent, chiefly because in a totally Christianized state the bishops were inserted into the regular juridical organization. The Code of Justinian gave bishops supervision over civil judges and provided that, in certain cases, they should take the latter's place. Justinian's Code assigned other administrative tasks to the bishops, particularly in the field of public works. Plainly, bishops were now considered, juridically, pillars of the governmental structure.

This increase of the bishops' power is not difficult to understand. First, there was the sacred character of their office as successors to the Apostles and bearers of their mission. Then, there was the continuity and stability of the office; a bishop's tenure was for life, and this gave it an obvious superiority over the precarious condition of the imperial and city magistrates. Another consideration was the prestige lent to the bishop by the wealth of which he disposed. And even more important was the manner of his election. According to custom, he was chosen by the clergy and people of the city, gathered together on the cathedral green. This made the bishop, theoretically, the representative and fiduciary of the entire population; no other magistrate had the same broad following or moral authority.

Throughout the 5th and 6th centuries and up to the Lombard invasion, the union between the bishops and the citizenry became increasingly close, so that ecclesiastical and civic affairs were closely intertwined. *Fidelis,* one of the faithful, was a synonym of *civis,* or citizen. Everyday life followed the hours of the liturgy, and both the geographical and the administrative layout centred around the cathedral. The assembly mentioned above dealt with both church and city affairs; the cathedral green or square where it met was usually a marketplace as well.

The Lombards did not break up this union; indeed, to some extent they reinforced it. In hard or dangerous times the people gathered around the bishops for comfort and protection. The bishops are chiefly to be credited for the conservation of the civil and civic traditions of Rome, including the principles of Roman law, inasmuch as they probably continued, officially or privately, to act as judges or arbiters of disputes among Christians. But this was essentially a period of transition. In the time of Liudprand, after the mass conversion of the Lombards, the bishops resumed their collaboration with the government, and this role was further strengthened during the age of the Carolingians. The Carolingian counts did not long succeed in stemming the power of the bishops, and the latter became, in most cases, masters of the cities. In the 9th and 10th centuries they obtained a succession of sovereign privileges, especially when it came to the construction and defense of city walls and the concession of markets (Bergamo, Mantua, Modena, Como, Vercelli, Cremona, and others). The bishops collected tithes from their own faithful and virtually exercised civil as well as ecclesiastical rule of the entire district. This explains why, toward the middle of the 10th century, some of them bore the title of count, which was simply an acknowledgment of the authority they had in civil affairs.

The episcopal vassals

For aid in their governing functions, the bishops needed trusted men, chosen from either clergy or laity. In the oldest times there existed a *defensor ecclesiae* (a layman) and then a *vicedominus* (an ecclesiastic). Later, in the prefeudal and feudal ages, there were *advocati, confanonerii,* or *vexilliferi,* and other minor officials who were chosen among families that favoured the bishop and who received fiefs, income, and benefices from church property. But this ever more numerous and demanding class of episcopal vassals, augmented by minor feudatories and by the *cives majores* or *boni homines*—that is, the petty landowners or lessees originating in the country and, in the city, exercising the specialized professions of judge, notary, doctor, or merchant—was the one that brought on a crisis of the bishops' rule. The way was opened to the formation of the communes, end results of a process of profound social change taking place in the 9th and 10th centuries.

Economy and society in pre-communal Italy. In the period roughly between 750 and 1000, during which East and West grew apart economically as well as politically, Italy was still, however, in the ambit of Byzantine power; that is, the south depended on Byzantium directly, while the Lombard and Carolingian north were linked to it indirectly. The fact is important, for it signifies that Italy never, or practically never, had a closed economy. On the contrary, commercial exchanges always flourished; there were many markets and an absolutely necessary circulation of money. Even the Lombards struck coins, among them the gold *tremissi,* minted in Lucca and Pavia. No comparison can be made between the economic structures of Italy and those of France in the 8th and 9th centuries. If the expression "economy of the *curtis*" may apply to continental Italy, it means only that most of the agricultural production was organized in *curtes*—that is, in farm units—and has no implication of autarchy or of doors closed to the outside world.

In southern Italy, trade relations with Byzantium and the East in general were quite active and favoured by political circumstances. Amalfi had colonies and warehouses at Constantinople, Antioch, and Durazzo; its ships sailed up the Adriatic and the Po River all the way to Pavia, put into the Arab ports of Sicily, and made their way along the Tyrrhenian coast as far as Rome, Pisa, and Genoa. On the opposite shore, Bari, which had become the capital of Byzantium's Italian possessions, had trade with Constantinople, Durazzo (Albanian Durrës), Greece, and the ports of Asia Minor and Syria. Along with Siponto, to the north, Bari was the embarkation point for pilgrims going to the Holy Land. And it was through Bari, Siponto, Trani, and Barletta that Byzantium got its supplies of Puglian grain. There can be no doubt, then, that southern Italy was a part of the Byzantine sphere, the economy of which was based on the exchange of goods and money. Even the Arabs contributed to the Byzantine economy, above all through their commercial dealings with Amalfi. Their gold dinar was as acceptable as the Byzantine bezant or hyperper.

Importance of the Po Valley

As for the north, the Po Valley became an important source of wheat and foodstuffs to the Byzantines after their loss of Egypt. The Lombards were not gifted for trade, but they did not prevent natives or other foreigners from engaging in it. An edict of King Aistulf of 754, concerning military service, put *negotiatores,* or merchants, on the same plane as landowners; both categories were divided into three classes, according to the amount of money and property in their possession.

The great artery, then, was the Po, navigable all the way to Pavia, and with it may be grouped the rivers flowing into it from the north—Ticino, Lambro, Adda, Oglio, and Mincio—all navigable for short distances. At the points where these entered the Po, there were small ports and customs offices; the great marketplace was the capital of the kingdom, Pavia. The first traders to sail up the Po were *milites* of Comacchio, bringing salt from the Lagoon to the interior; in 715 they obtained special privileges from Liudprand. The Venetians followed and then merchants from other cities with access to the river system—Cremona, Mantua, Ferrara, Piacenza, Milan, and Pavia.

The development of trade in the Po Valley cannot be understood unless notice is taken of the arrangement of land ownership. This was when the great ecclesiastical holdings were being put together; they had existed before, to be sure, but not in such numbers nor on such a large scale. The many new monasteries founded under the Lombard and Carolingian kings, the donations and legacies that they received, the exemptions and privileges that they enjoyed, the pressures and force of attraction that they exercised in regard to small landowners—all these contributed to enormous episcopal and monastic holdings, unlimited in size and tax free. Frequently, there was a surplus of agricultural production available to meet market demands. In brief, the motivation of the development of trade in the Po River Valley lay in the following factors: bishops and abbots found there the means with which to build and ornament their churches and also to forward their power politics; merchants made considerable profits, as did the government of Pavia and the other cities and potentates from the imposition of tolls and customs fees; new highways of trade were opened up and made operative through the Alps in the direction of central Europe, thereby making the fortune of such favourably situated centres as Milan.

It is not by mere chance that, at the marketplace of Pavia, there were warehouses belonging to the bishops of Lodi, Milan, Piacenza, Reggio Emilia, and, perhaps, Genoa, to the monasteries of St. Ambrose of Milan, St. Antoninus of Piacenza, Bobbio, St. Julia of Brescia, Nonantola, and even to such Frankish monasteries as St. Martin of Tours and Cluny. The operative merchants were not only from Venetia but from Gaeta, Salerno, and Amalfi as well; there were even some Anglo-Saxons among them. This constitutes further proof that the Po Valley was a meeting place of the Eastern and Western economic areas and that, for a time, Pavia shared with Venice the role of intermediary that Venice later played alone.

Secular holdings seem to have had no economic function equal to that of ecclesiastical ones. Probably only the *curtes regiae* had the same acreage. One need only recall that St. Julia of Brescia had 60 *curtes,* more than 700 serfs attached to the *dominicum,* and some 800 farming families to the *massaricium.* The *curtes regiae* did not have the same productivity, since they were made up largely of woods and fallow meadows. As for other secular estates, it seems reasonable to believe that they were, from the start, subdivided and increasingly so because of separate inheritances. The existence of a large fief, or feudal domain, did not necessarily make for economic capacity; lands enfeoffed to the second degree escaped the control of the overall owner. Even if he could count on the yield from widely scattered *curtes,* he could not set up a single management comparable to that of the monasteries.

It must also be noted that during the 9th and 10th centuries a class of peasant was evolving. There was a scarcity of prebendary serfs (*praebendarii*), those worst treated and assigned to the house and land of the master; parts of the *dominicum* were divided into *mansi* and turned over to *casati,* serfs, or *massari.* This made for an increase of the population attached to the soil and for a betterment of its condition, since obligations toward the master were limited and stipulated in writing. Another improved status was that of the *libellarii*—that is, the freemen bound by a written lease (*libellum*). Formerly, these men had been obliged to work the land with their own hands, but, little by little, they won the right to sublease and thus to collect rent. Many of them went to live in the cities. At the very beginning of the 11th century, a rising movement was taking place that was to change medieval society.

Italy under the Saxon and Franconian emperors

THE MAINLAND

The imperial restoration of 962. Toward the middle of the 10th century the Kingdom of Italy was torn by the struggles among the great feudal lords. Upon the death of Lothair II, son of Hugh of Provence (950), Berengar II of Ivrea, who, with his son, Adalbert, already held virtual power, was crowned king. But an opposition, headed by a courageous woman, Adelaide, daughter of Rudolph of Burgundy and widow of Lothair, called for help from Otto I of Germany. Otto, whom Berengar, when an exile in Germany, had already asked to intervene in Italian struggles for power, came to Italy (951), donned the royal crown at Pavia, and married Adelaide. Soon after, by a compromise, Berengar and Adalbert regained rule over Italy, though as vassals of the German king. But the marches of Verona, Friuli, Trent, and inland Istria (all lying across northeast Italy) were given to Henry, duke of Bavaria, to ensure the German king's access to Italy. Some years later, Berengar faced further opposition, and Pope John XII (originally Octavian, son of Alberic, prince of the Romans, and the first pope to change his name) asked Otto to intervene. Otto returned and, in 962, was crowned, together with Adelaide, in Rome. Berengar was soon captured and deported to Germany; later, John XII was accused of betrayal and was deposed, the new emperor replacing him with another Roman, Leo VIII.

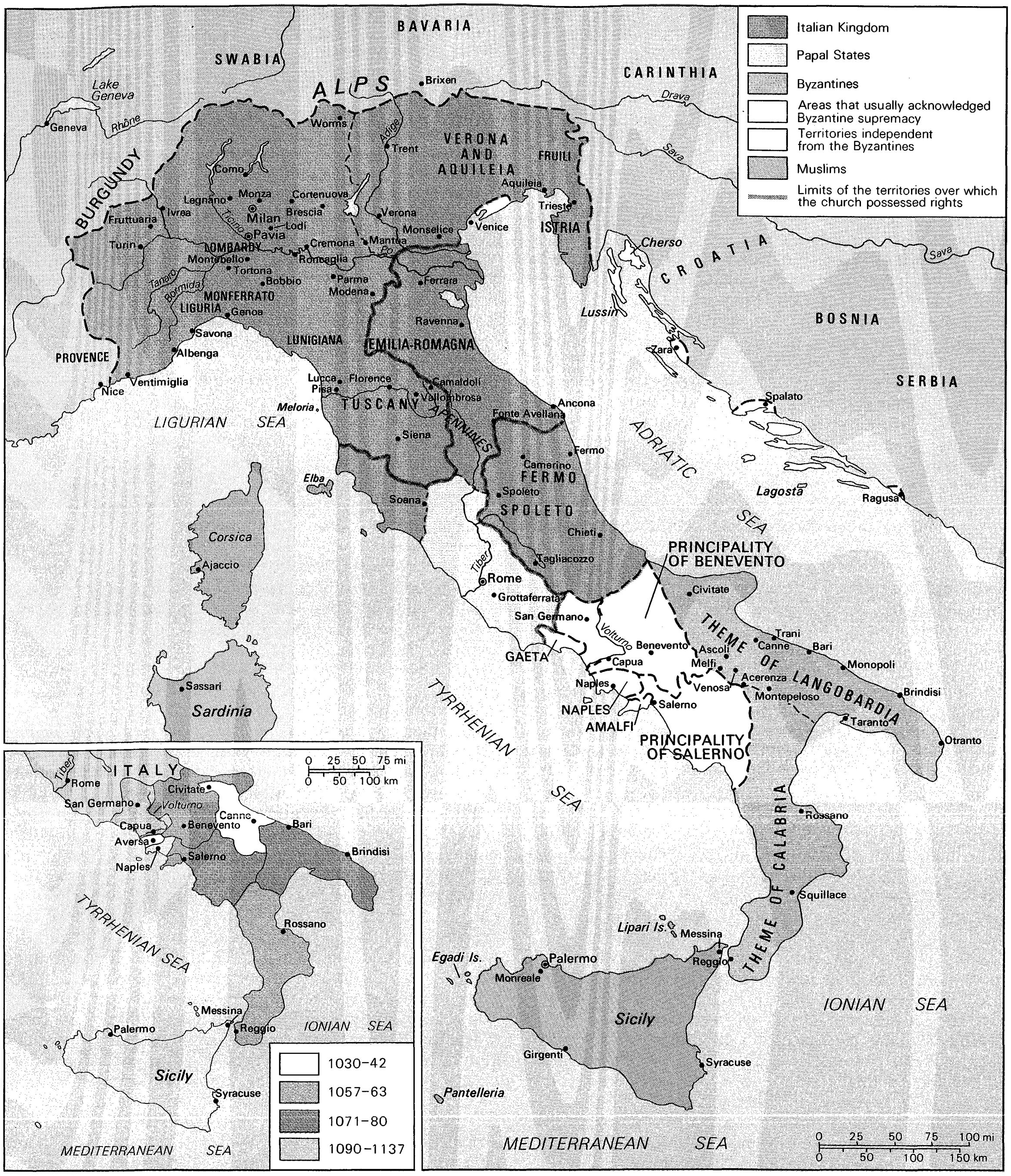
Intervention from Germany

Otto I was obviously appealing to the by now legendary tradition of Charlemagne and trying to restore the Carolingian empire. Otto, like Charlemagne, had won great prestige; he had ruled Germany firmly and, in 955, at Lechfeld (in Bavaria) had defeated the Magyars, preventing their invasion of western Europe; he had ably pursued a policy of penetration and conversion among the Slavs, creating the archbishopric of Magdeburg (in modern East Germany) and a chain of frontier marches from the Baltic down to Bohemia; and he had made himself felt in France, Burgundy, and Italy. Now the time had come to obtain a title symbolizing his rule over Europe.

Here, then, was a superstate, an *imperium plurimarum nationum* ("empire of many nations"); the idea of universality was still of the essence, but it corresponded less to reality than in Charlemagne's day because various countries, especially France, remained outside. From now onward the empire's universality was to be a symbolic aspiration rather than the exercise of power, except in Germany, Italy, and, after 1032, Burgundy, where the emperor held the crown and the *potestas* ("power") of a king. Thus, with Otto I, the empire assumed some of the characteristics that were to remain throughout the Middle Ages. The emperor had first to be the king of Germany, elected by the German princes and crowned at Aix-la-Chapelle (Aachen), whereafter he would have the further title of "king of the Romans" to indicate that he was a candidate for the imperial throne. Then he would be crowned king of Italy (at Pavia—and then at Monza or Milan) and king of Burgundy, after which he would receive the imperial crown from the pope in Rome. This was the lasting structure of what was to be called the "Holy Roman Empire of the German nation." Within this framework there later operated both the new imperial doctrine of Frederick Barbarossa (emperor in the 12th century), intended to strengthen the power and universality of the empire on the basis of Justinian law, and the theories of the Curia (the papal court), which sought to give the pope power to choose and, possibly, also to depose the emperor at will.

Shaping of the Holy Roman Empire concept

For the moment, the most obvious upshot of Otto's actions in 962 and after was the control he had won over the papacy, which went far beyond that of a Charlemagne or a Lothair I. For although the *privilegium Ottonianum* ("privilege belonging to Otto") of 962 confirmed the donations made by preceding sovereigns (except for the area around Ravenna), it also stipulated that the Romans should ask the emperor's approval of a candidate for the papacy and that the newly elected pope should pledge allegiance. Shortly thereafter, Otto assumed the right to nominate the pope—the logical consequence of his policy toward the church. In order to control the feudal lords he sought support from bishops and abbots, giving them fiefs and various privileges. To some bishops, German as well as Italian, he gave the title of count, forming a vassalage all the stronger because it was not hereditary. Thus the ecclesiastical hierarchy became increasingly tied to the



Italy during the second half of the 10th century and (inset) the Norman conquest of south Italy.

Adapted from *Enciclopedia Italiana di Scienze, Lettere ed Arti*, vol. 19; inset adapted from J.R. Strayer and D.C. Munro, *The Middle Ages, 395–1500*, 5th ed., p. 210 (copyright © 1970); by permission of Appleton-Century-Crofts, Educational Division, Meredith Corp.

feudal structure of society and the state and, hence, to temporal interests. The state, rather than the church and its discipline, benefitted; and herein lay the reason for the great conflict that arose between papacy and empire in the next century.

Expansion in south Italy and Otto I's successors. Otto I was also concerned with expansion in south Italy and struck an alliance, aimed against the Byzantines, with Pandulf, known as "Ironhead," prince of Capua and Benevento, to whom in 967 he ceded the duchy of Spoleto and the march of Camerino. His power politics were fruitless, but his diplomacy succeeded, resulting in a marriage between his son Otto II and Princess Theophano, daughter of the Byzantine emperor Romanus II, in 972. After his father's death in 973, Otto II pursued the same expansionist policy in the south, where he claimed new rights through his marriage. Preoccupied by the landing in Calabria (Italy's "toe") of some Arabs from Sicily, he led an army against them in 981. Unfortunately, his ally Pandulf died just then; and Otto, far from his bases, was defeated in 982, at Punta Stilo. He was organizing a revenge expedition when, aged 28, he died in Rome in 983.

Reign of Otto III

His son and successor, Otto III (983–1002), elaborated an imperial ideology mingling elements of the Roman-Byzantine tradition and Christian mysticism. Before Otto III came of age, the uneasy rule of the empire was wisely maintained by his mother, Theophano, and his grandmother Adelaide. The young prince was brought up among the armies guarding the eastern frontier; but he had more than a military education, having been instructed in religion, Greek, and Latin, as well as German, while he even tried writing. This culture, then unusual for a layman, explains his later behaviour. In 996, when he came of age, he went to Rome, named as pope his cousin Bruno of Carinthia (now southwest Austria), who took the name of Gregory V, and received the emperor's crown from him. By nominating the pope, Otto hoped, like his father and grandfather before him, to remove the papacy from Roman factions and restore its universal mission. There was, however, stubborn opposition; a powerful Roman patrician, John Crescentius, ran Gregory V out of the city and put an anti-pope in his place. Otto brought Gregory back to Rome, put down the rebellion, and had Crescentius tortured and killed (998). A year later, when Gregory died, Otto nominated his former teacher, Gerbert of Aurillac, archbishop of Rheims, who took the name of Sylvester II. This was the formative moment of the young emperor's religious and universal ideal. He settled the court in the imperial palaces on Rome's Palatine Hill and restored Byzantine offices and ceremonies. From a Rome that was again the centre of the empire and the *caput mundi* ("head of the world"), emperor and pope, in perfect accord, were to eliminate abuses, reform the clergy, and extend Christendom. The peoples of eastern Europe came under this plan, and it was now that the duchies (later kingdoms) of Poland and Hungary entered the orbit of the Catholic Church.

Otto III acquired intense but tormented religious feelings from such spiritual leaders as Adalbert of Prague, Nilus of Rossano, Odilo of Cluny, and Romuald of Camaldoli; but his political and religious dream was even shorter than his 22-year life. In northern Italy he had to face one Arduin of Ivrea, who was struggling against the local bishops and had executed one of them; in Rome, a new rebellion enforced Otto's flight, and he died in 1002 before he could return to the city.

Italian unrest. After Otto's death, Arduin had himself proclaimed king of Italy in 1002 by a hastily assembled group of nobles. But his adversaries, many of them bishops, produced a rival, Henry II, duke of Bavaria, of a collateral branch of the House of Saxony. Henry overcame Arduin in Italy in 1004 and was in his turn crowned king of Italy at Pavia, though a fight between local citizens and the German soldiers considerably damaged the city. Henry II immediately returned to Germany, leaving Arduin some freedom of action. He came back in 1013–14 to receive the imperial crown from Pope Benedict VIII; and in Rome, too, a violent anti-German reaction occurred, with Arduin's supporters participating. The failed revolt and other circumstances persuaded Arduin to retire to the monastery of Fruttuaria, where he died in 1015. Henry II returned for the third and last time in 1021. In agreement with Pope Benedict VIII, he concentrated on southern Italy, where shortly before had occurred a revolt, led by one Melus of Bari, against the Byzantines. Henry died in 1024 in Germany, whereupon the citizens of Pavia destroyed the royal palace.

Revolts against foreign rule

Feudal and ecclesiastical developments. The real significance of Arduin's action is in representing the reaction of feudal laymen to excessive episcopal power. Arduin had the support of the *secundi milites,* the minor vassals or vavasours, seeking better positions in episcopally controlled administrations. At first Henry II quite naturally sought episcopal support, installing trustworthy German clerics, some of them court chaplains or even members of his family, in such northern Italian sees as Como, Cremona, Trieste, Lodi, Turin, and Ravenna. Later on, under pressure from below, he tended to favour the vavasours and the *cives* ("citizens") and to limit the bishops' privileges.

Between the late 10th and the early 11th centuries several large marches, governed by powerful noble families, often originating north of the Alps or bound by relationship or interests to families of Provence, Burgundy, and Germany, came into being. These marches were not organic territorial units, made up of a certain number of counties; for cities had acquired some autonomy under episcopal rule, and in the countryside privileged ecclesiastical properties and new feudal or landed estates were forming. In any case, the noble families of the marches were powerful through having their own domains, even if scattered and not adjacent to each other. Only toward 1200 did great families tend to subdivide branches, each linked to a locality whence it took its name. For instance, the Obertinghi, descendants of a Count Obert (who acquired status first with Berengar and then with Otto I), possessed, besides their own march (composed of Genoa, Luni, Bobbio, and Tortona), the counties of Milan and, in the 12th century, Cavallo and Monselice. The descendants of the Obertinghi include the Estes, the Malaspinas, and the Pallavicini.

The division of the great families

The march of Verona, which included the whole of Venetia, Trentino, Trieste, and inland Istria, had a particular status, being part of Germany. Meanwhile, within it the ecclesiastical principalities of Trento, Aquileia, and Trieste were forming. Within the Kingdom of Italy were the new marches: Ivrea (or Anscarica, from the name of its local dynasty); Turin (or Arduinica), which included western Piedmont all the way to Ventimiglia and Albenga on the present Franco-Italian border; western Liguria (or Aleramica), later divided into the two marches of Monferrato and Savona; eastern Liguria (or Obertenga), extending southward from the Po and Tanaro rivers to include Genoa and the coast southeast of it all the way to Luni; and the march of Canossiana (or Attoniana), which extended from the Apennines near Modena and Reggio Emilia to beyond the Po and the vicinity of Mantua and Brescia. Besides these new marches were older ones such as Tuscia (Tuscany), soon to be joined to the Canossiana. Theoretically dependent upon Tuscia was Corsica, which actually was abandoned to Saracen invasions, but which in the 11th century became an independent republic called the Terra di Comune (Land of the Commune). There were also the duchy of Spoleto and the march of Fermo-Camerino (eastward from Camerino to the coast).

The House of Franconia and the struggle over investitures. The insurrection of Pavia in 1024 and the destruction of the *palatium* ("palace") had grave consequences. The authority of the Count Palatine was lost, and the royal court and the central administrative offices disappeared. With even more important results, so did the Royal Chamber, which had rigidly controlled the city's craft guilds and collected taxes and donations. Under milder episcopal government, citizens enjoyed greater economic and political freedom. The greatest advantage accrued, however, to Milan, which practically escaped from the domination of the nearby capital and was able to expand. The people of Milan seemed to be unified under their bishops' guidance and to have political aspirations, based on awareness of the city's economic development and on pride in their religious tradition, the Ambrosian rite.

A great Milanese archbishop, Heribert of Intimiano, set the trend of Italian politics in the critical period that followed the death of Henry II. He and his followers called upon the Duke of Franconia, later known as Conrad II, a Salian, who founded a new imperial dynasty. Conrad was crowned king of Italy at Pavia in 1026 and emperor in Rome the year after. Present at this second ceremony were two sovereigns: Canute (Knut), king of England, Denmark, and Norway; and Rudolph III, king of Burgundy, who thus bore witness to the prestige of the imperial crown.

Feudal upheavals. Conrad II rearranged the major Italian fiefs. First he joined together the marches of Attoniana and Tuscia, an exceptionally large and important territory destined to play a decisive role in future events, under the rule of the faithful margrave Boniface of Canossa. Then, in keeping with the German emperors' perennial need to control all the roads giving access to Italy, he detached the bishopric of Trent from the march of Verona and joined it to the duchy of Carinthia, establishing at the same time direct imperial rule over the patriarchate of Aquileia, which acquired a more clearly defined territorial and feudal identity and was governed by German patriarchs.

Redeployment of localities

In 1032, after the death of Rudolph III of Burgundy, Conrad II inherited his kingdom but had to make good his claim by military occupation. In this successful process (which led to his coronation as king of Burgundy in 1033) he was aided by an Italian contingent led by Archbishop Heribert and Boniface of Tuscia, together with a minor Burgundian noble, Humbert the Whitehanded, ancestor of the House of Savoy. The expedition had still other results. Upon their return to Lombardy the lesser vassals, or *secundi milites,* who had undergone considerable hardships and financial sacrifices in order to follow Heribert, revolted against him.

This rebellion, the first of many social upheavals, had its roots in the time when Landolph of Carcano was archbishop of Milan (983) and many properties of the Milanese church were enfeoffed to his supporters and members of his family. This made for a powerful and privileged class, called the *novitii capitanei* to distinguish them from others with powers derived from the breakup of the possessions of the counts, who took in the tax (*decima*) paid by country parishes and held jurisdictions (*districtus*). In their turn the *capitanei,* by sub-enfeoffing, had created a class of lesser vassals, bound to military service but without any hold on the land. These *secundi milites* had grown in numbers and, beginning with the time of Arduin, had been demanding a definition of their rights and possessions. Inimical to the *secundi milites,* or vavasours, were not only the *capitanei,* and hence Heribert, but also the *cives,* that is, the merchants, money changers, magistrates, doctors, and notaries in the city who owned nonfeudal houses and lands. The *cives* had, indeed, interests quite separate from those of the vavasours; and they saw Heribert as a man able to increase the city's political power and forward its economic development. The vavasours, on the other hand, found support in the rural nobility of the counties of Martesana and Seprio, which was already in conflict with the *capitanei* and Heribert; in the cities pitted against Milan, such as Pavia, Lodi, and Cremona; and, to some extent, in the *rustici,* that is, the peasants, who were equally oppressed by feudal holders of *districtus* and by allod (nonfeudal property) owners and lessees (*livellari*), who did not themselves till the land.

Unexpected results of intervening in class conflict

Such was the array of forces when Conrad intervened in behalf of the vavasours. With the Constitutio de Feudis (Constitution Concerning Fiefs) of 1037 he established the rule that no vavasour could be deprived of his fief without a sentence passed by his peers. A fief, moreover, was entailed to the owner's children or nearest relations. This was a radical change in an imperial policy that heretofore had favoured the bishops. But the results were not what Conrad had expected. The vavasours, once they were pacified, made peace with the *capitanei,* while, on the other hand, Heribert and many other bishops were antagonized. Conrad faced up to the new situation by striking an alliance with the secular lords. Then he returned to Germany, dying there in 1039.

Henry III and the ecclesiastical reform movement. Conrad was succeeded by his son, Henry III, who for several years was involved in a struggle against Bohemia and Hungary and had to neglect his Italian affairs. Meanwhile, a civic revolt took place in Milan. The *cives,* who had stretched their muscles in previous episodes of a similar kind, drove Heribert, the *capitanei,* and the vavasours out of the city. Then for three years (1042–45), under the leadership of the noble Lanzone, they stood off the siege of their adversaries. Eventually there was a reconciliation; *capitanei* and vavasours came to terms with the *cives* and returned to the city. It was no commune, but there were the bases of a commune's foundation: the deposition of the archbishop and the solidarity of the three main social classes, with the idea that there must be cooperation among them.

Henry III had to deal with circumstances very different from those familiar to his father, but he showed remarkable ability in adjusting himself to them, particularly in the understanding with which he treated the *cives* of Milan, Ferrara, and Mantua. He realized that the bulwark they provided against the encroachments of the bishops and their feudatories might strengthen the empire. At the same time he assured himself of the loyalty of the bishops, especially those whose sees straddled the roads to Germany, such as Como, Trent, and Verona, or which were the centres of large territories, such as Aquileia and Ravenna. To these sees and to the rule of various cities of Venetia he named trusted Germans. As a successor to Heribert in Milan he chose a country cleric, extraneous to the cathedral hierarchy, which was linked to the families of the *capitanei* and expected to furnish a candidate for the archbishopric. Obviously Henry shared his predecessors' rigid concept of the supremacy of the emperor over the bishops. But one cannot fully understand his policy unless one takes into account certain other local conditions that influenced him, above all the religious revival of the 11th century. For Henry was sensitive to demands for clerical reform and considered that he must eliminate abuses and make the church respectable.

Reorganization of the monasteries

In Italy there was much religious ferment. Already in the 10th century the Cluny reform had reached Italy, resulting in the founding or re-ordering of monasteries at Pavia, Rome, Fara Sabina, Pomposa (Ravenna), Fruttuaria (Piedmont), and Cava dei Tirreni (Naples). But the Cluny reform had lost its initial impact, and its strictly monastic ideal was not broad or strong enough; everywhere were complaints about unworthy prelates, the church's temporal interests and submission to secular power, and the clergy's practice of simony (traffic in ecclesiastical preferment) and concubinage. On the one hand, a reaction occurred in the rise of religious orders of hermits and cenobites founded by Romuald of Ravenna (at Camaldoli), Peter Damiani, his disciple and biographer (at Fonte Avellana), and John Gualbert (at Vallombrosa), all of them fired by asceticism; on the other hand (but a little later, beginning in 1057), there was a violent popular protest, represented chiefly by the reform party known as the Patarines of Milan and Florence, against bishops linked too closely to the empire and against the unseemly behaviour of the clergy. This latter movement was strictly secular and worked from below, its religious motivation having economic factors. The promoters were unwilling to leave to the clergy the profitable administration of the church's wealth, which they felt should be used by the whole community. In southern Italy there was a particular movement of reform of the monastic rule of the Basilians, which was inspired by Nilus of Rossano, founder of the monastery of Grottaferrata.

Henry III, in contact with the chief monastic reformers—Odilo of Cluny, Peter Damiani, Alinard of Lyons—favoured the reformed congregations and influenced the choice of bishops and even popes. He took drastic action on his first visit to Italy, in 1046: a schism in Rome had led to the election of three popes, and Henry voided all three elections and named a man of high moral calibre, Suitger, bishop of Bamber, in Bavaria, Germany, who took the name of Clement II. Subsequently he named three other reputable German bishops: Damas II, Leo IX, and Victor II.

The choice of Leo IX (1049–54) was of especial importance. As Bruno, bishop of Toul, in northeast France, he belonged to a noble Alsatian family related to that of the Emperor and had close connections with the reformers of Lorraine. No sooner was he elected than he undertook an energetic policy of reform. In Rome and Pavia, and later in France and Germany, he presided over a succession of councils that condemned the worst abuses—simony, concubinage, and the usurpation of ecclesiastical benefits. Unworthy prelates were deposed and there was a restatement of the laity's obligation of tithing. Such policies were extended even to southern Italy, where the situation was complicated by the presence of Normans and Byzantines. On various occasions the Pope and his supporters called attention to the necessity of respecting Canon Law in regard to the election of bishops. Leo himself gave the example when, even after the Emperor had chosen him, he insisted upon being regularly elected by the clergy and people of Rome. The reformers brought equal energy to bear on their support of the principle of the authority of the see of Rome over the universal church. Their defense of this principle came to a climax during the conflict with the patriarch of Constantinople, Michael Cerularius, which led to the definite schism between the Eastern and Western churches in 1054. The situation was changing: popes were no longer figureheads produced, amid intercity strife, by the counts of Tusculum; they were foreigners, aware of their worldwide responsibilities and bolstered by the emperor's power. The initiative of reform passed into their hands while Henry III was detained in Germany and after his death (1056), when his son Henry IV was a minor and the empire was in a long period of weakness. As the popes recaptured the leadership of Western Christianity, they perceived that the first step toward reform was the elimination of secular interference with church govern-

Incipient clash between emperor and pope

ment—hence the inevitable conflict with imperial power and the later "struggle over lay investiture." With Leo IX the clash was only latent; it did not break out until the time of his successors. Eventually the emperors' weapon of reform was turned against them.

Among Leo IX's successors Nicholas II (1059–61), of Burgundian origin, deserves special mention. At the Lateran synod of 1059 he decreed that from then on the pope should be elected by the cardinals, approved by the rest of the clergy, and acclaimed by the people, while the emperor could only give a generic assent after the election was over. By this decree the choice of the pope was removed not only from the emperor but from the Roman factions as well, and the form that it took then has since endured. This undermining of the imperial prerogative was possible only because Henry IV was not yet of age. And, as substitute for the German emperor's support, Nicholas II made, also in 1059, an agreement with the Normans, naming Robert Guiscard duke of Apulia and Calabria and receiving his oath of allegiance.

Further erosion of imperial power. After Nicholas' death, events followed a logical development. On one side there was the reform party, headed by the new pope, Alexander II (born Anselm of Baggio), representative of the intransigent and often violent popular movement of the Patarines of Milan. This party could count on such fiery preachers and debaters as Peter Damiani and Humbert of Silvacandida and on the interested support of the Norman princes. Against it stood a "conservative" party, composed of members of the imperial court, German bishops, and their Italian fellows linked to the empire and its traditional policy. This latter group had many "clients" and dependents, especially in Lombardy, Venetia, and Romagna, roughly, the area between Geora and Ravenna. Originally it refused to recognize Alexander II and put up an anti-pope, Cadalus of Parma (Honorius II), who was deposed in 1064.

In 1073 Gregory VII (formerly Archdeacon Hildebrand of Sovana), a trusted counsellor of his predecessors and the soul of the reform movement, became pope. He was intensely moral and uncompromising, determined to attack abuses at the root, and in 1075 forbade, under penalty of excommunication, any secular power to concede the investiture of abbeys or bishoprics. This signified a definite negation of the hierarchical feudal system upon which the empire and its single states were based and affirmed the church's complete independence from lay domination. It was a revolutionary act, to which Henry IV replied, after some hesitation, by the Diet of Worms of 1076, which branded the Pope as unworthy and illegitimate and removed him from the throne. This was the beginning, on the basis of the political differences outlined above, of the investiture struggle, which had its beginnings in Italy but soon spread to the rest of Europe.

The humiliation of Canossa

Gregory VII made a rapid and effective reply. He excommunicated the Emperor, released his subjects from their allegiance to him, and thus favoured a revolt of the German princes. Henry IV found himself in a critical situation and came precipitately down to Italy with a small escort in order to obtain the Pope's pardon and removal of the excommunication. The meeting took place in February of 1077 at the castle of Canossa, where Gregory VII was the guest of Matilda, countess of Tuscany, a strong supporter of his policy. Henry had to humble himself to receive absolution, but the danger temporarily blew over. There was a continuation of the crisis, however, in 1080, when the Pope once more excommunicated the Emperor, relieved his subjects of their vows, and recognized Rudolf of Rheinfelden, long since chosen by the German princes, as German king. After this, events became favourable to Henry. At a council meeting at Brixen, where many bishops from northern Italy were present, he had an anti-pope elected (Guibert, archbishop of Ravenna, who took the name of Clement III) and defeated Rudolf of Rheinfelden in a battle in which the latter received wounds that led to his death. In 1081 Henry returned to Italy with an army, defeated the militia of Matilda of Canossa, and for three years besieged Rome. In 1084 he was crowned emperor by the anti-pope but soon after was driven from the city by Robert Guiscard and his Normans, who had come to rescue Gregory VII, who died at Salerno a year later.

The struggle was resumed, after a brief interval, by the next pope, Urban II (1088–99), of French origin. He was supported by Matilda's militia and won success for the reformist, or "Gregorian," party among the cities and bishops of the north. Henry's last resistance was overcome, and, frustrated and embittered by the rebellion of his own sons, he died in 1106.

Henry V, his son and successor (1106–25), and a new pope, Paschal II (1099–1118), tried to settle the controversy peacefully. The Pope proposed the abolition of imperial investiture and, in return, the clergy's renunciation of fiefs and benefices. The idea was Utopian because too many private interests were involved, and, indeed, the German bishops bitterly opposed it. Tension continued, complicated by the fact that Countess Matilda, upon her death in 1115, left all her properties—both fiefs and allods—to the Roman Church. The former, because they had been granted by the empire, should rightfully have returned to it; hence another conflict.

The investiture question finally reached a compromise solution in the Concordat of Worms (1122), drawn up between Henry V and Calixtus II. It was stipulated that the emperor should no longer concede a religious investiture with the symbols of ring and staff. In Germany, an elected bishop or abbot was to receive from the sovereign only temporal investiture (with a sceptre), which would be followed by religious consecration; in Italy, this consecration was to precede temporal investiture and the latter was to take place without the sovereign's presence.

Beginning of the papacy's paramountcy

After 70 years of bitter struggle the system remained substantially the same as before. Yet great social and political changes had taken place. First of all, the Church of Rome had taken the place of the empire as the pilot of Western Christendom; the most obvious proof is the First Crusade, an imperial sort of enterprise but eminently the work of Pope Urban II. The prestige of the Roman see was very high; and the Normans of southern Italy, the Christian states of Spain, England, Hungary, Croatia, and even the princedom of Kiev recognized in it some sort of feudal superiority and had recourse to it for protection. Within the church itself there was a radical transformation. The necessity of closing ranks in the struggle against secular interference caused the popes, in particular Gregory VII, to strengthen the central authority, that is, that of the apostolic see over the bishops. This determined what was to be for centuries the structure of the church, rigidly hierarchical and centred in the papal Curia, with a clergy to which all sacramental prerogatives and disciplinary powers were entrusted and a laity that had no more than a passive role. At the same time a new, theocratic theology was developed. The temporal was subordinated to the spiritual; and it became the popes' right and duty to intervene in such events as the election or deposition of an emperor, conflicts among states, and violated oaths that imperilled peace among Christians and their souls' salvation. Gregory VII was the great artificer of these ecclesiastical and political doctrines, which inspired the project of collecting and defining the articles of canon law.

Another important aspect of this period was the impulse given to civic life and liberty, especially in central and northern Italy. The investiture struggle was actually not a mere political and economic conflict between two great powers. It had also an ideological content, which aroused the participation of the masses. Popes and emperors vied for the citizens' support by the concession of exemptions and privileges. In the cities and fortified castles, groups, alliances, and various *conjurationes* (agreements) were constantly forming. Writers distributed polemical pamphlets discussing such matters as the validity of a sacrament administered by an unworthy priest, the sale of church offices, the grant of contracts in return for political favours, resistance to authority, people's rights, the *libertas ecclesiae* ("freedom of the church"), the articles of Roman law, and laymen's participation in religious life. A world was in ferment, opening up to new ideas and experiences acquired by living together. The most important end result was to be the establishment of the communes.

NORMAN SICILY

The Normans' arrival. Small groups of Normans arrived in southern Italy in the early 11th century. They were adventurers and skilled men-at-arms, seeking their fortune in a smiling land and one divided into so many small, conflicting states that it was easy to enlist in the service of one or another. Gradually the Norman leaders got themselves lands of their own and settled down, pursuing at the same time a policy of expansion. After 50 or 60 years the newcomers constituted an important political and military power, which proceeded to place southern Italy and Sicily under its rule.

In 1017 Pope Benedict VIII called upon a contingent of Norman knights to support the revolt of Melus of Bari against the Byzantines. Subsequently, other Norman mercenaries took part in the wars between Naples and Capua. Sergius IV, duke of Naples, in order to win over the Norman leader, Ranulph Drengo I, gave him his sister for a wife and made him count of Aversa (1030). This first territorial acquisition was followed by others in Apulia and its vicinity, effected by another Norman group headed by the Hauteville brothers (William "Iron Arm" the first among them). The cities and towns in question were Ascoli Satriano, Venosa, Lavello, Monopoli, Trani, Civitate, Canne, Montepeloso, Acerenza, and, above all, Melfi. The possession of these places was legalized by feudal investitures granted by Prince Gaimar IV of Salerno in 1042. Only four years later Henry III gave imperial investiture to Raymond of Aversa and Drogo of Hauteville, brother of William. Probably this was the confirmation of preceding investitures conceded by Gaimar, together with recognition of the superiority that Drogo had asserted over all the other Norman knights who had settled in Apulia.

Resistance of Leo IX to the Normans

Actually, the continuous state of war favoured the concentration of power in a single hand, such as that of Humphrey, successor to his brother Drogo, who came up against the hostility of Pope Leo IX. This pope is important not only in the history of the reform of the church but also for his policy in southern Italy. Fearful of the Normans' growing power, as it threatened the city of Benevento (35 miles northeast of Naples), which he had recently added to his possessions, he strove to bring about an alliance between the two emperors—German and Byzantine—against the Normans. Failing, he went in person to war but was defeated and taken prisoner in the battle of Civitate in 1053. Humphrey had the intelligence to give him honourable treatment and to set him free. Thus, there came into being an agreement between the papacy and the Normans, furthered at Melfi in 1059, when Robert Guiscard, Humphrey's younger stepbrother and successor, pledged allegiance to Pope Nicholas II and assumed the title of *dux Apulie et Calabrie et futurus Sicilie* ("duke of Apulia and Calabria and the future duke of Sicily"). The same pledge was made by Richard of Aversa on behalf of the principality of Capua. These were acts of great historical importance, inasmuch as the Normans legitimized their conquests, past and future, and the popes established their feudal sovereignty over all of southern Italy and Sicily, thus making concrete a long-standing aspiration to political and religious control there and establishing a diplomatic and military base for the coming struggle over investiture.

Norman expansion. Robert Guiscard and his younger brother Roger (the "Great Count") rapidly extended their conquests, beginning with Calabria. In 1060, Roger captured Reggio and soon after Messina, which served as a beachhead for the conquest of Sicily. The island was divided and politically weak, but the Arabs put up a stubborn resistance, which was not overcome until 1091. Meanwhile, on the mainland, just above Italy's "heel," Bari (1071) and Salerno (1077) fell into Norman hands; and the whole territory was united under the rule of Guiscard. Only the Norman principality of Capua-Aversa and the Byzantine duchy of Naples kept a certain autonomy, the former until 1156, the latter until 1137.

In conquering Sicily, the "Great Count" Roger acted as vassal and representative of his brother Robert and hence with a certain de facto independence of the Church of Rome. Moreover, he staved off the formation of too-powerful feudal lordships such as the Norman knights had set up in the former Byzantine and Lombard territories, keeping a large part of the conquered lands and using viscounts and army leaders—officials directly dependent upon him—to govern the island. In short, Count Roger managed to create a centralized and efficient governmental structure linked only in name to the duchy of Apulia and its feudal superior, the papacy. It further happened that, in 1098, Pope Urban II conferred upon Count Roger and his successors an "apostolic delegation," by which they were legally recognized as the pope's representatives for Sicily's ecclesiastical affairs. This amounted to official recognition of the power that Roger had acquired over the Sicilian churches during and after the conquest. The situation was very different from that of 1059, when Robert Guiscard had to recognize the pontiff's jurisdiction over the churches of Apulia and Calabria.

The importance of this became obvious when Roger II (son of the "Great Count") brought about the unification of the Norman territories (1127) and created the new Kingdom of Sicily, with its capital in Palermo (1120). This kingdom was born as a solidly organized and centralized state, with an efficient bureaucratic administration that had authority over both clerics and feudatories. The only thing that Roger II did not completely eliminate was the feudal overlordship of the papacy, which he sought to reduce to a mere formality.

The authority of the Norman king

Norman administration. The Norman rule of southern Italy and Sicily has a historical importance comparable to that of Britain (set up at the same period) for its solidity and duration. The king's authority was conceived as absolute and deriving directly from God, a concept influenced by that of Roman–Byzantine autocracy, as can be seen in the body of law of the assizes of Ariano, applicable to the entire kingdom, put out in 1140. The sovereign was assisted by a *magna curia* composed of the top officials of the kingdom, the princes of royal blood, and the chief prelates and feudatories. In spite of the presence of the last-named category, the curia did not represent the feudal class. It was open, on the grounds of position or career, to men of diverse social ranks and conditions, primarily Normans (and later Franco-Normans and Anglo-Normans), then Italians, Greeks, Lombards, Arabs, and Jews. The major officials were five in number: the Seneschal, in charge of everyone connected with the court; the grand chamberlain, who watched over finances; the chancellor; the "protonotary," head of the notaries; and the admiral, leader of both land and sea forces. The very titles have mixed Norman, Roman–Byzantine, and Arab origins.

As for the ordering of the provinces, the kingdom was divided into circumscriptions ruled by justiciaries and chamberlains, the first having administrative functions and feudal and penal jurisdiction, the second controlling finances and civil lawsuits. Lands belonging to the royal domain and lands enfeoffed both came under the provincial administration; obviously the barons' immunities were quite limited. Equally limited were those of the bishops. The Byzantine imperial–papal model, joined to the traditions of the Norman conquest of Sicily and the privilege of apostolic delegation, had resulted in a rigid subordination of clergy to the state. Feudatories and bishops, tightly controlled as they were by the sovereign, took part in the great assemblies of the kingdom (*curiae generales* or *colloquia*), called together for the proclamation of laws, though they played no active role except at moments of crisis for the monarchy (as in 1189).

Status of the cities

The cities, especially those of the Apulian and Campanian coasts, had by 1000 achieved some autonomy and economic development. Gradually they were incorporated, through treaties and agreements, into the Norman state and legally passed into the royal domain. They were considered mainstays of the governmental structure, but this was too rigid and authoritarian to allow them their former independence. Their autonomy was lost, except for brief intervals, until the statutory revival of the *universitates* (the whole number of citizens) at the end of the Angevin and the beginning of the Aragonese period (in the 16th century). The domain cities entered the *curiae generales* only in 1208 and were permanently incorporated from 1232.

The Norman dynasty continued after Roger II by direct descent, with William I ("the Bad") and William II ("the Good"). When the latter died, leaving no heirs, there was a violent struggle between a nationalist faction that proclaimed the illegitimate royal prince Tancred king and the supporters of Constance, daughter of Roger II, who had married Henry VI, son of Frederick Barbarossa. The latter won, and the Norman crown passed into the hands of the House of Swabia.

The Norman rule over Sicily was not only the embodiment of an orderly and strong government; it provided a place for the meeting and fusion of different traditions, which together produced a vigorous and original civilization. The liberal cultural tradition of the Arab emirs was continued by Roger II. What is left of Arab–Norman architecture (S. Giovanni degli Eremiti, the Zisa, and the Cuba of Palermo, northwest Sicily; and the cathedrals of Monreale, southwest of Palermo, and Cefalù, on Sicily's north coast) and of Byzantine mosaics (in the Palatine Chapel and La Martorana of Palermo and the cathedral of Cefalù) indicate a high level of civic life and culture.

(Gi.Ma.)

ITALY AND SICILY IN THE 12TH AND 13TH CENTURIES

The Hohenstaufen emperors. The conflict (mentioned above) between the empire and the papacy began soon after Frederick I Barbarossa, the second German king of the Hohenstaufen dynasty, had been crowned emperor (1155). The decrees of Roncaglia (1158), issued at the beginning of his second Italian expedition, with the assistance of members of the new Bolognese law school, placed imperial policy on a new basis. Frederick's predecessors Lothair II (or III) the Saxon and Conrad III had only rarely intervened in the affairs of northern and central Italy. Conrad III never came to Italy at all. By the time of Frederick Barbarossa's first expedition (1154–55) the communes had greatly increased their power, and no organized attempt had been made to check them. Frederick tried to strike at the core of the problem by claiming for the empire the royal rights (regalia) that had in so many cases been usurped by the communes.

The real victims of this usurpation, however, had been the bishops and the nobility, who for a long time had been in possession of most of the regalia. The innumerable grants of royal rights had formed an integral part of the traditional system of imperial government. Frederick intended to replace it, to a wide extent, by one of direct control. Imperial officials were to administer town and countryside; and if the communes were to be left some of their liberties, they were to owe them entirely to the emperor. In fact, the Emperor began soon to differentiate between the communes; those that took his side were granted considerable concessions, such as free elections of the consuls. The violent resistance of Milan and other towns was largely responsible for this turn of Frederick's policy. It was significant that Milan's enemies, such as Lodi and Cremona, vigorously assisted in the siege of that town and shared in its destruction in 1162.

The Veronese and Lombard leagues

On the other hand, the revival of the struggle between empire and papacy provided the anti-imperial towns with a powerful ally. The double papal election of 1159 led to a schism; and while Frederick recognized Victor IV, Alexander III was prepared to support any communal reaction against the Emperor. After the fall of Milan, Venice, threatened by the extension of imperial power, had taken the initiative in founding the Veronese League (1164); and in 1167 a second league was formed between several Lombard towns. In the same year, the two leagues joined and entered into close contact with Alexander. In the meantime, Frederick had taken Rome; but an epidemic had decimated his army (1167), and his return route to Germany was almost cut by the Lombard revolt (1168). As a result, imperial authority practically collapsed in Lombardy and was much weakened in Tuscany.

After his return to Italy, Frederick was prepared to give up the execution of the Roncaglian decrees, as was shown in the negotiations of Montebello (1175); but these negotiations were ineffective, and in 1176 Frederick was defeated by Milan near Legnano. In 1177 Frederick concluded the separate peace of Venice with Alexander III, and an armistice with the Lombard League was culminated after six years by the peace of Constance (1183). The Lombard communes were left the regalia inside the towns and, on certain conditions, in the territory, and they retained nearly all their liberties. But the consuls were to be invested by the emperor (a right of which he seems hardly ever to have made use). The peace, however, concerned only Lombardy. In Tuscany, Spoleto, and the Marches, the imperial position had been greatly strengthened during the previous years; and the administrative reorganization that had failed in Lombardy was carried out with considerable success in that region. Moreover, Frederick was preparing the extension of German rule to the kingdom of Sicily.

Imperial ambitions in Sicily

The emperors had made repeated attempts at conquering the Norman kingdom. In 1137 Lothair II had achieved a short-lived success in Apulia; in 1166–67 Frederick had planned an attack on the kingdom. In 1184 the Sicilian heiress Constance, King Roger's daughter, was betrothed to Frederick's son and successor Henry, an event of far-reaching importance for Italy and the empire.

At Frederick's death in 1190, Henry VI had already begun to assert his and his wife's claims to Sicily, King Roger's grandson William II having died in 1189. But resistance led by Tancred, an illegitimate grandson of King Roger, was strong; and it was not until 1194 that Henry succeeded in conquering the kingdom. He left the Norman regime unchanged; the highly centralized Sicilian state was an invaluable addition to the resources of the empire in Italy.

Henry's unsuccessful plan to make the empire hereditary rather than elective would have led to permanent union, since the Sicilian crown was hereditary. Placing imperial rule on a stronger basis than it had ever possessed, it would have been fraught with dangers for the political independence of the papacy. Thus the papacy made a determined effort, after Henry's death in 1197, to destroy the union of Sicily and the empire.

Henry's death was followed by a disputed imperial election; in Italy imperial administration disintegrated rapidly, the communes recovering everywhere what they had lost. Pope Innocent III (1198–1216) took full advantage of this reaction and substituted papal government for imperial administration in the duchy of Spoleto and in the march of Ancona, thus once more extending the Papal States to the Adriatic Sea. Constance of Sicily renounced the imperial crown on behalf of her son Frederick, the new king of Sicily, and appointed Pope Innocent III to be his guardian after her death (1198). The next step in the separation of Sicily from the empire was the renunciation of that kingdom by the German king Otto IV (1201), a Welf who had been recognized by Innocent against the Hohenstaufen Philip of Swabia. But after his imperial coronation in 1209, Otto turned to its conquest, the continuity of imperial policy proving stronger than promises and a change in dynasty.

In Sicily, the preceding years had been marked by internal disorders that threatened to destroy the work of the Norman monarchy. In 1211 Pope Innocent decided to play off the young king Frederick against Otto by supporting his new election as future emperor in Otto's stead. Crowned as German king at Mainz (1212), Frederick II grew rapidly more powerful; and the Battle of Bouvines (1214) sealed Otto's fate. But the success of papal policy was only temporary, for Frederick did not keep his promise to separate Sicily from the empire.

Between 1220 and 1250 Frederick continued with great vigour and much success the policy of his German and Norman ancestors in Italy and Sicily. The years after his imperial coronation in 1220 were devoted mainly to rebuilding and consolidating the structure of the Norman monarchy. Further reforms took place in later years, and the Constitutions of Melfi of 1231 admirably reflected the spirit and working of the highly centralized and bureaucratic Sicilian state. In northern Italy, active intervention began in 1226 and immediately led to the revival of communal resistance; and when in 1227 the new pope, Gregory IX, excommunicated Frederick after his failure to keep the date appointed for the crusade, the pattern of the reign of Frederick I was reproduced, the papacy allying itself with the Lombard towns against the Emperor.

The Peace of San Germano (1230) between Pope and Emperor was of short duration. Open war with the Lombards broke out again in 1236; the great defeat of the league at Cortenuova (1237) was not fully exploited by Frederick; and in 1239 Gregory once more excommunicated the Emperor. The capture of Rome then became a major objective; Frederick may have been on the eve of attaining it at the time of Gregory's death (1241). In 1244 peace negotiations with Innocent IV and the league broke down, and Innocent deposed the Emperor at the Council of Lyons (1245). In the following years the struggle continued with unprecedented violence. Despite numerous setbacks, the imperial cause seemed to be in the ascendant when Frederick suddenly died in 1250.

Changes in mainland government

After 1237 he had reorganized imperial administration in northern and central Italy, introducing and adapting Sicilian methods of government. Vicars general with wide authority governed new provinces, while under them local officials administered towns and countryside. The preponderance of natives of the Sicilian kingdom in the Italian administration significantly contrasted with the leading role played by Germans under Frederick I and Henry VI. Sicily had become the main pillar of imperial rule in Italy, and its great resources held out a substantial promise of success. At the same time, Frederick tried to preserve the support of the German princes by far-reaching concessions (1232).

Frederick's death was a turning point in the history of Italy; it marked the end of the Hohenstaufen policy of placing the country under a centralized monarchical government. The conflicts between Frederick's successors and the papacy continued until 1268. In 1265, Pope Clement IV invested Charles of Anjou with the Sicilian kingdom, where King Manfred, Frederick's illegitimate son, had been consolidating his power. In 1266, Manfred was defeated and killed in the Battle of Benevento. Two years later, Frederick's young grandson Conradin made a supreme effort to save the fortunes of his house; after a triumphant entry into Rome he was beaten by Charles at Tagliacozzo (1268) and executed at Naples.

During the struggles between papacy and empire Italy had become divided into two parties, papalist and imperialist, which in the course of the 13th century assumed the names of Guelf and Ghibelline. This had not only affected relations between states but had also divided the population of the towns, thus giving fresh impetus to local factions. The names remained a tragic legacy of the Hohenstaufen period to the political life of Italy.

Monarchies and communes in the 13th century. By the end of the 12th century, the communes had triumphed in Lombardy and Tuscany; but although the communal movement extended also to other parts of the country, it did not prevail everywhere. Large regions of Italy retained or developed monarchical institutions: foremost, the Sicilian kingdom, but also the papal states Piedmont and Sardinia. Feudalism formed an important element of monarchical Italy, which represented also in this respect a contrast to the antifeudal policy of the communes. With other European countries, monarchical Italy had in common the development of assemblies of estates (*parliamenti*), in which the towns were represented. Originating primarily from the earlier feudal assemblies, the parliaments of the 13th and 14th centuries reflected the attempt to give the towns, together with the feudal classes, an influential place in the political structure of the monarchical states. They thus constituted also a new development in the position of the towns, which often enjoyed considerable local autonomy under monarchical control. Conditions varied considerably, however, both with regard to the functions of the parliaments and the rights of the towns. Thus the Papal States, comprising many communes that had been only recently acquired, contrasted with the Sicilian kingdom, in which the towns had remained strictly subjected to the monarchy from the time of the Norman conquest, whatever the privileges that had been granted to them.

In communal Italy the internal conflicts that began to disrupt the communes in the 12th century created new and far-reaching problems. Caused by rivalries and feuds within the ruling oligarchies and already, sometimes, by social conflicts, they led, toward the end of the 12th century, to the institution of a single executive magistrate (*podesta*). This innovation, however, did not put an end to internal strife. While the consular government disappeared at the beginning of the 13th century, it became the general rule for the *podesta* to be a citizen of another town so that the executive could no longer be the object of family rivalry. The struggles between the municipal parties, led by powerful families from which they often took their names, became a permanent feature of communal politics; and strife was intensified by the custom of the blood feud (*vendetta*) and by the expulsions of the defeated party. The growth of Guelfism and Ghibellinism gave the local parties endless possibilities of outside support and added fresh violence to their struggles. But they were no longer alone in their desire for political control.

The *podesta*

The increase of economic prosperity and the growth of the town populations led to a challenge of the virtual rule of the aristocracy in the communes. The prosperity was primarily the result of the expansion of Italian trade and the development of Italian banking and industry. The Crusades provided Venice, Pisa, and Genoa and, indirectly, other towns with new trading centres on the fringe of Asia; the Fourth Crusade, which established the Latin Empire at Constantinople in 1204, gave Venice a colonial dominion in the Levant, the economic value of which was immeasurably great. The restoration of the Byzantine Empire under Michael VIII Palaeologus in 1261 was not a serious blow to Venice's new dominion, but it gave Genoa vast commercial opportunities at Venice's expense. The Sicilian policy of the papacy after 1250—the offers of the kingdom first to Edmund, son of Henry III of England, and then to Charles of Anjou—provided Italian bankers and merchants with new fields of action. Hand in hand with increasing prosperity went immigration into the towns, the new citizens being recruited from all classes.

The *popolo*

The rise of the merchant and craft guilds was closely related to these developments, another form of organization of the *popolo* being military companies. The *popolo* roughly corresponded to the middle classes, and the final stage of its political formation was reached when it was organized like a commune, with an executive, councils, and statutes of its own. Conflicts between the aristocracy and the *popolo* began early but generally did not reach full strength until the 13th century. A first attempt to settle them led in Milan and other Lombard towns to the government's being shared between the two classes. Later in the 13th century, the organized *popolo* sometimes succeeded in establishing control of the commune, as in Florence in 1250 and 1282.

Instability and civic strife were generally traits of the 13th-century communes. The fate of the vanquished was bitter, the political exile becoming a typical figure; and the victor's desire for permanent power was determined not only by ambition but also by the fear of the consequences of defeat. To grant special powers to the leader of one's party or of the *popolo* might appear the easiest method to achieve this end; but the breaking down of constitutional limitations opened the road for the establishment of despotic government.

Interstate conflicts

Interstate conflicts, which had existed even before the rise of the communes, were a constant feature of Italian politics from the 12th century onward. Caused by quarrels over boundaries, by commercial and political rivalry, or by Guelf and Ghibelline loyalties, they present a bewildering picture, which nevertheless reveals some patterns. Thus the conflicts between Pisa and Genoa, the principal rivals on the Tyrrhenian coast, both aiming at control of Sardinia, continued intermittently throughout the 12th and 13th centuries and culminated in Pisa's decisive naval defeat off Meloria in 1284, which sealed its decline as a great maritime power. In the meantime, the struggle between Genoa and Venice over the eastern trade, beginning in the 12th century, had been intensified by the establishment of Venice's Levantine dominion in 1204 and the restoration of the Byzantine Empire in 1261. Commercial rivalry must have been largely responsible for

Milan's traditional enmity with Pavia and Cremona. In Tuscany, the struggles between Pisa and Lucca began in the 11th century. Lucca, once the most powerful Tuscan town, had been overtaken by Pisa, the chief Tuscan port. Florentine expansion led to mortal enmity between Florence and Siena from the 12th century onward, whereas Florence's growing economic and political power brought about struggles with Pisa from the beginning of the 13th. One effect of these conflicts was that the emperors always succeeded in enlisting the services of communes against other communes. Another result of the conflicts and of the territorial policy of the communes in general could be the acquisition of city-states by more powerful neighbours. But in this respect, too, the rise of the *signori* (seigniories; see below) proved a new departure. (Ed.)

Italy in the late Middle Ages and the Renaissance

SOCIAL AND POLITICAL DEVELOPMENT

The 14th and 15th centuries coincided, in the history of Italy, with the age of the Renaissance; hence, Italian social and political development in this period has been an object of special interest. Not only did it supply the context for brilliant achievements in artistic and literary culture; in addition, these centuries saw the emergence in Italy of patterns of social and political organization that have been conventionally taken as marking the end of the Middle Ages and the beginning of the modern era for the rest of the European continent. According to this view, Italy, in the phrase of Jacob Burckhardt, from his classic *Die Kultur der Renaissance in Italien* ("The Civilization of the Renaissance in Italy"), published in 1859, was "the education of Europe."

Withdrawal of imperial and papal authority. Fundamental to the history of Italy during this period was the virtual withdrawal from the peninsula of both universal powers, the empire and the papacy, whose long struggle, although for centuries disrupting Italian political life, had at least given some unity to its history. The emperors of the 14th and 15th centuries, chiefly concerned with promoting their dynastic interests in Germany, were generally indifferent to Italy. Although they occasionally descended into the peninsula for the prestige of a coronation in Rome and to derive income from selling titles and privileges to local powers, none made a significant effort to impose imperial rule in Italy. And, with the exception of a rare and anachronistic idealist such as the poet Dante, who hailed the visit of Henry VII to Italy in 1310 as the onset of a golden age of peace under orderly imperial rule, Italians were equally indifferent to this traditional source of political authority. Towns and nobles were concerned only to exploit a connection with the emperor in order to consolidate their merely local power or to extort from him the recognition of their own claims as the price of a support that was little more than nominal.

At about the same time, the efforts of the papacy to transcend local politics by a strong reassertion of its universal authority over the powers of Europe met with a major defeat that effectively removed the pope as a strong presence in Italian affairs for most of the 14th century. When Philip IV of France insisted on his right to tax the French clergy to finance war with England, Pope Boniface VIII delivered a stern rebuke in the bull *Clericis Laicos,* which was followed by an emphatic assertion of papal authority in a second bull, *Unam Sanctam* (1302). The French king accepted the challenge. He accused the Pope of heresy for claiming to stand above all secular rulers and dispatched a small army under Guillaume de Nogaret, from the French base in Angevin southern Italy, to the papal residence at Anagni, with the aim of seizing the Pope and carrying him off to France to be tried and deposed. Nogaret forced himself briefly into Boniface's presence and announced his errand; and, although the Frenchman was expelled from Anagni before he could carry out his mission, the aged pope died a few weeks later. The consequences were disastrous for the papacy as an Italian institution. French pressures on the demoralized papal court secured the election of a new French pope, Clement V (1305–14), who soon moved the papacy from Rome to Avignon, in southern France; and there it remained until 1377. The popes in Avignon by no means forgot Italy, but their absence from the peninsula substantially reduced the role they could play in its affairs.

Papacy at Avignon

General characteristics of Italian society. With the two great powers thus absent from the scene, the peoples of Italy were left to determine their own destinies during the 14th and 15th centuries largely without external interference; some historians have seen in this period a great but missed opportunity to construct for Italy a national state comparable to the monarchies developing north of the Alps. This view seems, however, to be based on reading back into the remote past a conception of Italy as a latent political unity that largely reflects the presuppositions of the unification movements of modern Italian history. In fact, since its forcible unification under Roman rule in antiquity, Italy had disintegrated into innumerable local units, and the withdrawal of both the imperial and the papal powers now deprived it of the little coherence its political history had revealed in earlier centuries of the Middle Ages. Italy differed from states such as France or England in its social composition, in its geographical divisions, and above all in its lack of a monarchy around which national institutions and sentiment could be organized. Without such a cohering power, the Italian towns and principalities were neither unified nor in any way checked in their claims to independence; nor were they any longer forced into mutual alliance—as during the time of the Lombard League (formed to combat the Hohenstaufen emperors of the 12th and 13th centuries)—by a mutual need to resist its encroachments. Thus, the history of Italy in the age of the Renaissance consists first of all of the separate histories of a large number of particular political entities, often widely different from each other, that happened to be located on the Italian peninsula and were drawn into relationships with one another less by a sense of community than by their mutual antagonisms.

Some conditions and experiences were, nevertheless, common to most of the inhabitants of Italy. One was precisely this peculiarly local orientation of Italian life, which, to a larger degree than elsewhere in Europe, was centred in towns. These had a certain family resemblance to one another, above all in the intense loyalty they commanded; Italian townsmen saw themselves primarily as inhabitants of particular communities, not as divided members of an Italian nation. Their ardent local patriotism found expression in devotion to local saints, in similar myths about the origins and uniqueness of their towns and in literary compositions praising them, and, negatively, in traditions of enmity with other towns. Their devotion to local independence was manifested in constant appeals for liberty, which to them meant above all the right of self-determination, of freedom from control by any external authority. In the 14th century this concept found juridical definition in the work of commentators on Roman law, common to most of Italy, and notably in that of Bartolus of Saxoferrato (1314–57). Although they still admitted the theoretical supremacy of the emperor, the lawyers adapted theory to practical reality by developing the principle that any community has an original and inborn right to govern itself. In addition, Italy, during these centuries, increasingly shared a common language. Although local dialects continued to flourish, Tuscan, the language of Dante—partly because of the distinction with which he had used it—increasingly became the customary speech of educated men.

Spread of Tuscan language

The Italian towns not only found means to resist the intrusion of imperial authority into their affairs; they also generally worked to reduce that of ecclesiastical authority. Without attacking the theoretical supremacy of the pope, townsmen tended to regard the church, like their secular governments, as for all practical purposes a local affair. Communities erected their own church buildings and quite naturally adopted a proprietary attitude toward them, and in the 14th century they were taking various steps to bring the clergy under local control. In Venice, for example, the Senate appointed bishops, and the government restricted papal taxation of the local clergy and tried clerical lawbreakers in secular courts. In Florence the government also controlled the operations of the Inquisition and abolished many clerical privileges. Such actions were accompanied by a defiance of the traditional sanctions by which the papacy had been accustomed to discipline the faithful to obedience. Florence between 1376 and 1378 and Venice on several occasions chose to ignore a papal interdict. The particularization of Italian life was increasingly ecclesiastical as well as political.

These tendencies reflected not only pressures for local autonomy but also the increasing laicization of Italian society. The precocity of Italian economic development meant that in Italy laymen were rich, educated, ambitious, and assertive to a larger degree than elsewhere in Europe. They were particularly inclined to resent the claims of the clergy to special privileges such as exemption from lay courts or from taxation and to resist any effort by the ecclesiastical authorities to control social and political life—for example, by applying the church's usury laws, which restricted the lending of money for interest. They saw priests not as superior to other men but as primarily the servants of the communities whose spiritual needs they were supposed to meet; in some areas of Italy it was common for parish priests to be elected by the more substantial laymen of the parish. This should not be taken, however, as a sign of any decline in religious fervour. The 14th and 15th centuries were, in fact, a peculiarly devout age in the history of Italy, but Italian devotion now took on a special quality. It found expression in spontaneous and local confraternities of laymen for the purposes of performing pious works and devotional exercises together. Numerous Italian saints, often with lay backgrounds, arose during this period. Lay heretical movements, often strongly anticlerical, were also still active, especially in the 14th century; an example is the Fraticelli, a radical spiritual branch of the Franciscan Order, who were treated with some indulgence by the Florentine and some other secular authorities.

Increased size of cities

The special piety of the age was related to developments in 14th-century economic and social life, which, in spite of wide local variation, also affected most Italians in this period. At the beginning of the 14th century, Italy was reaching the climax of a long period of prosperity, based on commerce, that had accelerated since the start of the Crusades. This commerce had nourished the growth of urban population. At the beginning of the 14th century, three of the cities of Italy and Sicily—Palermo, Venice, and Florence—had populations in excess of 100,000. These three were the largest cities in Europe. Milan and Genoa, with well over 50,000, were not all that far behind; and Bologna, Padua, Siena, and Perugia, with populations of between 20,000 and 50,000, were also sizable cities. Numerous lesser communities were also regarded, by the standards of the time, as major towns.

Crises of the 14th century. The long period of growing prosperity was brought to an end in the middle decades of the 14th century by a great catastrophe that reflected basic weaknesses in the medieval economy. In the decade after 1340, Italy—along with other parts of western and central Europe—was afflicted by successive waves of pestilence, above all by the Black Death (bubonic plague) of 1347 and 1348. Every part of Italy was affected, rural areas as well as towns, the rich along with the poor. In the major cities of Italy for which reasonable estimates are possible, the death rate seems to have been as high as 50 or 60 percent within a period of only a few months. This disaster, furthermore, proved to be only the beginning of a prolonged demographic crisis. Subsequent epidemics followed, in a regular cycle, about every 10 or 15 years, so that the rapid population recovery that usually follows periods of high mortality was regularly wiped out. The danger of death from disease—not only plague but also dysentery, cholera, typhus, typhoid, or smallpox—now became the normal condition of life, as it had scarcely been earlier; and this was to remain the case until well into the 17th century.

Most historians have explained this development as a result of the pressure of expanding population on a limited food supply. By the early 14th century most arable land had been brought under cultivation; intensive exploitation of the soil had reduced the productivity of older regions; and it is possible that a slight change in the climate, bringing cooler weather and unwanted heavy rains in the growing season, may also have been a contributory factor. Thus, agricultural surpluses could no longer tide people over bad years, and the result was periodic undernourishment, which increased susceptibility to disease.

Economic results of plague

The results of so fundamental a modification in the condition of human life brought a considerable change in the atmosphere of Italy in the 14th century. One result was psychological: the optimism of the preceding period came to an end, giving way to a climate of fear and anxiety that was also increased by political disasters, such as the steady expansion of Turkish power at the expense of Christendom in the eastern Mediterranean, and by the disorders of civil life resulting from endemic local warfare. This sense of insecurity was reflected in an intensified religious attitude. Not only were laymen more pious, but the numbers of Italians in holy orders increased, and piety and morality grew more rigid. The results of the pestilence and economic regression were also serious for the highly developed economies of the major Italian cities. Since the demand for goods fell with the depletion of the population, and labour costs increased for the same reason, both prices and profits declined; and the level of business activity fell sharply, especially in that international commerce in which Italy had established its leadership. The famous business enterprises of the Medici in 15th-century Florence were substantially smaller than those of the great Florentine banking and commercial firms of the earlier 14th century, the Bardi and Peruzzi. This reflects the fact that the recovery both of population and of economic activity was slow. Although a distinct upturn was evident by the earlier 15th century and was more pronounced in some parts of Italy and in some segments of the economy than in others, most of this period was characterized by relative, though probably not absolute, economic depression. The brilliance of Italian Renaissance culture was based partly on the restricted opportunities for business expansion, so that wealthy men had both the leisure and capital to devote to other interests.

Economic difficulties doubtless intensified the internal struggles that, continuing from the previous period, increasingly disrupted the life of most of the towns of Italy in the 14th century. Individuals, families, economic groups, and social classes engaged each other in a long struggle that dramatized the need for more effective government that would somehow be able to subordinate competing special interests to the general welfare. By the 14th century the disorderly tendencies of the old feudal nobility had been largely contained, and this group had been generally assimilated into the life of the towns. Another group was now dominant: that of the great merchants, bankers, and industrialists who, organized in their guilds, directed the most profitable economic enterprises of their communities. But the significance of their triumph for the quality of urban life in Italy should not be exaggerated. In sharp contrast to the rest of Europe, there was in Italy no radical distinction between the life-styles and the culture of nobles and merchants. Old aristocratic families often built palaces, settled in the towns, and even engaged in business, instead of remaining proudly aloof from urban life; and merchants tended to absorb their values, buying estates of their own in the countryside and investing in agriculture and reading chivalric romances as well as the classics. Jousting was a favourite diversion of Italian townsmen. Intermarriage between the two groups was also common.

Urban violence

But, in spite of this social and cultural amalgamation and perhaps in part because of the pretensions and militant traditions of the nobility, life in the towns of Italy was increasingly violent. Members of the ruling groups engaged in constant struggles for power with each other, and they also came into regular conflict with groups lower on the social scale, which they everywhere tended to exploit. These lesser elements in Italian society included the guilds of skilled artisans and small tradesmen, which resisted efforts to reduce their political rights, and an unorganized mass of city dwellers, usually of peasant origin, who formed a growing urban proletariat that was denied participation in the political life of the town. Between these groups and also among factions within them, there was constant tension, which exploded in periodic and often bloody civil strife. The result was a high degree of instability, rapid changes in political fortune, brutal seizures of power, conspiracies and aggressions, insecurity, and disorder.

THE ITALIAN STATES IN THE 14TH CENTURY

These conditions prepared the way for a characteristic development in the Italian towns of the 14th century, the rise to power of governments dominated by individual despots (signori). This process, already well under way in the previous century, now became general, especially in the north of Italy, where the disorderly republican rule of the communes gave way, in town after town, to government by one man. In some communities the dominant group imported an outsider, known as a *podesta,* to maintain order; the lordship of the Este family in Ferrara was established in this way. More usual, however, was the appointment of a captain of the people, an officer originally intended to check the growing power of the urban patriciate on behalf of the lesser guilds. Given resources to accomplish this purpose, he was likely to extend his authority by degrees, until it amounted to virtual control over the town and could then be made hereditary. This was the road to power for the Della Scala family in Verona, for the Carrara in Padua, for the Gonzaga in Mantua, and for the Visconti in Milan. Eventually, such lords might detach themselves altogether from the popular origins of their powers, perhaps buying from the emperor the title of imperial vicar or duke in order to emphasize their independence from popular control.

Once established, the signori consolidated their power by the centralization of the agencies of government in a purely personal regime. The lord, though he might continue to respect some of the forms of communal government, in practice exercised unrestricted authority over his subjects. He stood above the law, and his power was limited only by the danger of overstepping what was tolerable to his subjects, who, together, might be more powerful than himself. But he generally had the advantage of a monopoly of military force, for the old communal armies had largely disappeared. Even in republics, ruling groups feared to arm the discontented populace, and busy merchants had little interest in bearing arms themselves. The result was a general tendency to rely on mercenary armies led by military entrepreneurs (*condottieri*), who sold themselves to the highest bidder. Thus, armed force escaped popular control, and the unreliability of such armies, together with the disorders they often provoked, contributed to the increasing decadence of Italian military power in the 14th and 15th centuries.

Mercenary armies

It was once supposed that the rise of the signori was almost universal and was therefore the essential element in Italian political life during the age of the Renaissance. But it has now been clearly demonstrated that a major group of towns escaped conversion into despotism—notably Venice, in the north, and the more important communities of Tuscany, including Florence, Siena, Lucca, and Pisa. In these places merchant groups were usually too firmly in control to need the help of a strong man to preserve order. They could accomplish this, for the most part, by themselves, and so the republican forms of government, which they could dominate, were preserved. In these communities government still rested in theory with the whole body of citizens, though practice was far from democratic. Citizenship was variously defined, but participation in politics was regularly the monopoly of older and more substantial families. Republican governments were, nevertheless, based on what has been called the ascending theme in politics. According to this view, power is not imposed from above but resides initially in the community itself, the ends of government are defined by the community in accordance with its sense of its own special needs, and ruling authority is only delegated to public officials, who remain responsible to those with whose affairs they are entrusted. Such conceptions, the direct antithesis of the descending theme that dominated most medieval political thought, were elaborated in the *Defensor pacis* ("Defender of Peace") of Marsilio of Padua (1324), a work that makes it clear that, even on a theoretical level, the significance of the Italian achievement in politics was not limited to the construction of despotism.

In one respect, however, despotisms and republics were alike: both types of government tended to expand by absorbing their smaller neighbours and gradually constructing larger regional states. Thus, in the 14th and earlier 15th centuries, the chaotic pattern of innumerable petty political units in northern and central Italy gave way, largely through conquest but on occasion by purchase, to a few much larger units. Verona, for example, had by 1335 absorbed Vicenza, Treviso, Padua, and Reggio; and by the end of the 14th century Florence had taken over much of Tuscany. But such regional empires were likely to be unstable and at best were often a mixed benefit. Although they could be exploited economically, they also posed problems of administration and control that further taxed the political and military resources of the major states. The citizens of the absorbed towns resented external rule and were likely to revolt at every opportunity. Moreover, the extended territories to be defended and eventually the rival ambitions of the larger powers to absorb lesser powers brought them into dangerous confrontations with each other.

Milanese despotism. The most aggressive Italian state of the 14th century was Milan, whose history may be taken as additional illustration of many of these generalizations. The city had long suffered from the same complicated struggles and internal disorders that plagued other communities, and its troubled populace finally turned for protection to the Visconti, a noble family with large lands outside the city. Beginning as captains general of the people, the heads of the family also exploited a paper allegiance to the emperor to establish an increasing independence from popular control. But, as with other successful despotisms, the chief basis of their power was the support of substantial groups in Milan who, at least initially, appreciated their ability to preserve order, together with their increasing control of military force. As a result, they were able to dominate one area of government after another: legislation, taxation and expenditures, the judicial system, and foreign policy.

The Visconti family

The rise of the family began with a division of the Milanese patriciate in the 13th century into factions led by the rival families of the Della Torre and the Visconti; these bore, respectively, the old Guelf and Ghibelline labels denoting the pro-papal or pro-imperial parties, though, as elsewhere, these terms were becoming increasingly unreal. The Ghibelline faction was led, after 1277, by Ottone Visconti, archbishop of Milan, whose family claimed aristocratic origins going back to the early 9th century, owned large estates just outside the city, and had long exploited local ecclesiastical office on behalf of its younger sons. Ottone succeeded in defeating his rivals, became virtual ruler of the city, and secured the election of his nephew Matteo as captain general in 1287. Matteo was responsible for a close alliance with the Emperor, an alliance that played a large part in the history of the family. He was appointed imperial vicar of Lombardy, and, although he was briefly expelled from the city by the partisans of the Della Torre, he was able to return after the imperial expedition of Henry VII in 1310, which he had supported. The Della Torre were now permanently crushed, and Matteo was made captain general for life. In 1317 he was able to make his position hereditary, and his successors extended their power step by step, notably under Azzo Visconti (1328–39).

This process was brought to a climax by Gian Galeazzo (1351–1402), who bought the title of duke from the Emperor in 1395 and married a daughter into the princely French House of Orléans, an alliance with unhappy consequences for the future of Italy. Developing his government into something like a modern bureaucracy, he emphasized his power and remoteness from the people with an elaborate court etiquette and replaced the honourable title of citizen so long borne by the Milanese with the more ambiguous name of subject. And, with so much power at home, he came close to conquering and uniting into a single state much of the north of Italy; by 1402 he was threatening to subjugate even Florence, when he was suddenly carried off by the plague.

The effectiveness of Milanese despotism has sometimes been taken to illustrate one of the most significant aspects of the Italian political achievement in this period. In this view, the tyrants of the Italian Renaissance pointed the way to modern politics by being the first rulers to conceive of the state and its government as, in Burckhardt's famous phrase, "a work of art"; that is, a product of rational planning, deliberate calculation, and the careful adaptation of means to ends. Thus, disregarding their own theoretical subordination to the emperors, the Visconti dukes claimed an absolute authority over all their subjects, nobles and townsmen alike. They replaced elected officials with their own men, upon whose loyal service in enforcing obedience to themselves they could depend; they taxed and spent at will; they imposed uniform laws over their state; they took possession of all fortified points, dispossessing the local nobility; and they were even able to institute a censorship over mail and a kind of passport system to control travel. Such an accumulation of measures is indeed impressive and certainly suggestive for the techniques of later European despotism; but it is doubtful that they really reflected a kind of blueprint for the state. As was done elsewhere, the dukes in Milan actually improvised their policies piecemeal to meet particular problems or take advantage of special opportunities as they arose.

Such government, nevertheless, had obvious attractions in a period of general disorder, advantages that were widely advertised throughout Italy by Visconti propagandists such as the Humanist Pier Candido Decembrio (1392–1477). But the people of Milan paid dearly for Visconti order with the loss of their freedom, a loss the significance of which was often highlighted by the ducal government's abuse of its power. Gian Galeazzo's predecessor Bernabò, for example, extorted huge sums from his helpless subjects, and some of his successors in the 15th century were notorious for an arbitrary ruthlessness against which there was no recourse. The rule of despots such as the Visconti was merely personal, and it never succeeded in creating the modern type of institutional state, independent of the individual ruler and able to rely for its stability on the loyalty of the subject. Selfish special interests were kept in check, to be sure, but they were subordinated to the special interest of the ruler, not to the general welfare. And the memory of communal self-government and the resentment of the suppressed special interests persisted in the Milanese state. When the last of the Visconti died in 1447, the Milanese people made a pathetic effort to restore republican government. But the Ambrosian Republic, as it called itself in memory of the great 4th-century Milanese saint Ambrose, could not solve the problems republics had failed to deal with earlier; above all, it could not control the restive Milanese territory beyond the confines of the city. Thus, by 1450 the republic had been overthrown by Francesco Sforza, one of the great *condottieri* of the age, who had served Milan in the past, and a new despotism replaced the old. Milan therefore continued as the outstanding representative of Italian despotism among the Italian states, in both its strengths and its limitations.

Florence in the 14th century. The history of Florence in this period is better known than that of any other place in Italy. This is partly because of the richness of the Florentine archives but chiefly because the importance of the city for Renaissance culture has attracted special attention to its political and social development. That attention has focussed especially on the survival of republicanism in Florence, in which it contrasted strikingly with the despotism of so many other communities, notably that of Visconti Milan. Because they continued to participate in politics and to take some responsibility for the general welfare, Florentines tended to develop explicit loyalties and a habit of participation in public affairs, and their practice can be seen as an important precedent for posterity. Yet the survival of the republic was often precarious in the 14th century, and, in much of the 15th, republican forms of government were little more than a facade for the personal rule of the Medici. Thus, the history of Florence reveals both similarities to and differences from that of Milan.

Florentine republicanism

The independence of Florence was protected by its strength as the largest city of Tuscany and by its surrounding circle of mountains; and its prosperity was nourished by a relatively diversified economy. This economic activity had developed later than the enterprises of the maritime cities and even of Milan. Still a small town at the end of the 12th century, Florence at the end of the 13th had only recently become prosperous. Because of its inland location, it did not specialize in international trade, though, situated on the major north–south route of the peninsula, it actively participated in it. In addition, Florence had become a major centre for cloth manufacturing, especially of woolens, taking advantage of the decline of the cloth industry in Flanders. Since Florence was a major power in the Guelf alliance, its merchants began, in the 13th century, to lend money to popes, for whom they also served as tax collectors throughout Europe, and also to nobles and other rulers. During the 14th and 15th centuries the great business firms of Florence engaged simultaneously in these and in other activities, such as mining. They established a network of agencies abroad, extending from the eastern Mediterranean to England and the Low Countries. The huge Bardi enterprise of the earlier 14th century, for example, had branches in all the larger towns of Italy and in Antwerp, Bruges, Paris, London, Avignon, Rhodes, Cyprus, and Constantinople. The great Florentine merchants also acquired estates in the surrounding countryside, the *contado,* which they actively supervised and from which they drew their food. This connection with the land was typical of the relation between town and *contado* in Italy.

The disorderly Florentine nobility had been largely excluded from political life by a new constitution in 1282, which vested the government in an elected council whose members, called priors, served for very short terms. This meant frequent elections and changes in the membership of the government, and it encouraged a high degree of public interest and participation in politics; it also made for uncertainty. But the restriction of participation in the political life of Florence to members of the organized guilds gave effective control over the government of the city to the more substantial business interests; the great majority of the priors in the earlier 14th century came from just three guilds, which represented the wealthiest men of Florence: the cloth finishers, the wool merchants, and the bankers, all of whom tended, like others in their position, to favour policies advantageous to themselves. Thus, they taxed property in the surrounding countryside but not in the city itself (where their own possessions chiefly lay), while, within the city, taxes were levied largely on necessities, especially food, consumed by the lower classes. The ruling group retained its old Guelf orientation, a residue of the 13th-century alliance with the papacy against the empire, not as a token of political subordination to the popes but, on the contrary, because this tradition represented the freedom of the city from any external control. This orientation also had continuing practical value. Dedicated to maintaining the Angevin rule over Naples that had been arranged by the Pope between the years 1265 and 1268 in order to exclude the Hohenstaufen emperors from Italy, the Florentine ruling class maintained close relations with southern Italy, facilitating the economic exploitation of the Kingdom of Naples by Florentine businessmen, who collected its taxes, monopolized its grain trade, and reaped huge profits. To idealists such as Dante, the result of such prosperity was a gross materialism in which the traditional values of a simpler age were in decay.

The three ruling guilds

Despite its injustices, the system worked well enough during good times, and the peace of the city was only occasionally disrupted by factional disputes in which those who lost (such as Dante himself) were exiled from Florence. But the political organization tended to break down in a crisis, and after 1340 Florence was in serious trouble. As a leading centre of international commerce and finance, it was badly hurt by the general economic decline that began in the following decade, which in its case was aggravated by special circumstances. The leading business houses of the city had made huge loans to England to finance Edward III's war against France, and much of the capital of Florence was tied up in this dubious enterprise. Thus, the subsequent repudiation of his debts by Edward III brought general disaster to the city, and meanwhile the ruling group had been further discredited by the failure of its attempt to conquer the neighbouring city of Lucca. In these circumstances, in an effort to preserve its control over Florence, it imported a French adventurer, Walter of Brienne, who called himself duke of Athens, to preserve order. But, in a manner familiar elsewhere, this man chose to rule for his own ends, and soon all groups in the city united to expel him and save the republic. Indeed, the failure of this experiment was followed by several decades of more popular government as new men came into Florence from the countryside to replenish the losses of population during the Black Death and, in a time of relative social mobility, rose in Florentine society.

The social dislocations caused by the plague thus combined with the strains of the expansion of Florentine dominion over much of Tuscany during the 14th century to produce important changes in Florence. The newcomers from the countryside were needed to replace those who had died; and, as business again slowly began to improve, some of them did well enough to rise in the world and to challenge the older ruling group. The expenses of prolonged warfare gave them an opportunity: the government badly needed money, and any man with the resources to lend it was able to claim the political influence that accompanied being a creditor of the state, regardless of his family background. In the interest of protecting its investments, this latter group of moneylenders was also concerned to make government more efficient. As a result, the procedures of government became more bureaucratic, professional, and impersonal.

Another decade of turbulence after 1375 imposed a further test on Florence, and again it emerged with its republican institutions stronger than before. This period of crisis began with a war against the papacy, a result both of the growing territorial expansionism of Florence and of the disorders in the Papal States during the Pope's absence in Avignon. These tended to spread into adjacent areas of Tuscany, until the Florentines felt compelled to intervene, thus antagonizing the papacy. The ensuing struggle was known as the War of the Eight Saints (1375–78), so called after the committee that supervised it on the Florentine side, and it raised the most serious ideological questions. The Pope, by imposing an interdict on the city and excommunicating its leaders, converted the conflict from a localized and purely political war into an alleged rebellion against ecclesiastical authority. The implied suggestion that Florence had no basis for its existence as an independent and secular state and no right to conduct a policy based on a sense of its own interests brought into question the fundamental issue of Florentine liberty. Florence found a spokesman in its chancellor, the learned Humanist Coluccio Salutati, whose broad propaganda campaign represented the war as a struggle "for the salvation of Italy and the liberty of all." The whole episode was accompanied by vast republican enthusiasm. Although the war itself was inconclusive, Florence came out of it with a deepened sense of the value of its liberty and also of the essentially secular quality of politics.

Revolt of the *ciompi*

This crisis gave way to a new one in 1378 with the revolt of the *ciompi*, workers in the Florentine wool industry, an event that has usually been interpreted as an uprising of the working class against the business group that had long controlled the government as well as the economy of Florence (and it may be that the recent war, waged in the name of liberty, had aroused some radical democratic ferment). More recently, however, this view has proved incorrect; the true leaders of the Ciompi Revolt, it is now clear, were not workers but disaffected members of the old ruling group itself. They managed to form a new government, with a somewhat more democratic constitution, that retained power until 1382. But the new group was unable to maintain order, and the previous rulers then returned to power in a strengthened position; and, in the period that lay ahead, political power gradually contracted once again. Florence remained in this situation until the end of the 14th century, and the continuation of the same group in control of the government through the first third of the 15th century indicates that the republic had attained a new level of stability. Doubtless, a gradual, though incomplete, economic recovery was helpful. In addition, the brief Florentine experiments with dictatorship and revolution also probably contributed to the stability of the republic. They interrupted the tendency of the merchant oligarchy to abuse its position, displaced groups that had long enjoyed authority, and opened up opportunities for new men to rise into the ruling class. The narrowly oligarchical character of the government remained, but in Florence the oligarchs seemed to have learned something from events and to have developed both a broader understanding of their own interests than elsewhere and greater sensitivity to public needs. Thus, faced with a financial crisis in 1427, the government adopted a new and more equitable form of taxation, based, after a careful survey of individual property, on wealth. In this, they displayed a willingness to assume a major responsibility for the support of the state, instead of shifting it to other and poorer men. It is understandable, therefore, that Florentines remained reasonably united in their support of the state; and their city survived as a republic while many other Italian cities were turning to despotism.

Venice in the 14th century. One other great Italian city remained a republic and for reasons partly similar to those that influenced the history of Florence. This was Venice, in which the domination of a merchant oligarchy was even more complete than in Florence. But in other respects the history of Venice was different. Its rise to economic power was not of recent origin but extended back over many centuries; and its island location, detached from the mainland, freed it from the problem, so troublesome to many other states, of imposing discipline on a disorderly landed nobility. The society of Venice was, therefore, unusually homogeneous, and its ruling patriciate had a tradition of solidarity very different from that of individualistic and turbulent Florence. This solidarity was reflected in the active role of the government in the organization and regulation of all aspects of the Venetian economy. The state built a large part of the Venetian merchant fleet in its Arsenal; it organized and directed the convoys in which Venetians transported their commodities on the seas; it imposed standards on Venetian manufactured goods and inspected them for quality in order to maintain the competitive position of the city. It also regulated prices and wages, for, in Venice, guild organization lacked the strength and independence it had in Florence. The relative internal peace of Venice, so widely admired elsewhere, depended largely on the combination of this solidarity with the great prosperity it produced and in which most Venetians shared, though the tranquillity of Venice, known as the Serenissima, "most serene city," and the high degree of personal freedom that went with it—including a tolerance for Jews, Greek Christians, and even Muslims that shocked contemporaries—were generally attributed to the wise arrangement of its institutions and the rigour and equity of Venetian justice.

Venetian tolerance

Political rights in Venice had been restricted in 1297 to those families at that time sitting in the Great Council. Occasional gestures of discontent with this arrangement were ruthlessly suppressed by the Council of Ten, an agency established early in the 14th century that proved singularly effective throughout the long history of Venice in protecting the established government of the city. Since the Great Council was too large to function as an effective governing body, the chief legislative and policy decisions of the republic were the work of the Senate, a body with great prestige, most of whose members were elected by the Great Council. Despotism from above was avoided by such close restrictions on the doge, the executive head of the state, that he was little more than a figurehead except in times of special crisis. The Venetians tended to choose for this post not vigorous leaders but old men for whom it was a reward for a long career of services to the state.

But the republican institutions of Venice were little threatened from either above or below. Venetian society was united in an immensely profitable commerce with the eastern Mediterranean, in which Venice had long been the chief middleman in meeting the needs of all Europe. Although the 14th century brought an end to the regular movement of Venetian galleys into the Atlantic and northward to England and the Low Countries, Venice remained one of the busiest ports of Europe, and its overland trade through the Alps into central Europe continued to expand. The 14th century also saw the decisive triumph of Venice over Genoa, its great rival in the commerce of the East. A series of wars finally culminated in a great naval victory at Chioggia in 1380. Genoa's fleet was destroyed, and it never fully recovered from the defeat. Relieved of this competition and with the gradual improvement of business, Venice entered, in the earlier 15th century, into perhaps the most prosperous period of its history. Its internal stability owed much to this prosperity.

The Venetian empire

Venice also differed from the other states of Italy in one crucial respect. It was not a purely Italian power but the possessor of a great empire consisting of a string of commercial bases extending down the east coast of the Adriatic to the islands of the Aegean and including the great island of Cyprus. A good deal of this empire was acquired in the 14th century in the course of the wars with Genoa. Much of the attention of the republic was directed to its administration, for which Venice developed a body of able patrician officials that has been compared with the British colonial service of more recent centuries. And, concerned with the maintenance of this empire on which its trade depended, Venice could not confine its attention, as did the other Italian states, to the Italian peninsula; indeed, throughout most of the 14th century it remained largely aloof from the affairs of the mainland. The main interest of Venice, in fact, was directed to the larger politics of the eastern Mediterranean. This preoccupation became even stronger in the 15th century, when its interests there began to be challenged by the expanding Ottoman Empire. Against this power Venice fought a long series of delaying actions, in which periods of active warfare were regularly interrupted by intervals of peace and friendly commercial intercourse, at which the official conscience of Christendom professed to be scandalized. Self-righteous indignation against Venice on this score, combined with resentment at its detachment from Italian affairs and envy of its wealth, made it from an early point unpopular with other states. At the same time, Venice acquired a singular reputation for political prudence.

The Papal States. Although Milan was a despotism, and Florence and Venice were republics, they were alike in that each was based on a city that, at least in practice, insisted on its absolute independence from any external control. In this respect, the other major political entities of Italy were somewhat different. The Papal States, extending southward from the River Po, across the Apennines from Tuscany, and then cutting across the centre of the peninsula, included a considerable variety of political units. In the north and centre, Emilia and Umbria had a number of major towns, among them Ferrara, Rimini, Bologna, and Perugia. These were similar in important respects to the towns of northern Italy and Tuscany. Ferrara had evolved in the familiar way from a free commune into a relatively stable and well-governed despotism under the Este dukes; because of its location, it tended to fall under the political and cultural influence of Venice. Rimini was taken over by the Malatesta family. Bologna, though remaining nominally a republic, was dominated by the Bentivoglio family. Perugia fell to the Baglioni family. But these communities, whatever their forms of government, also owed obedience to the pope, whose authority over them involved an ambiguous mixture of spiritual and secular claims. And the pope, though often in no position to do so, was constantly concerned to exact from them what he considered due both to himself and to St. Peter.

Much of the central and most of the southern part of the Papal States consisted largely of feudal domains, some of considerable size, whose proprietors were as likely to rebel against papal control as against that of any secular lord. And, in the absence of the papacy during its residence in Avignon, the Papal States, never very firmly ruled by the popes, disintegrated almost completely. The despots in the towns expelled papal officials and ruled in complete independence; since papal taxation had been heavy and papal authority an irritating infringement on local liberty, these actions were often popular. The nobility fought each other, terrorized the countryside, and did as they pleased, and bandits also made the region everywhere unsafe.

Rome in the 14th century

The restiveness of the Papal States extended also to the city of Rome, which in the early 14th century and in the absence of the papal court was little more than a small provincial town, now overshadowed by the ruins of its glorious past. Economically dependent primarily on the exploitation of pilgrims, Rome had for some time been dominated by a struggle for control between the two great families of the Orsini and the Colonna. Rome nevertheless provided, in the mid-14th century, the unlikely setting for a curious attempt, inspired by memories of its past greatness, to restore a republican government that could lead Christendom in a general movement of moral recovery. It was led by a young notary, Cola di Rienzo, who headed a revolution in 1342 in which the great nobles were expelled and a republic established. But, after obtaining papal approval of this action on a visit to Avignon, Rienzo's idealism grew increasingly extravagant. It became gradually apparent that he saw in the rebirth of the Roman Republic the start of a general rebirth of order and virtue in the Western world. He developed an increasingly messianic conception of himself and dispatched letters to the various cities of Italy and to European princes invoking their support for his program of world reform. But his pretensions eventually antagonized the distant pope, and Rienzo lacked political talents of a practical kind. His enemies combined against him: in 1354 he was overthrown and killed, and Rome returned to its old ways. The episode is, nevertheless, instructive for the mood of Italy in the difficult middle decades of the century. It reveals something of the tension, the despair over the condition of Italy and of the world, and the apocalyptic hope for a dramatic change in the spiritual and political climate that agitated the peoples of Italy during this unhappy period.

The Great Schism

The turbulence of both Rome and the Papal States provides a partial explanation for the long residence of the papacy in Avignon, in spite of the fervent appeals of such figures as the poet Petrarch and St. Catherine of Siena for the return of the pope to Italy. The pope's absence from his traditional home was a major element in the pessimistic mood of the age. Yet the papal court had never ceased to be concerned with the condition of the Papal States; and, in the decade following the death of Rienzo, Gil Álvarez Carrillo de Albornoz, a Spanish cardinal acting as the pope's legate, began the difficult task of reducing them to obedience by a shrewd mixture of force and diplomacy. By 1377 his successors had made enough progress to permit the pope's return. But this achievement was largely thrown away by the Great Schism of 1378 to 1417. The return of Pope Gregory XI to Rome in January 1377 had been greeted with deep joy in Italy, but the aged pope died early the next year. The cardinals were largely Frenchmen who would have preferred to remain in Avignon, but the Roman populace was determined that the next pope should be an Italian. Accordingly, when the Sacred College assembled for this crucial election, a mob gathered outside their meeting place and threatened violence unless its wishes were met. Under these conditions, the cardinals chose the bishop of Bari, who assumed the name of Urban VI. But, on the ground that this choice had been coerced, a group of dissident French cardinals withdrew to Fondi in the shadow of the French-dominated Kingdom of Naples, held a second election, and chose one of their own number as pope. Calling himself Clement VII, he soon moved back to Avignon, and there were now two popes. Under such circumstances the authority of the Pope in Rome was again seriously weakened. Rebellious elements in the Papal States were able to play off one pope against another and to disregard the claims of both, and the Papal States once again fell apart. Another consequence of this situation was the dependence of the popes in Rome on the support of other Italian powers, at times Naples but, more significantly, Florence. The enmity that had produced the War of the Eight Saints (1375–78) gave way to a close friendship that was all the more needed in Rome, since the rivalry with the French papacy in Avignon was accompanied by periods of tension with the Angevin rulers to the south. Humanist scholars from Florence were now regularly employed at the papal court.

Naples and Sicily. The history of Naples and Sicily in this period is largely a story of dynastic changes within the framework of a backward and feudalized society. The effective control imposed by the Hohenstaufen emperor Frederick II collapsed after his death, and the habit of rebellion against central authority long encouraged by the papacy made monarchy in both places constantly precarious. Government in both regions has been described as despotism tempered by revolt. The power of great noble families also kept towns in a condition of political inferiority, and for the same reason class distinctions remained strong. Southern Italy experienced little of that mingling of nobles and townsmen so characteristic of society in the north. Thus, the contrast between southern Italy and the urbanized north was even greater than in the case of the Papal States. Yet, because of their dynastic ties with great powers outside Italy, Naples and Sicily regularly played an important part in the affairs of the peninsula.

The union of Naples and Sicily under the French House of Anjou, arranged by the Pope in the 1260s, had ended in 1282 with the popular revolt known as the Sicilian Vespers. Peter III of Aragon was invited to become king of Sicily on the strength of a distant family connection with the Hohenstaufen line, and Sicily was henceforth under Aragonese rule. The Angevins remained, however, in control of the Kingdom of Naples, supported by the popes, whose feudal suzerainty they continued to acknowledge; this connection became even stronger during the residence of the papal court at Avignon. The direct Angevin line ended in 1382 with the death of Queen Joan I; the one positive accomplishment of her disturbed reign had been the recognition of Aragonese rule in Sicily by a treaty of 1372. Her will had named Louis, duc d'Anjou, brother of Charles V of France, as her heir, and this bequest was to be the basis of future French claims to the kingdom. But, with the support of the Pope in Rome in his capacity as feudal overlord of Naples, the will was set aside in favour of a junior branch of the family, represented by Charles of Durazzo, who became King Charles III of Naples. Angevin rule continued until 1435; then, once again, the direct dynastic line was extinguished with the death of Queen Joan II. She, too, tried to pass on the rule of the kingdom to the French House of Anjou, but she had also made an earlier bequest to Alfonso V the Magnanimous, the Aragonese ruler of Sicily, who won control of the kingdom by 1442. Thus, Naples and Sicily, separated for a century and a half, were reunited under a single head. They remained so only until Alfonso's death in 1458, but both regions thereafter remained under different branches of the House of Aragon.

Other Italian states. It has been convenient for historians to portray Italy in the age of the Renaissance largely in terms of Milan, Florence, Venice, the Papal States, and Naples and Sicily. But it should not be forgotten that there were other parts of Italy whose territories were not always and in some cases were never included in these larger entities. Thus, in the northwest Alpine region were the feudalized territories of the House of Savoy, divided into three branches, which ruled Savoy, Vaud, and Piedmont. In view of the future importance of this dynasty for Italian history, it should be noted that Savoy was not yet considered an Italian power, being still oriented to France and Switzerland; and it remained during the 14th and 15th centuries largely apart from Italian affairs. To the south of Piedmont lay Genoa, but, in spite of its commercial power of an earlier age, it, too, especially after its final defeat by Venice, played no great independent part in the political events of the peninsula. Without a surrounding province like that controlled by Florence, Genoa was frequently at the mercy of the more powerful adjacent states, and its external weakness was compounded by its singular internal instability. The struggle for control of its republican government produced frequent revolutions. An effort in 1339 to produce some order in Genoese affairs by instituting a doge, on the Venetian model, proved unavailing. In the summer of 1393, for example, this office changed hands five times. Such turbulence was an invitation to conquest; Genoa fell to the French in 1396 and later to Milan. And, though its citizens were constantly prone to rebellion against foreign rulers, they never developed the civic spirit of the Florentines or the cohesion of the Venetians.

The House of Savoy

Between Milanese-dominated Lombardy and the Venetian lagoons lay another urbanized area that, during the 14th century, preserved an existence separate from that of the great powers in the north. The most important political centres here were Verona, Padua, and Mantua, all republics that had gone the standard way toward despotism in the later 13th and earlier 14th centuries. Mastino Della Scala, a leading citizen of Verona, began the process there, first becoming captain of the people and then securing his independence from the commune by acquiring the title of imperial vicar. This dignity was then passed on to his brother and his nephews. In the 14th century his family was succeeded in the lordship of the city by the Scaligeri, who were notable as patrons of the arts. Paduan republicanism lasted longer: the city gave up its freedom only in 1318, when Jacopo di Carrara became its lord. Mantua had a similar history. Before the end of the 13th century, it had fallen under the control of the Bonacolsi, who remained in power till 1328, when, with some popular support, they were overthrown by Luigi Gonzaga. The firm rule of the latter established his family securely in a control that lasted until the early 18th century.

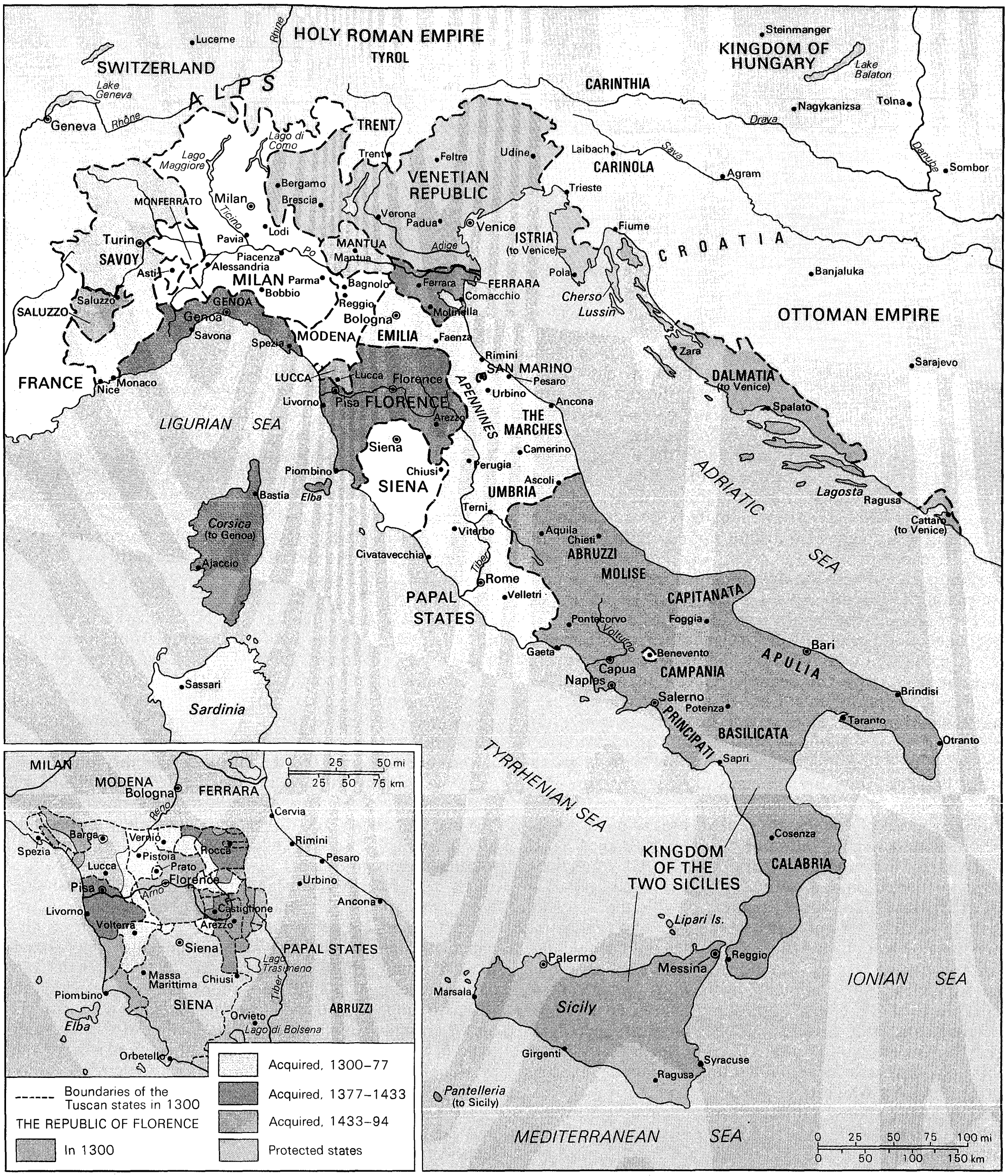
THE ITALIAN STATES IN THE 15TH CENTURY

Expansion of the major Italian powers. Even the degree of 14th-century Italian political consolidation that makes it possible for historians to present the later history of Italy in terms of the five major powers should not be exaggerated. Some lesser powers, such as the republics of Lucca and Siena, managed to preserve their independence intact, while the coherence of territories subjugated by the major powers remained limited and their stability uncertain. This was notably true in the case of Naples and the Papal States, but it was also often true of the territories gathered together under the rule of Florence; and the conquests of any individual Visconti duke of Milan were always liable to fall apart at his death.

This weakness notwithstanding, there was, in the late 14th and earlier 15th centuries, a significant shift in the interests of the major Italian powers: all, apart from the papacy, now paralyzed by the Schism, sought to expand their territorial authority. As a result, the concern of governments tended to shift from the internal struggles of an earlier period, first to conflicts with neighbouring powers and eventually to wars on a larger, sometimes peninsular scale. This was as true of the republics as it was of the despotic states. Florence, its commercial expansion inhibited by an inland location, sought a seaport on the western coast. This was the reason for its conquest of Pisa in 1406, and the addition of Livorno (Leghorn) by purchase from Genoa in 1421 gave it full control of the Tuscan coastline. By the next year the first Florentine galleys were heading directly to the Levant and soon thereafter to the European Atlantic ports.

Territorial expansion

Meanwhile, Venice had been abandoning its long isolation from the mainland and decided to conquer and organize for itself a substantial dependent state. Its motives were twofold. First, it needed to make secure its overland trade routes, which meant that it could no longer countenance the existence of a strong and hostile power between the head of the Adriatic and the passes through the Alps. Second, it required a nearby agricultural province under its own permanent control as a source of food, especially since Turkish conquests had made imports of grain from the Black Sea increasingly uncertain. The danger that Milan might move into the territories adjacent to the Venetian lagoons also impelled Venice to act. In the first decades of the 15th century it conquered the lands of the tyrants of Verona and Padua, who had been levying heavy duties on Venetian goods passing through their territories and occasionally actually cutting off Venetian food supplies. The wisdom of these conquests was long debated in Venice, whose power had so long been based rather on



Italy in the 15th century and (inset) Florentine expansion.

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York, revision Copyright © 1964 by Barnes & Noble, Inc.

the sea than the land; although it now had its own farms and had achieved secure access to the northern passes, the results were not altogether advantageous. Venice was henceforth far more involved in the political struggles of Italy, with new responsibilities and new demands on its resources. And its new conquests on the mainland further antagonized other powers already alarmed by its vast wealth. Before long, charges would be heard that Venice aimed to conquer the whole of Italy—indeed, to establish an empire over all Europe.

At the same time, the broadening of its contacts with the rest of Italy brought Venice into closer contact with the cultural movements of the Renaissance. Up to this time it had remained a self-centred, materialistic, and culturally backward community of merchants. But the conquest of Padua gave it control of the liveliest university centre in Italy, and its ablest young men proceeded to take full advantage of the new opportunities that this presented. Educated at the University of Padua, they brought back literary and scientific interests to Venice. By the later 15th century, aided by its overseas contacts with the Greek East and above all by its development as a major printing centre, Venice had become the capital of Greek learning in Europe as well as a point of diffusion for the Latin classics. Meanwhile, leading painters from the mainland had begun to visit the city, and Venetian painters began to absorb new ideas and to develop a Venetian school of Renaissance art.

The crisis of Florentine republicanism. For the time being, however, the greatest danger to the peace of Italy and the independence of other powers was Visconti Milan, especially under Gian Galeazzo, created duke of Milan by the German king Wenceslas in 1395. His conquests had played some part in influencing Venice to expand onto the mainland, but he posed a particular threat to Florence. Indeed, because of its central location on the Italian peninsula, Florence found itself regularly confronting aggressive princes seeking expansion from either the north

Milanese conquests

or the south. It had been particularly alarmed by the conquests of Gian Galeazzo, who expanded his empire steadily southward after 1385, taking special advantage of the disarray of the Papal States during these years of the Schism. By 1400 he had gathered in much of the northern domains of the popes, and Lucca, Pisa, and Siena had accepted his lordship; Bologna fell to his armies in 1402. He seemed invincible, and Florence—next in his path and fighting alone—appeared doomed. Florence was immediately saved by his unexpected death in 1402, but this did not end the danger from Milan. Gian Galeazzo's younger son, Filippo Maria Visconti, attempted to reassemble the Milanese empire after 1420. Once again, Florence was in danger of conquest by a despot, though this time it did not fight entirely alone: in 1425 it concluded an alliance against Milan with Venice, its sister republic, and the threat was once more contained. This alliance was also decisive for Venice; henceforth, it would play an active part in the larger affairs of Italy. Meanwhile, between the two onslaughts from Milan, another danger to Florence had come from the opposite direction. King Ladislas of Naples saw in the troubles of the Papal States an opportunity for his ambitions. Early in the 15th century he began to meddle in their affairs, and in 1404 a popular revolt in Rome—encouraged by him—forced the Pope to turn to him for protection. He succeeded in dominating the Papal States, and from this base he twice sent his armies into Tuscany, in 1408–09 and 1412–14.

On each occasion the Florentines had professed to see in the threat to themselves an attempt to subjugate the whole of Italy under a single ruler. Doubtless, they exaggerated both the intentions of the conqueror and the possibility that this might be accomplished. Neither Ladislas nor the Visconti dukes seem to have intended to assemble more than a strong regional state, and their resources were certainly inadequate for the control of the large, diverse, and divided Italian peninsula. Nevertheless, the long period of danger to Florence coincided with the emergence among its citizens of a new political mentality.

Hans Baron, a historian of the Italian Renaissance, has argued that this crisis of Florentine liberty, especially in its most acute phase, between 1400 and 1402, was the major watershed between an age still essentially medieval and the beginnings of a characteristically modern intellectual and political culture. Basing his argument on the chronology of a series of significant Florentine documents, Baron shows that before 1400 most thoughtful Florentines, wearied by constant disorder, had often longed, like medieval thinkers such as Dante, for a strong autocratic government that could preserve the peace of the state and allow them to devote themselves to the private satisfactions of a contemplative life. Already enthusiastic students of the classics, they had, like other medieval men, idealized the benevolent despotism of imperial Rome; their hero was Caesar, who had overthrown the disorderly republic. But now, faced with the loss of their liberty and the prospect of absorption into a larger despotic state, they became increasingly conscious of their heritage of republican freedom and the human values it fostered. Thus, the Humanists of Florence, led by Leonardo Bruni (1369–1444), began to praise the human values of freedom and the obligations of active citizenship. They found their model now in the Roman Republic, in which medieval thinkers had taken little interest, and they applauded not Caesar but Brutus and Cassius, who had assassinated him in the name of liberty. In this way, the citizens of Florence began to formulate a new political ideal of peculiar importance for the future of politics. In addition, conscious, through their awakened love for Florence, of its special identity, they began to consider its development in time; and from this crisis of embattled Florence there emerged the rich tradition of Florentine historiography.

Florentine historiography

The histories composed by Bruni and his successors, in a long series of works that reached a climax with Niccolò Machiavelli and Francesco Guicciardini in the next century, reveal the importance of the contribution of Italy in this period, and especially of Florence, to the formation of modern political attitudes. They exhibit two characteristics hardly present before in the European mind. One is the assumption that historical development proceeds through a succession of natural causes, with the implication that these may be understood by men and to some degree controlled by intelligent and well-informed action. But, perhaps even more important, these historical writings also express a feeling for the particular political community as a concrete and continuing entity that is independent of the men and governments in power at any given time and worthy of human affection, loyalty, and support. In this sense, the historical experience of Italy helped to bring modern consciousness of the state and modern patriotism to birth.

There has been a good deal of resistance to Baron's understanding of the significance of the Florentine experience. Some of it has come out of a reluctance to attribute any major shift in fundamental attitudes to a particular set of episodes concentrated within a very few years. It has also been argued that Baron's dating of the documents is wrong or that the Humanists who gave such eloquent expression to Florentine ideals were only paid propagandists who had no personal commitment to what they wrote. But Baron's case has largely withstood these attacks, and there has been no convincing alternative explanation for the remarkable power of Florentine political and historical thought.

The Papal States in the 15th century. During the 15th century, substantial changes took place in the domains of the pope, following the settlement of the Schism in 1417. Once again the papacy was faced with the difficult task of restoring order in possessions that had fallen apart. And now the difficulties were even more serious. Despots once more controlled the major towns of the Papal States; *condottieri,* mercenary leaders, were carving out principalities for themselves; and, meanwhile, other powers on the peninsula were constantly fishing in these troubled waters. The successes of Gian Galeazzo Visconti had been facilitated by papal weakness, and Milan remained a potential danger. Venice was extending its sphere of influence southward toward Ferrara, one of the more independent towns of the papal domain. There was constant friction along the borders of the Florentine state, and the continuing tension between Anjou and Aragon in the south was a matter of regular concern to the pope. He had also to keep an eye on the possibility of further republican uprisings in Rome. Nevertheless, Martin V (1417–31) made a substantial beginning toward the recovery of papal authority, and his successor Eugenius IV (1431–47) continued the process. Eugenius' decision in 1442 to recognize Alfonso of Aragon as king of Naples, although it resulted in a cooling of papal friendship with Florence, strengthened the security of the Papal States; and by the middle of the century the pope had enough real power to be treated as an equal among the princes of Italy.

But much remained to be done, and control over their Italian domains remained a problem for the popes throughout the century. Even the great Pius II (1458–64), though primarily concerned with restoring papal authority in all Europe and organizing a crusade against the Turks, was forced to devote a large share of his time to the rule of the Papal States, raising armies and negotiating alliances against his own rebellious subjects. Notable among these was the notorious Sigismondo Malatesta, tyrant of Rimini, who was at last brought to obedience, though even he had to be left in possession of that town with the title of papal vicar. Popes of the later 15th century also made use of members of their own families, especially vigorous young nephews, to control their possessions. Unlike most local nobles, such men could be trusted to obey the pope, although the practice led to charges of nepotism that increased the discontent of religious reformers. Sixtus IV (1471–84) was particularly given to nepotism; thus, he made his nephew Piero Riario a cardinal at the age of 25. Meanwhile, the ambitions of such relatives of the pope to carve out territories for themselves also promoted the recovery of papal control. But the task moved slowly, though it received impetus through the conquests of Cesare Borgia, the illegitimate son of Pope Alexander VI (1492–1503).

Papal nepotism

The despotisms of the 15th century. Although dynasties changed, the internal histories of the despotic states of Italy were little different in the 15th century from what they had been in the 14th. The reign of Alfonso I the Magnanimous nevertheless gave southern Italy after 1442 a period of unusual strength. Not only did he rule over both Naples and Sicily, which were thus reunited for the first time in a century and a half; he was also, as Alfonso V, king of Aragon and thus a power in the whole western Mediterranean. Strengthened by papal recognition, he was sought as an ally by the other princes of Italy. He also displayed a personal dignity, an interest in Renaissance culture that made Naples briefly a major centre for literature and the arts, and a strength of character long absent among the rulers in the south. But on his death in 1458 his possessions were divided. Aragon and Sicily went to his brother John, the Kingdom of Naples to his legitimized bastard Ferrante (Ferdinand I; 1458–94) and subsequently to Ferrante's son Alfonso II. This division once more left the kingdom in its earlier state of weakness. In addition, Ferrante proved treacherous, cruel, and incompetent, so that his more powerful subjects, always close to revolt, thought again of resurrecting the old claims of the French House of Anjou.

The new Sforza rulers of Milan behaved much like the Visconti had done. Achieving power in 1450, Francesco Sforza was an able ruler who conquered Genoa in 1463 and meanwhile cultivated closer relations with France; he dispatched his son, Galeazzo Maria, to aid Louis XI in his war against the rebellious French nobility. But Galeazzo Maria, duke of Milan from 1466 to 1476, lacked his father's competence. Cruel and tyrannical, he was assassinated by a group of republican conspirators, although republican sentiment was generally dead in Milan after so long a period of princely control. The assassination was therefore not followed by a popular uprising, and the infant son of the dead duke, Gian Galeazzo II, succeeded his father under the regency of his mother. But the nominal rule of a minor under a female regent was too precarious to survive. In 1480 the young duke's uncle, the ambitious Lodovico il Moro, with the support of both the Pope and the French king, managed to seize control of the ducal government, displacing the boy's mother as regent. Given the unscrupulous habits of the age, this was ominous for Gian Galeazzo II.

Florence under the Medici. While the Papal States were being centralized, and princely government was entrenching itself further in southern Italy and Milan, republicanism was also faring less well in Florence, which, under the concealed dictatorship of the Medici, tended to become increasingly like the despotisms of the peninsula. The old Florentine oligarchy, led by Rinaldo degli Albizzi, had prepared the way for this development by its own ineptitude; between 1429 and 1433 it had failed disastrously in another (and unpopular) attempt to conquer Lucca. This project had been opposed by Cosimo de' Medici (the Elder), a prominent banker of the city, who had made what seemed to the old ruling families a dangerous appeal for popular support, and they had accordingly sent him into exile. But, discredited by defeat, the oligarchy was overthrown in 1434, and Cosimo returned in triumph to assume control of Florentine affairs. He remained a private citizen, however, governing Florence more like a modern big-city political boss than a Renaissance tyrant. The election of officials loyal to himself was assured by eliminating his opponents from the lists of those eligible for office, although the forms of republican government were retained. Initially a popular choice to control the government, Cosimo continued to command broad public support. He maintained order; the lavish expenditures from his private fortune on the patronage of literature, the arts, and especially architecture made him popular with the Florentines; and his ability to keep Florence at peace after so many years of warfare particularly endeared him to the public. In spite of this popularity, however, his dominance saw the beginning of a significant shift in the political climate of the city. Government by an active and concerned citizenry gradually gave way to rule through a bureaucracy responsible only to Cosimo and his successors. The Humanist Leonardo Bruni, now chancellor of the republic, spent much of his later years reading Plato instead of celebrating the benefits of republican freedom.

Cosimo de' Medici (the Elder)

Florentine republicanism was by no means dead, and the indirect nature and tact of Medici rule was a tribute to the continuing vitality of the old republican tradition. Indeed, Cosimo understood that he could disregard it only at his peril. The old families that had previously controlled Florence were constantly restive, and Cosimo felt the need to send some of their leaders into exile. Republican sentiment also gave support to occasional conspiracies against Medici rule. When Cosimo died in 1464, the leadership of Florence passed to his son Piero (1464–69), despite an abortive attempt to restore popular control by a return to free elections. Two years later the republican enemies of the Medici struck at them in the Pitti Conspiracy. But its failure left the Medici more firmly in power than before, and on Piero's death the government was inherited by his young sons, Lorenzo and Giuliano.

The plot against the Medici

Once again an opportunity seemed to present itself for a return to the old order in the city. Enmity had been growing between Florence and the papacy of Sixtus IV over lands claimed both by Florence and by one of the Pope's nephews. It reached a climax in a plot involving both Rome and the enemies of the Medici in Florence, under the leadership of the Pazzi family. In 1478 the conspirators attempted to assassinate both the Medici brothers during a mass in the cathedral of Florence. Giuliano was stabbed to death, but Lorenzo escaped and henceforth ruled alone; meanwhile, Medici partisans hanged the conspirators in the streets, among them the Archbishop of Florence. To revenge this sacrilege, the Pope excommunicated Lorenzo, placed the city under an interdict, and declared war, in which he was joined by his vassal Ferrante (Ferdinand I) of Naples. But at this juncture Lorenzo carried out a sudden diplomatic coup. He made a quick personal visit to Naples, where he persuaded Ferrante to abandon the papal alliance and sign a treaty of friendship with Florence. The crisis was finally resolved with Lorenzo's public apology to the Pope, and matters proceeded as before. Under Lorenzo, known as the Magnificent as much for his personal style as for a patronage of learning and the arts that exceeded even the generosity of his grandfather, the rule of the Medici resumed its development toward something resembling the princely governments elsewhere in Italy. Lorenzo married into the aristocratic Orsini family of Rome, in an alliance that symbolized the acceptance of the Medici by the great nobles of Italy; and, as his negotiation with Ferrante illustrates, he was able to deal on equal terms with other princes. Yet even at this point the rule of the Medici differed from that of the naked despots elsewhere. Lorenzo continued to respect republican institutions even as he controlled them, and it is significant that much of the power of the family, as well as its ability to dazzle Florentines by its generous support of culture, depended on profits from the wide business interests of the Medici. Cosimo had been an astute businessman as well as a politician; and Lorenzo kept a hand in the extensive enterprises of the Medici bank, although, distracted from full attention to business by his political and cultural activities, he allowed too much freedom to the managers in his branches throughout Europe. (The imprudence of the latter led to the decline of the firm, and later representatives of the family were compelled to depend on other sources of income.) But, meanwhile, the ties between Florentine business activity and government persisted, and this helps to explain continuing support for the Medici. In addition, dread of the inconveniences that would arise from further violent changes in government also contributed to their support. Florence had experienced enough disruption in the past, and a sense of relief at the maintenance of order at home and peace abroad, both attributed to Medici rule, worked against further change. Later Florentines would look back on the period of Medici domination as a golden age of prosperity and tranquillity.

Venice in the 15th century. Only Venice, among the great powers in 15th-century Italy, remained true to the substance as well as the forms of its republican constitution, though Venetian society, too, displayed significant changes after the conquests on the mainland. Its wealthier families began a process, which would accelerate in the next century, of withdrawal from the city to newly acquired estates on the mainland, and this group began to develop a way of life similar to that of the ruling groups elsewhere. Meanwhile, the relative equality of the patriciate gave way to an increasingly wide division between poor nobles and a few rich, powerful, and increasingly aristocratic families. Discontent was still kept in check, however, by the general prosperity of this period and by an effective system of poor relief and other social services. Moreover, as Venice came more regularly into contact with the other Italian powers, its ruling group showed signs of an increasingly self-conscious republicanism that contrasted strikingly with the eclipse of republican sentiment in Florence. Complacent and secure, Venice had been backward, compared with Florence, in the development of an articulate political culture. But by the middle of the 15th century Venetians were beginning to celebrate the virtues of their republican constitution, which they interpreted as explaining the remarkable stability of Venetian life, in contrast to the turbulence common to the rest of Italy. These discussions eventually culminated in Gasparo Contarini's *De magistratibus et republica Venetorum* (1543; "Concerning the Magistrates and the Republic of the Venetians"), an early classic of republican and constitutional thought that was widely read throughout Europe. About the middle of the century Venetians also began to take an interest in the history of their own city somewhat similar to that displayed by the Florentines of an earlier generation. The Venetian histories of Marc'Antonio Sabellico and Bernardo Giustiniani marked the beginning of a long and distinguished tradition of Venetian political historiography. During the same period Venice also began its development as a major centre of European artistic and musical life; here, too, it had earlier been remarkably backward. The optimism of this period in the history of the Venetian Republic was only slightly disturbed by occasional wars with the Turks, although the struggle of 1463–70 resulted in the loss of the island of Euboea (modern Evvoia), in the Aegean; Argos, in the Peloponnese; and Scutari (modern Shkodër, Albania).

The rise of Venetian republicanism

Changes in Italian society. Centralization in the Papal States, the long continuation of princely rule in other parts of Italy, and Medici rule in Florence were together bringing profound if gradual changes to Italian society. Habits of dependence on despotic princes became ever more deeply engrained, court ceremonial became increasingly formal, class divisions became more and more rigid, and the way of life of the ruling circles gathered around princes was increasingly differentiated from that of other men. A culture of citizens was slowly being transformed into a culture of courtiers. Men increasingly developed a personal style appropriate for those attendant upon princes, who spent much of their time on country estates and who often cultivated the extravagant ways and even the physical skills of an earlier nobility. Baldassare Castiglione's treatise on courtly manners, *Il cortegiano* (1528), gave eloquent expression to this new human ideal, so different from that of the republican citizen. It has been argued that the social and political changes of later 15th-century Italy therefore prepared the way for the reception of Italian influence at the great royal courts of western Europe in the 16th century.

Italy as a political system. Meanwhile, the conflicting interests of the increasingly consolidated major states of Italy kept them in close contact with each other, and during the 15th century Italy exhibited many of the features of a miniature international system. Some scholars have seen in this system a significant anticipation of the modern principle of the balance of power, a persistent theme in later international relations. Indeed, by the end of the century, Italian observers of the shifting political scene were explicitly using the language of equilibrium to describe its workings.

The emergence of Italy as a kind of system based on the five major powers was possible only after it had become clear that none of them was, in fact, strong enough to absorb the others. Yet even the dangers of Visconti expansion of Milanese territory had led, in the first half of the 15th century, to the creation of a fairly clear alignment of powers: Florence and Venice joined in a republican alliance against Milan, while the Visconti, after Aragonese rulers replaced the Florence-oriented Angevins in the south, found support in Naples.

The alignment became even clearer, though on a somewhat different basis, after the middle of the century. When Francesco Sforza seized power in Milan in 1450, he promptly became embroiled with Venice, which had taken advantage of the preceding period of confusion in Milan to seize some minor territories on the border. At this point Cosimo de' Medici, persuaded that the growing power of Venice was beginning to pose an even greater danger to the interests of Florence than the aggressions of Milan, abruptly switched alliances by supporting the new despot of Milan. The shift may also be taken as a symptom of the decline of republicanism in Florence; with the triumph of the Medici, the differences between Florence and Milan were less important, although the alliance with Milan was unpopular with many of Cosimo's subjects. The diplomatic revolution was completed when Venice turned to Naples, and general war seemed near when Pope Nicholas V intervened as peacemaker. After taking Constantinople in 1453, the Turks seemed poised to invade Italy; and it was also possible that France might intervene in Milan on the basis of claims arising from the marriage of Gian Galeazzo Visconti's daughter Valentina with Louis de France, duc d'Orléans, in 1389. Pope Nicholas V therefore managed to persuade the Italian states of the necessity for mending their differences so that they could present some common front to the outside world. The result was the Peace of Lodi in 1454, in which the coup of Francesco Sforza was recognized by all, and peace was maintained on the basis of the new balance, which aligned Florence and Milan against Venice and Naples, with the papacy as a kind of counterweight. This peace managed a precarious survival for the next 40 years, although imperilled again and again as one state or other attempted to secure particular advantages.

The Peace of Lodi

The result was a series of crises of increasing gravity, which have been compared with those in the 20th century that brought Europe to both world wars. They were so serious because it was increasingly apparent that Italy was not alone in Europe and that great outside powers were more and more inclined to intervene in its affairs. And there were pretexts enough. The Spanish House of Aragon and the French House of Anjou were still rivals for the control of southern Italy, and France had old claims on Milan. Thus, the crises among the Italian states not only illustrated the weakness and division of the peninsula but invited the attention and ultimately the intervention of outside forces of far greater strength.

The first crisis came four years after the Peace of Lodi. On the death of Alfonso the Magnanimous in 1458, Pope Calixtus III (1455–58), incited by Francesco Sforza (who, in turn, was trying to strengthen his own position by promoting the interests of his French allies against his Aragonese enemies), was disposed not to recognize the accession of Ferrante on the ground that an illegitimate son could not inherit the Kingdom of Naples. This brief crisis was ended, however, by the Pope's death. Pius II, his successor, alarmed by the possibility of a French intervention in Naples, recognized Ferrante as king. But in 1460 an even more serious situation developed. An expeditionary force, representing the interest of the Angevin claimant René and based on Milanese-controlled Genoa, invaded the Neapolitan kingdom; with the help of some of Ferrante's own perennially rebellious subjects, it won a series of early victories. Ferrante was saved this time by the arrival of mercenaries from Albania and a revolt in Genoa, and by 1464 the Angevin forces had given up and returned home. But Naples was not the only area of danger. In 1467 a famous *condottiere,* Bartolomeo Colleoni, attempted to carve out a state in northern Italy at the expense of Florence and Milan. Long in the service of Venice, he had probably received Venetian encouragement. But his ambitions were blocked by a rival army under Federigo of Urbino, and the otherwise indecisive battle of Molinella put an end to his hopes.

The Florentine Pazzi Conspiracy

Even more serious was the Florentine Pazzi Conspiracy against Lorenzo de' Medici in 1478, which, given the alignments of the peninsula, had serious implications for the general peace. Lorenzo's success in detaching Ferrante of Naples from the Pope averted general war. A further factor in inducing the Pope to make peace with Florence was the Turkish seizure of Otranto, in the south of Italy, a move that was taken by contemporaries as a preparation for a larger Turkish effort to conquer the peninsula. After 1480, Venetian ambitions led to still another crisis. A quarrel had broken out between Venice and the city of Ferrara over the control of salt production in the northern Adriatic. Ferrara was supported by Naples, Florence, and Milan, all alarmed by the growing power of the Venetians. Venice was backed by the Pope, who wanted to assert his authority over Ferrara, and by Genoa, again revolting against Milanese control. In the course of the ensuing war, troops from Naples and Florence invaded the Papal States. Partly for this reason but partly because he was himself uneasy over Venetian victories in the north, Pope Sixtus IV switched sides. Nevertheless, although now fighting almost the whole of Italy, Venice did well enough to keep some of its conquests when peace was made at Bagnolo in 1484. The most alarming aspect of this episode, however, was the interest in the war displayed by outside powers, an interest that Venice had encouraged. At their moment of greatest danger, the Venetians had tried to attract the new French king, Charles VIII, who had also personally inherited the claims of Anjou to Naples, to invade Italy; and they had promised him their help in the conquest of Naples. Meanwhile, King Ferdinand II of Aragon was negotiating with both sides. Conflicts within Italy were clearly providing increasing temptations to the great monarchies of western Europe.

There were obviously, therefore, serious weaknesses in the workings of the Italian system. Its individual members had no dependable sense of the common interest; Italy was for them, quite literally, only a geographical expression. Indeed, even particular states were badly served by the political situation. The interests of dynasties too often took precedence over the needs of their peoples and certainly over the interests of Italy as a whole. But, above all, the Italian system was not self-contained. Insecure, disgruntled, or ambitious elements in Italy tended regularly to look for support outside. However novel the political history of Italy may have been in some respects during this period, therefore, and however suggestive for the future of European political development, its modernity should not be exaggerated.

Yet, if the Italian states did not entirely anticipate the later conduct of international relations by means of a balance of power, there is no doubt about the contribution of Italy to the techniques of diplomacy. The articulation of Italy into a group of self-consciously independent states that ignored their theoretical unity in a larger Christian commonwealth—whether under pope or emperor—had prepared the way for this development, since diplomacy in the modern sense can be conducted only by fully sovereign states. The needs of commerce and the feverish political actions of the 14th and 15th centuries had impelled the various Italian powers to create suitable instruments for dealing with other powers: foreign offices staffed by able men, which collected information, kept records, and carried on an extensive correspondence; and, above all, a system of permanent ambassadors residing in foreign capitals, commissioned to report on conditions abroad and to negotiate on behalf of the states they represented. The diplomatic machinery developed by Venice was particularly efficient, though it was by no means unique. Venetian ambassadors were carefully chosen for regular three-year terms and periodically transferred from one place to another. Their duties remarkably anticipated those of modern diplomats. They received detailed instructions on being sent abroad; they were expected to maintain a high standard of living, in keeping with the dignity of the republic; they entertained and paid ceremonial visits; and they prepared elaborate dispatches and reports, which are still today among the historian's richest sources of information about all aspects of European society for several centuries. In its development of standards for diplomacy and international communication in a new political world composed of sovereign states, 14th- and 15th-century Italy served as a model for the rest of Europe.

The French invasion. No diplomatic skills, however, could save Italy from the consequences of its weaknesses, which in the end brought about the long-impending tragedy of foreign invasion that was largely to end the independence of the Italian peoples until the 19th century. In the last decade of the 15th century Lodovico il Moro, uncle to the legitimate Sforza duke of Milan, was eager to take the place of his nephew, whom he held a virtual prisoner. The young duke, however, had recently married the granddaughter of Ferrante of Naples, Isabella of Aragon, who in 1490 gave birth to a son. The disposition of the Milanese duchy was now of direct concern to Naples. To solve his personal dilemma, Lodovico, oblivious to the larger interests of Italy, invited the French into the peninsula, in the expectation that they would deal with his enemies in the south and thus open the way to his assumption of the ducal title in Milan.

Charles VIII's ambition

Charles VIII of France was attracted to Italy by various considerations. In addition to the French claims to lordship over both Naples and Milan (which latter claim Lodovico had chosen to forget), he seems to have been influenced by the medieval ideal of a mission for the French nation, on behalf of all Christendom, to set Italy to rights and purify the church; from Italy he dreamed of then leading a crusade against the Turks. Among his advisers, most of whom had more material ambitions, was Cardinal Giuliano della Rovere, a disappointed candidate in the recent papal election and an enemy of Alexander VI. And spiritual impulses emanating from Italy itself may have encouraged such ideals. Lorenzo the Magnificent had died in 1492 and was succeeded by his less competent son Piero. In these circumstances a republican reaction again gathered in Florence and found a leader in the Dominican friar Girolamo Savonarola (1452–98). A powerful and demagogic preacher of repentance, Savonarola began by denouncing the wickedness of his times, daring even to include the Medici rulers in his indictment; he predicted terrible catastrophes as a result of God's wrath on Italy; and he called for reforms that, as became increasingly clear, involved the restoration of the Florentine republic on a basis more democratic than had ever before been established in the history of the city. Since he also strongly denounced the Pope, it is not surprising that Alexander VI (1492–1503) soon became one of Savonarola's greatest enemies. A member of the Spanish Borgia family, Alexander was shameless in exploiting his papal office to promote its interests; and at the same time he was particularly aggressive in imposing his authority over the Papal States, a task for which he was employing his natural son Cesare. Against this—to contemporaries—scandalous pontiff, who eventually excommunicated him, as well as against the more general wickedness of Italy, Savonarola called for the intervention of a foreign "scourge of God," whose invasion and chastisement of Italy would open a new age of righteousness.

Lodovico had probably hoped that the mere threat of French invasion would be enough to deter his Aragonese enemies but that, if it did come, it would move by sea from Genoa; and, indeed, Charles VIII prepared a fleet there under the Duc d'Orléans. But the main thrust of his attack, coordinated with a revolt of pro-French forces in the Kingdom of Naples under Antonello Sanseverino, prince of Salerno, was by land. His army, 30,000 strong, which included Balkan, Swiss, and German mercenaries, as well as heavy artillery of a kind not before used in Italy, entered the peninsula through Milanese territory in October 1494. This development was shortly followed by the death of the young duke Gian Galeazzo, perhaps by poison, and Lodovico was promptly proclaimed duke.

The reaction of the rest of Italy was irresolute and nicely illustrates the failure of the various powers to consider the larger interests of the peninsula. Venice remained entirely aloof, while Florence and the papacy wavered before siding with Naples (and even then they offered only token resistance). Meanwhile, the French advance down the peninsula was rapid. The forces of the few Italian *condottieri* that presented themselves to resist the French were used to a less aggressive style of warfare and quickly collapsed. Soon after the middle of November 1494, Charles had reached Florence. From there he proceeded quickly to Rome, where the Pope promptly came to terms with him in a treaty that allowed the French unhindered passage through the Papal States, and by February 1495 he was in possession of the city of Naples. The rest of the Neapolitan kingdom had rapidly fallen to pro-French forces, in spite of Alfonso II's abdication in favour of his more popular son Ferrantino; and the French king was able to take possession without fighting a major battle.

The consequences of these events were especially momentous for Florence. Discredited by his oscillations and then by a policy of cooperation with the invader, which involved the surrender of important Tuscan strongholds to the French, Piero was overthrown and fled from the city during a revolt that left Florence under the leadership of Savonarola. The Friar then proclaimed the restoration of the republic and sponsored a new constitution, more democratic than Florence had ever known. At first the new regime was immensely popular. It was accompanied by a wave of patriotic and moral fervour marked by dramatic renunciations of the vanities of Renaissance culture and by an enforced purity of manners that briefly transformed Florence into a city of saints regarding itself as a model for the reform of all Christendom. But the extravagances of Savonarola's program, his excommunication by the Pope, and the inability of the new government to recover Pisa, which had with French support rebelled against Florentine rule, led to growing opposition against Savonarola. Eventually he fell from power and was executed in 1498, though the reconstituted republic continued in existence.

Florence under Savonarola

But meanwhile the French were running into difficulties. Their economic and political exploitation of Naples and their brutality proved that the new master was no improvement over the old. Revolts, encouraged by the former Aragonese rulers who had taken refuge in Sicily, had broken out even before Charles started back to France. Even more serious, for the rest of Italy as well as for the French king, was the intervention of Ferdinand II the Catholic, king of Aragon. This had both diplomatic and military aspects. Encouraged by Ferdinand, the states of Italy (with the exception of Florence, whose external affairs remained under French control) at last recognized the need for solidarity. At the end of March, the Pope, Venice, and Milan joined the Spanish king and the German emperor in the League of Venice. Its general purpose was the defense of Italy against aggression, its immediate aim the expulsion of the French. But the adherence of Spain and the empire brought these powers, as well as the French, now regularly into the politics of the peninsula; it meant that, henceforth, Italy would no longer be able to control its own affairs.

The league immediately put an army into the field, which met the returning French forces early in July 1495 in the Battle of Fornovo. Although the Italians on this occasion fought well, the French were able to continue their retreat from the peninsula, and both sides claimed victory. But in the meantime, Aragonese forces had been recovering control of Naples. With the surrender of the French garrison that Charles had left behind, less than a year later nothing remained of the French conquests but an unhappy legacy of intervention that, in the next generation, ended the independence and the liberty of the peoples of Italy until the movement of national unification in the 19th century. A major chapter in the history of Italy had ended.

THE LESSONS OF HISTORY

Machiavelli. The effort to define the significance of this chapter in Italian history, at once so full of promise and in the end so tragic, has been a major concern of European historians of all subsequent generations. It began immediately in the writings of Niccolò Machiavelli, who, born in 1469, had lived through many of the disasters that were to spell the end of the freedom of Italy and, out of an intense patriotism, was concerned to understand their causes. His analysis is of particular interest since it reflects the experience of a direct and highly sophisticated participant in major events. Machiavelli had served as a diplomat and secretary in the restored Florentine Republic after the downfall of Savonarola. When the Medici were reinstated in 1512, he lost his official position; and while in retirement he set down his reflections on history and politics in a number of famous works, particularly *The Prince,* the *Discourses on the First Ten Books of Livy,* and the *History of Florence.* The first of these consists of advice to a ruler on how to secure absolute control over a state, together with a celebrated "Exhortation to Liberate Italy from the Barbarians"; the second includes eloquent passages on the superiority of a republic over all other forms of government, while the third seems not to achieve any definite conclusion. Historians have, therefore, long discussed the mutual relationship of these works, their apparent inconsistency, and the extent to which each may represent Machiavelli's true thought. It may be, however, that their apparent contradictions really indicate his uncertainties about the proper course of action in a bewildering and disorderly world. In any case, these works are instructive as an early effort to evaluate the troubled political history of Italy in the 14th and 15th centuries, for the remedies they prescribe for the maladies of the peninsula, and for their usefulness in revealing the importance of Italian history in the development of the European political understanding.

The Prince

Contemplating the past disorders of Italy, its present vulnerability to foreign intervention, and perhaps most directly the recent instability of his own beloved Florence, Machiavelli saw clearly that something had gone seriously wrong. Comparison with other, more successful polities, especially with the Roman Republic, helped him to identify the trouble. The Italians of the 14th and 15th centuries, he decided, had failed to preserve the political virtues, the decisiveness, and the sense of civic responsibility that had so long characterized the Romans and accounted for their political effectiveness. Their religious fervour, the most effective of social bonds, had declined and for this he blamed the intrusions of the papacy into politics. Since the early Middle Ages, he noted, popes had regularly invited foreigners into the peninsula to serve their own political ends; the result had been both the degradation of the spiritual power and the weakness and disunity of Italy. Furthermore, the rulers of Italy had employed unreliable mercenary armies to do their fighting instead of creating loyal citizen armies; hence, military power in Italy had been too decadent to oppose the challenge from without. And their failures of leadership and their struggles with one another had opened the peninsula to invasion.

Machiavelli then turned to a consideration of possible remedies. Although he considered a republic superior to all other types of government, experience had made him a pessimist. Looking back on what had happened to republics in the past, he developed a view of history according to which the selfishness of men will regularly subvert the state, reduce it to chaos, and require strong and ruthless leadership to set it to rights again. This cycle would recur again and again, human nature never changing, and it seemed obvious to him that in his own time Italy was passing through the most disorderly phase of the political cycle, in which the most urgently needed quality was leadership. These views are probably the explanation for the republican author's flirtation with the idea of a tyrant: an extraordinary problem required extraordinary measures, perhaps even the most cynical and brutal actions, for the restoration of political health. If successful at home, moreover, the prince might be able to organize a general Italian effort to expel the barbarians, though it seems unlikely that Machiavelli, a Florentine to the core, envisaged the formation of a united Italian state. In the long run, however, the prince would play his proper role in the historical cycle if, through sound laws and wise discipline, he prepared his subjects for the restoration of an effective republic—the only kind of political organization capable, Machiavelli believed, of the greatest achievements. Notable here, too, however, is a degree of hope that suggests the inadequacy in his grasp of the contrast between the resources of Italy and the vast power of the French and

Machiavelli's cyclic view of history

Spanish monarchies. It was already much too late for such reforms as Machiavelli dreamed of.

Implicit in Machiavelli's reflections are attitudes toward politics that demonstrate the value of the peculiarly Italian exposure of educated townsmen to the problems of political life. Machiavelli obviously believed that it is useful to analyze political situations and problems, to draw lessons from historical experience, and thereby to establish the principles on which sound political calculations and decisions can be based. In his view, man, by taking thought, can add a cubit to his political stature, at least in the short run. The political virtues can be encouraged through deliberate action by governments; rulers can control events and solve problems and can thereby triumph over the bludgeonings of fortune. To this extent Machiavelli makes explicit that tendency in the Italy of his time to conceive of government as a series of problems in the adaptation of means to ends, as a matter of rational calculation based on a knowledge of men and of the workings of institutions. For this reason (though in other respects he was too passionately committed to warrant such a title), he has been hailed as the father of modern political science. But equally important was his concern with the welfare of the state, conceived as an end in itself. The good was, for Machiavelli, quite simply what serves to preserve and strengthen the state; the bad is whatever tends to destroy it, since states are the only effective source of order in human affairs and, hence, of the happiness of men. From this standpoint all religious and ethical criteria are irrelevant to politics; and "reason of state," the famous phrase now permanently associated with the great Florentine and the source of much of the opprobrium heaped upon him by posterity, is the only measure of political wisdom. In thus making the state independent of all ideal considerations, Machiavelli's thought to some extent paralleled the growing tendency of the Italian states since the 14th century to pursue particularist interests regardless of the common good. Machiavelli, while a witness to the political failure of 14th- and 15th-century Italy, also reflected its most significant political achievements; and he was able, from his study of Italian events, to formulate basic political principles that other, more homogeneous states were later more effectively to pursue.

Later estimates of the period. While Machiavelli was something of a political scientist as well as a historian, Francesco Guicciardini (1483–1540) abandoned the effort to extract generalizations about political behaviour from history. He was a somewhat younger Florentine who had also seen much active political service. By his time the crisis of Italy had become desperate, and thus in his experience the world seemed too disorderly and unpredictable to warrant Machiavelli's type of reflection. But his great *History of Italy,* which reviews the 15th century before concentrating on the events of his own lifetime, exhibits many of the same concerns and the same cool skills in the analysis of events and the understanding of their causes as those apparent in the works of Machiavelli. Guicciardini's picture of the 15th century is highly idealized: it is, for him, an age of unequalled peace and prosperity for which he gives major credit to the Medici. The sequel he represents as a tragedy, for which he blames the blind passions, the selfishness, and the errors of individual rulers—especially the pope and Lodovico il Moro; and he shows in great detail how their machinations brought foreign invasion to Italy. The interpretations of Machiavelli and Guicciardini, widely read throughout Europe, were to become the classic account of Italian history in this period.

Guicciardini's *History of Italy*

This account was little changed until the 19th century. During the long domination of the Italian peninsula by foreign powers, historical composition generally languished. Even less than earlier was it possible to conceive of Italy as a unity about which it was possible to write an integrated history; and students of the Italian past were unable to go beyond erudite compilations of historical data that made sense only in local terms. This was true even of the *Annali d'Italia* (1744–49; "Annals of Italy"), by the great 18th-century scholar Ludovico Muratori, which details the events of Italy year by year but gives little sense of their meaning as a whole.

The modern understanding of Italian history in the 14th and 15th centuries begins with Simonde de Sismondi's *Historie des républiques italiennes du moyen âge* (1807–18; "History of the Italian Republics in the Middle Ages"). Inspired by the romantic liberalism of the earlier 19th century and beginning to think in national terms, Sismondi attributed all that was great in the life of Italy to the freedom of the medieval communes. From this standpoint the 14th and 15th centuries seemed a period of tragic and progressive decline, in which republican liberty was everywhere undermined by tyranny. The failure of the communities of Italy, corrupted by despotism, to unite had opened the way to foreign domination. Sismondi's vision of Italy in the period was also reflected in volume 7 of the French historian Jules Michelet's *Histoire de France* (1833–62) entitled *La Renaissance.*

Sismondi's republican emphasis was largely displaced by the great work of the Swiss historian Jacob Burckhardt, *The Civilization of the Renaissance in Italy* (1859). For Burckhardt, Italy had made a distinct break with its medieval past at the end of the 13th century and thereafter pointed to the modern world in a number of highly significant ways: in the amoral calculations that characterized its political life, in the interest in the human personality and external nature that characterized Renaissance culture, and in a paganism and immorality that pervaded many aspects of Italian life. These tendencies were, however, all expressions of a deeper quality, a fundamental individualism, which Burckhardt considered to be the central feature of the age in Italy. Its cause he found essentially in political conditions: most notably in the anarchy of the Italian peninsula in the later 13th century; he held especially that the dissolution of the traditional sources of order, papal and imperial authority, had created an atmosphere of insecurity and unrestraint that was favourable to the emergence of ruthless individuals. Thus, tyrants, whose power depended on personal gifts rather than on a legitimate relation to larger patterns of traditional order, came to dominate Italian society, making common cause with the Humanists, whose eminence similarly depended on their unique individual gifts. Although Burckhardt called his book an essay, the breadth of its vision of Italian culture as a whole made it a model for a new kind of synthetic history.

Burckhardt's analysis

If its interpretation depended above all on the political conditions of the peninsula, its scope gave special influence to his understanding of this aspect of Italian life. For most of the following century, writers on this period of Italian history tended to follow Burckhardt.

Burckhardt had seen that Italy in the 14th and 15th centuries was not a political unity but a congeries of particular entities united chiefly by common tendencies in political life and also by a largely common culture. But, even while he was writing, the political unification taking place in Italy was producing in Italians a tendency to regard the Italy of earlier centuries as a political whole containing in embryo the national state of the future. The result was, among Italian historians of Italy (though less commonly among outsiders), a revolutionary new vision of the past. Historians such as Carlo Cipolla, in his *Storia della signorie italiane dal 1313 al 1530* (1881; "History of the Italian Lordships from 1313 to 1530"), and Pietro Orsi, author of *Signorie e principati* (1900; "Lordships and Principates"), followed Burckhardt's emphasis on despotism but, lacking a modern concern with those social and cultural elements in Italian life that were common to much of the peninsula, tried to present the political history of Italy as a unified narrative. This effort has persisted in the more recent works of Luigi Simeoni, *Le Signorie* (1950; "The Lordships"), and of Nino Valeri, *L'Italia nell'età dei principati dal 1343 al 1516* (1949; "Italy in the Age of the Principates from 1343 to 1516"). (The latter, however, makes a far more effective attempt to integrate social and cultural with political history.) Such works are characterized by an uneasy tension between their authors' concern to present the history of Italy as a whole and the need to do justice to the intricate wealth of local detail provided in the histories of separate states.

(W.J.Bo.)

Italy in the 16th–18th centuries

EARLY MODERN PERIOD, 16TH AND 17TH CENTURIES

Expulsion of the French. The restoration in Naples of Ferdinand II in 1495 was through the combined effort of military forces furnished by the Venetians, who occupied several important cities in Puglia and meant to remain there; of Ferdinand the Catholic, who sent Gonzalo de Córdoba from Sicily to Calabria; and of Ferdinand II himself, who, landing at Naples, strove to regain the hereditary lands of his ancestors. Defeated in several battles and unable to receive supplies from their homeland because the Spanish fleet controlled the seas, the French finally abandoned southern Italy. Before they left they signed an armistice (February 27, 1497) with Frederick I of Aragon, uncle of King Ferdinand (who had unexpectedly died the previous October).

French abandonment of southern Italy

French acquisition of Milan. The new king was a moderate with humanist leanings; he wanted to pacify the kingdom and consolidate his power. But neither Charles VIII nor his successor, Louis XII (ruled 1498–1515), had given up the idea of acquiring Naples, and they made an agreement with the Spanish to garrison a number of fortresses there. Indeed, Louis XII, bent on enforcing his claim to the Duchy of Milan by a war of conquest, made concessions to the monarchs with whom his predecessor had negotiated the acquisition of Naples. Torn by internal discord, without allies, and poorly defended, Milan easily succumbed (1499).

Franco-Spanish division of Naples. This turmoil had fateful repercussions in the Kingdom of Naples. Neither by lenience nor by arms could Frederick appease the recalcitrant feudal lords, headed by the pro-French House of Sanseverino, to whose branches Charles VIII had restored vast feudal estates. Meanwhile, the acquisition of Milan had put Louis XII in a more favourable supply position; his diplomacy aimed at partitioning the territory of the kingdom with King Ferdinand the Catholic, as agreed to in the Treaty of Granada (1500). The Spaniards had not taken kindly to the fact that Alfonso V had given the Kingdom of Naples to his illegitimate son Ferdinand I in 1458. More seriously, Naples' present weakness stimulated the French and Turks, thus jeopardizing the security of Sicily, to which the kings of Aragon attached the highest importance.

King Frederick approached the Ottoman Empire for help. The latter, emboldened by its successes against the Venetians in the Aegean Sea, seemed ready to spill over into the Mediterranean. Thereupon Pope Alexander VI publicly proclaimed a crusade and called upon the Christian nations to participate in it (1493); this furnished Frederick's enemies with a pretext to invoke the Treaty of Granada.

French and Spanish invasion

Invaded by the French from the north and the Spaniards from the south, in 1501 Naples bowed to the conquerors, who proceeded to divide it according to the prearranged agreements: Louis XII gained Campagna with Naples and the Abruzzi; Ferdinand the Catholic obtained Calabria and Puglia. Frederick of Aragon spent his remaining days in France on a feudal estate and with a pension granted him by Louis XII, to whom he had surrendered his rights to the lost kingdom.

French losses in Italy. But territorial and fiscal differences soon developed between the occupying armies, which degenerated into a war. The Spaniards, led by Gonzalo de Córdoba (el Gran Capitán), forced the French to return to their native land. The two rival monarchs agreed on a three-year truce (March 31, 1504), which held firm; the French, beset by more pressing problems, preferred to allow the fate of the Kingdom of Naples to remain an open diplomatic question.

France attached utmost importance, however, to its possession of Lombardy, because of its high level of culture and because it was the gateway to Italy. The French fought long and ruinous wars for Lombardy, withstanding a coalition formed by Pope Julius II (reigned 1503–13), consisting of the Papal States, Venice, the Habsburgs, and the House of Aragon. Julius viewed France's presence in Lombardy as the real threat to the freedom of Italy, which he identified with the territorial independence of the Holy See.

The military superiority of the Spanish Habsburg bloc, led by Emperor Charles V (ruled 1519–56), prevailed at the Battle of Pavia (1525), and the French were driven out of the Duchy of Milan. Restored provisionally to the last heir of the Sforza dynasty, the duchy reverted to Spanish rule after his death (1535) and remained a feudal dependency of the Holy Roman Empire.

Italy under Spanish domination. Sealed by the treaties of Barcelona (1529) and Cateau-Cambrésis (1559), Spanish Habsburg domination of Italy lasted until 1700, when, as that royal line died out, the French Bourbons and the Austrian Habsburgs vied for the Spanish Habsburg inheritance. The treaties of Utrecht (1713) and Rastatt (1714), acknowledging the transplanting of a branch of the Bourbons in Spain, allotted to the Austrian Habsburgs—for balance-of-power reasons—the inheritance of Ferdinand II of Aragon and of Charles V.

Ruling several states (Milan, Naples, Sicily, Sardinia) by direct rule and maintaining a protectorate over others (including Genoa and Florence), Spain considered Italy a part of its world empire and a rampart against the Ottoman Empire and its satellites, the Barbary States of North Africa.

Economic decay in the 16th century

This situation coincided with a slow general decay that developed in Italy during the 16th century. This decay resulted from diverse causes. The economies of the mercantile states were harmed by the shift in the centre of world trade from the Mediterranean to the Atlantic, following the geographical explorations and discoveries of the 15th–16th centuries. Moreover, industrial, merchant, and banking capital began to develop in central and western Europe, thus eliminating the Italian economic traders, who were now reduced to regional proportions within their own country. The ruin of many public and private fortunes went hand in hand with a depletion of the creative energies that had flourished in Italy at the height of the Renaissance and with a marked decline in civic virtues. Added to this was Spain's political domination—part cause and part effect—which restricted the already limited mobility of such healthy states as the Venetian Republic.

Absolutism, characteristic of the European monarchies of that day, drove Spain to consolidate its rule in Italy. Spain aimed to centralize its administration, even if it was unable to improve the conditions of the people. The old privileged classes found their political influence weakened, yet retained their juridical and fiscal privileges on their huge estates. But these estates themselves, no longer run by watchful and diligent feudal lords, most of whom had been drawn to the cities, now were in the less able hands of managers eager to get rich and climb the social scale.

Spain showed no desire to make serious changes in the administrative apparatus of the state, although demands were urgent. (Studies by experts disturbed by mounting poverty called for reforms in legislation, taxation, social welfare, food distribution, and public health in order to renovate and move this closed, stationary, indolent world off dead centre.)

The authoritarian attitude based itself on the need to protect the state from disturbing confrontations, in both the political and the religious fields. In religious matters this covered not only questions of morality but also intellectual manifestations, judged in terms of formal logic and theological dogma, in which knowledge and faith were held to be inviolably one. Serious breakdowns in morals and discipline had disturbed the Catholic Church, and many requests had arisen for internal reform. In an atmosphere of change, imbued with a feeling for freedom that was inherent in Renaissance culture, the ideas of the Reformation had aroused widespread interest in Italy. The lowest common denominator of this movement was the free examination of sacred texts; in Italy there arose groupings of dissident monks, some of them also political in nature. With the Council of Trent (1545–63), the church carried out its long-projected reform: the doctrinal authority of church teachings and traditions was restored.

Thus Catholicism, having been reformed and having become a sponsor of socially beneficial works, was now adamantly defended, with church and state in complete accord. Press censorship and the tribunals of the Inquisition became the dreaded instruments of the church's rule. A good many intellectuals who had breathed the free air of Renaissance thought were the victims; two of the most noteworthy were Giordano Bruno in philosophy and Galileo in science.

Accord between church and state

Although state and church defended religion as the spiritual cement of the social community, other points of friction disturbed their relations. In the Spanish-ruled states, disputes arose because of the rulers' tendency to chip away at the church's immunity in jurisdictional and financial affairs, in line with absolutist practice. But in Venice, as a result of a conflict with Pope Paul V (1605), there emerged the modern rational principle of the secular state (see below *The Republic of Venice*).

With the poetry of Torquato Tasso, in the second half of the 16th century, the flourishing period of great literary creation in Italy came to a close. Empty formalism gained the upper hand, mirroring the arrogance, ostentation, and frivolousness of the leading classes. Some patriotic poems were anti-Spanish in tenor, but they could not generate and arouse popular feelings of revolt. In the figurative arts the Baroque was a pleasing and novel way of expressing beauty; but this vein soon dried up, giving way to indefinite, sensual, heavy-handed virtuosity in painting and sculpture.

As the 17th century faded, there appeared the first signs of a reawakening of community awareness of the country's needs. Criticism of the administrative structure grew increasingly sharp. Contacts were resumed with the cultures of more advanced European nations. When the Spanish Habsburg line died out with Charles II in 1700, Spain, weakened and helpless, realized that the spread of this reawakening meant the crumbling of its power in Italy.

Spanish Habsburg rule in Naples, Sicily, Sardinia, and Milan. The basic aim of Ferdinand the Catholic and of his nephew and successor at the head of the Spanish states (1516), the Habsburg Charles I (Emperor Charles V after 1519), was to consolidate power in the Kingdom of Naples, eliminating the sources of its internal and external weakness, whether the unruly feudal barons, France's ambitions, or the Ottoman and Berber threats in the Mediterranean.

The Kingdom of Naples. France's futile attempt to conquer southern Italy in 1528 gave the pro-French barons their last chance to rebel. Charles V ordered his forceful viceroy, Pedro de Toledo (served 1532–53), to root out any desire for political power on the part of the feudal aristocracy. Resistance in the capital city of Naples, led by a segment of the rebellious nobility, prevented the viceroy from introducing the Inquisition into the kingdom, a move he had sought in order to crush political opposition as well as Protestant-inspired religious dissidence; but in general he established centralized, absolutist rule. It was tightened by his successors, more concerned with the interests of the king of Spain than with conditions in the Kingdom of Naples.

Traditionally, the Naples Parliament—made up of two *bracci* ("branches"), one feudal and the other appointed by the crown—had been authorized to grant the government power to levy ordinary as well as special taxes, in exchange for various *grazie* ("favours"), but after 1642 it was no longer summoned. Instead, the practice developed of looking upon the municipal government of the city of Naples as the representative of the kingdom. This was a medieval-type aristocratic administration: the executive consisted of six representatives of the city's *seggi* ("districts")—five noblemen and one chosen by the upper and middle bourgeoisie. The one elected representative was generally a tool of the viceroy. Food was the chief worry of the public authorities. With its heterogeneous population continually increasing, Naples became one of the most populous cities in Europe, its lower classes swarming, poverty-stricken, coarse, and quick to riot.

Municipal government of Naples

In the provinces the bureaucratic and military apparatus was a far cry from the needs of a centralized, absolutist regime. Old families, deeply francophile, declined or disappeared; but others took their place, most of them Genoese in origin, rewarded by Charles V and Philip II for their financial aid. Continuous demands for funds forced the government to dispose of more and more crown lands, thus extending the feudal area to almost two-thirds of the kingdom. Meanwhile, the management of the landed estates deteriorated by the large landowners moving to the cities and farming out their estates to contractors who exploited them recklessly. There was a rural and urban middle class in the most important centres of the kingdom, but it was not an independent social force mindful of the common good.

This was a sluggish, inert society, culturally cut off from the forward-moving nations of Europe. Forced to contribute financially and militarily to an empire staggering under its own weight, the Kingdom of Naples could not escape the general decline that characterized the multinational complex ruled by Madrid. These, along with such woes of its own as epidemics, natural disasters, Berber incursions, and princely abuses of power, afflicted the people, intensifying their tendency toward apathy, resignation, and religious fatalism.

Occasional popular explosions of wrath did occur, as when taxes were increased or prices of basic necessities rose steeply. The most significant of these revolts broke out in Naples in June–July 1647, provoked by price increases in a number of staple foodstuffs. It bore the name of Masaniello (Tomaso Aniello), a young fisherman who first led the uprising; but Masaniello was actually a tool in the hands of a clever lawyer, Giulio Genoino, who hated the nobles because of their dominant position in the Municipal Council and vowed to raise the people's representative from his subordinate role. This frightened the viceroy, despite the insurgents' proclamation of loyalty to the King of Spain; he appeased them by granting various concessions, which he revoked as soon as he could strike back. An early victim of this counterattack was Masaniello, who was assassinated on July 16, 1647. The movement then began more and more to assume the character of open rebellion against Spain, while in the provinces the rural and urban masses revolted against the diehard feudal lords with unusual violence. As the crisis worsened, Cardinal Mazarin, prime minister of France under Louis XIII, used aiding the rebels as a pretext for attacking Spain in the Kingdom of Naples. But before preparations were well under way, Mazarin was forestalled by the Duc de Guise, heir to the rights of the French House of Anjou to the Neapolitan throne and a favourite of the French king. The rebels' extremist faction in Naples had proclaimed a republic and invited de Guise to take command, in dictator fashion, of the armed forces mustered to defend it. But de Guise and the republic were swept away by a Spanish expedition to southern Italy; and in the countryside the barons joined with government forces to repress the revolts and restore law and order.

Rebellion against Spain

Still lurking within the hearts of the feudal aristocrats, however, was resentment of Spanish domination and centralized power. Indeed, once Spain had fallen into decay and the conflict sharpened between Louis XIV and the Austrian Habsburgs for the succession to Charles II, the last and heirless descendant of Charles V, a section of the Neapolitan aristocracy conspired on behalf of the Habsburgs (1700; the so-called Conspiracy of the Prince of Macchia). The nobles hoped to restore the kingdom's independence by placing a Habsburg prince at its head and to regain the privileges of the feudal nobility.

The Austrian Habsburgs occupied the Kingdom of Naples in 1707 and had its conquest ratified by the treaties of Utrecht and Rastatt. But they maintained Naples as a viceroyalty and, faithful to their own brand of absolutism, paid no heed to the Neapolitan barons. The Austrian Habsburg government developed no program of reforms, contenting itself with bringing a certain amount of order into the central administration. Moreover, it had fewer financial problems than did Spain, so it could nourish the ambition of controlling all Italy (Sicily was added to Naples in 1718 as a result of the anti-Bourbon coalition of the Quadruple Alliance).

A new crisis developed in Europe in 1733 with the War of the Polish Succession. In 1734 Don Carlos, son of Philip V of Spain and Elizabeth Farnese, was crowned king of Naples and Sicily.

Sicilian loyalty to Spain

The Kingdom of Sicily. Sicily developed institutions and aspects of community life akin to those of Naples. But there were differences, which arose because the island had followed a different historical path after the revolution of 1282 (Sicilian Vespers). After Sicily's absorption by Aragon in the early 1400s, its longing for independence had survived for a time and at moments had even flared up. To dampen any such desire, John II of Aragon consented to make his firstborn (the future Ferdinand II the Catholic) king of Sicily (1460). A subsequent plot to place the island under French rule—hatched in 1523 by the Imperatore brothers and encouraged by the French monarch while the Habsburg Charles V, then king of Sicily, was at war with France—was brutally crushed. The long reign, the prestige, and the political moderation of Charles V, and common problems of security in the Mediterranean, combined to stabilize relations with the Sicilians and make them loyal subjects of the Catholic kings. Spain attached great importance to Sicily, not only because it considered the island a bulwark of its power in the Mediterranean but also because Sicily produced much wheat and was a good market for Spanish goods.

The socioeconomic organization of the island was strictly feudal, with the upper aristocracy, owners of huge landed estates and with hosts of retainers in every segment of society, the dominant class. The barons, maintaining full control of their estates and wielding power in the Sicilian Parliament by means of the feudal *braccio* ("branch"), became part of the constitutional fabric of the regime and one of the pillars of the state.

The Parliament, composed of three *bracci*—feudal, ecclesiastic, and royal—was charged with voting the amounts required for ordinary and special taxes, a task it performed along traditional lines but with the feeling that in this domain it was sovereign and represented the nation. Aristocratic interference was evident even in the administration of the big cities, except for Messina, which was a busy trading centre controlled by a wealthy bourgeoisie. Because the island remained loyal, Spain had no intention of disturbing the existing system; indeed, it even reinforced the spirit of its inherently conservative policies. Nor were there significant changes under Victor Amadeus II of Savoy, who held royal title to Sicily in 1713–18, under the Austrian Habsburgs, who controlled it in 1718–34, or under the Bourbon regime established in 1734.

Sardinia. Sardinia was closely linked with Spain, its firm ties the outcome of a deep-going process of assimilation. Many Spaniards had come to Sardinia and been assimilated into this patriarchal society led by a powerful feudal class, whose chief source of wealth was sheep raising. Neither the crown nor the island's ruling class felt any inclination to alter a system based on a solid feudal-monarchist regime—the viceroy was generally a Sardinian and the Parliament was divided into three *stamenti* ("branches"). The population, peace-loving and inured to harsh living conditions, had only limited relations with Italy. Under such circumstances, Sardinia in 1720 went over to Victor Amadeus II of Savoy, together with the royal title, in exchange for Sicily (see below *The Austrian government in Italy in the 18th century*).

The Duchy of Milan. One of the most prominent states in Italy in the late Middle Ages, Milan was attached to Spain in 1540. When Charles V took this step, after long and fruitless negotiations with France, he was prompted by considerations of Milan's strategic importance as well as its economic development in agriculture, industry, and trade.

Supreme power was in the hands of a governor, assisted by consultative councils; the Senate, patterned by Louis XII after the Parlement of Paris, remained unchanged and controlled the whole administrative apparatus of the state. But overall directives and guidance came from Madrid. To this end, Philip II in 1558 set up in Madrid the Council of Italy, whose members included two councillors from Milan and two from Sicily.

The church had tremendous influence on the government and social life of Milan. In the wake of the Council of Trent, Cardinal Charles Borromeo enthusiastically fostered reforms in Milan; his nephew, Cardinal Federico Borromeo, followed his example. The zeal accounted for many institutions that were socially beneficial—some new, others revived. But when the church claimed full jurisdiction over these institutions, it collided with the Spanish-oriented political rulers, and the old church–state struggle reappeared.

The highborn, on the other hand, lost much of their vigour. No longer involved in their former productive activities, they masked their moral and economic impoverishment with bizarre and ostentatious pomp, called *spagnolismo* ("Spanishism"). And Spain's concern for the economy could not prevent a decline. The process was hastened by military levies and requisitions (in the wars for Monferrato and the Valtellina, which Spain sought to annex), famines, plagues, and the soaring prices of basic commodities. The people gave vent to their discontent by rioting on various occasions. Thus, there was no sense of shock when Lombardy passed from Spanish rule to that of the Austrian Habsburgs, who had long coveted it.

Spain and the independent states of Italy. Relations between Spain and those Italian states that remained independent followed a different course. Functioning in a system that underwent little change until the downfall of the dominant power, some of them may be considered satellites of Spain, others independent entities.

The Spanish satellite states

Disregarding the independent territories in central-southern Italy that were small in area and of minor political importance—*e.g.,* the duchies of Modena, Reggio, and Ferrara of the dukes of Este, the Duchy of Mantua and Monferrat of the Gonzagas, the Duchy of Parma and Piacenza of the Farnese family, the Republic of Lucca—the satellite group included the Duchy of Savoy in Piedmont, the Republic of Genoa in Liguria, and the Duchy (later the Grand Duchy) of Tuscany (Florence) in Tuscany, ruled by the Medici family; the Papal States and the Republic of Venice were independent of Spain.

The Duchy of Savoy. With the Treaty of Cateau-Cambrésis (1559), Savoy, hitherto occupied by the French, was restored to Emmanuel Philibert, victor at the Battle of St. Quentin, with the pledge that he remain neutral both toward the French, guarding the Alps, and the Spaniards, masters of Lombardy. Philibert (ruled 1559–80) rebuilt and strengthened Savoy, maintaining equal distance between the two powers. His son Charles Emanuel I (ruled 1580–1630), avid for territorial expansion, joined in the wars of the period variously allied with Spain and France but without profiting from his participation. At his death the dukedom fell under the ever more oppressive rule of France, which the Bourbon Henry IV restored to the rank of a great power. Not until the advent of Victor Amadeus II (ruled 1675–1730) did the duchy recover from its humiliating subjection.

The Republic of Genoa. Genoa was reduced to a protectorate under Charles V and subsequently under Spain, which looked upon it as a base from which to control the Tyrrhenian Sea. This actually came to pass when Andrea Doria, a powerful Genoese shipowner, defected from the King of France and, with his fleet, entered the service of the Habsburgs (1528), convinced that such a move would favour the economic interests of the city as well as his own. Events vindicated his decision. Doria gained the favour of Charles V by rendering great service in the Emperor's later Mediterranean campaigns; and he prepared a constitutional reform by which the powers of the Genoese Republic would be concentrated in the hands of the faction loyal to him. In 1547 the opposing faction, the pro-French Fieschi, tried to oust Doria's nephew Giannettino; but the movement failed, as did other anti-Spanish uprisings in Italy that same year.

Genoa's weaknesses

The spirit of factionalism died down, but at the same time the mercantile spirit had lost its vitality—the old maritime leaders, losing interest in the sea, began to invest in landed property with a sure return; unemployment swelled; the state staggered under a load of debts. An especially thorny matter was the insubordination of Corsica, harshly exploited by the Banco di San Giorgio, an organ of the Genoese government that administered the island. The dukes of Savoy played on Genoa's weaknesses and its people's distaste for aristocratic government by fomenting numerous conspiracies in the 17th century; similarly, they found Liguria a tempting prize. The republic foiled these manoeuvres; but having refused to break away from Spain, it was bombarded and blockaded in 1684 by Louis XIV. The following year, abandoned by its Spanish ally, Genoa had to submit to Louis XIV's exacting terms.

The Duchy of Tuscany. The Republic of Florence also came into Spain's orbit. Charles V, implementing his agreement of 1529 with the Medici pope Clement VII, conquered the republic by force of arms. Then, in 1531, Charles named Alessandro de' Medici duke of Florence with hereditary rights. Alessandro was assassinated in 1537 in an atmosphere of hatred spawned by his own dissoluteness and of intrigue by republicans and those who favoured an oligarchy. The Florentines ran a serious risk that Spain, using the pretext that the French would intervene to foster a resurgent republic, might occupy Florence. The danger was avoided when, in the same year, Cosimo de' Medici occupied the throne as Cosimo I; he was swiftly recognized by the Senate of Forty and accepted by the people.

The reign of Cosimo I

Cosimo, with farsighted realism, allied himself with Charles V. With the latter's consent he waged war on the Republic of Siena (1552–55), which was pro-French, and annexed its territory except for five seaports that Spain kept for itself; these five constituted the Stato dei Presidi ("State of the Garrisons") and were placed under the dependency of the Kingdom of Naples. The acquisition of Siena extended Florence's rule over Tuscany and enhanced Cosimo's prestige. Moreover, Cosimo worked tirelessly to restore, consolidate, and modernize the state. Inequalities between the capital and the annexed towns were reduced, finances reorganized, land reclamation carried out, ports built, and a strong navy and regular militia created. Attesting to his esteem, Cosimo I had the title of grand duke conferred on him by Pope Pius V in 1569. His successor, Francesco I, was confirmed in this title in 1576 by the Holy Roman Emperor.

Nevertheless, Florentine industry, banking, and trade slowly declined, and Florence lost its former European outlook and became provincial in nature. Land became the pivot of the region's economy and the source of financial profit. Attempts under Cosimo I and his immediate successors to forge an economic policy independent of Spain failed.

The Papal States. The popes Paul III (reigned 1534–49) and Paul IV (1555–59), bent on safeguarding the independence of the Papal States, could not check the expansion of Spanish power in Italy. The papacy saw a lessening of its power as a supranational state as the process of secularization in international relations advanced in Europe. Even its role as head of a major Italian state declined. In 1563 the Council of Trent finished its task of synthesizing the Catholic Counter-Reformation. Implementing the council's resolutions, the popes laboured to reform and re-establish the church, and religious matters occupied most of their time and attention. Pope Pius V (reigned 1566–72) promoted the alliance of Spain, Venice, and the other Italian states that defeated the Turks at the Battle of Lepanto (1571), thus ending Turkish expansion on the high seas. Pope Gregory XIII (reigned 1572–85) reformed the calendar and gave it his name. Pope Sixtus V (reigned 1585–90) also distinguished himself at the helm of state. Pontiffs such as these repressed brigandage, reorganized the court system, and built magnificent public works in Rome. The city took on new beauty as a result of the artistic flowering of the period. Furthermore, the territories of Ferrara, Urbino, and Castro were again brought under the direct rule of the church.

In the never-ending antagonism between Bourbons and Habsburgs in Catholic Europe, the popes were proponents of a balance of power. Yet it was a troubled time for the church because of jurisdictional disputes with Spain (even though that nation was a bulwark of Catholicism), because the papacy's good relations with France deteriorated markedly under Louis XIV, and because the power of the Papal States was eroding.

Papal disputes with Spain

The Republic of Venice. Venice, at the outset of the 16th century the most powerful state in Italy, suffered from defeats in 1509 by the League of Cambriai and from Spain's subsequent takeover of Lombardy, the crowning triumph in the Spanish struggle with France for hegemony beyond the Alps. Venice—a rich, solid, and unified political entity—was far from exhausted. But under pressure from the Spanish Habsburgs straining to expand from Lombardy into Venetia, and from the Austrian Habsburgs jealous of its supremacy in the Adriatic, Venice could no longer count on its alliance with France or with the other Italian states; it weakened politically as well as economically. The growing use of the Atlantic as a sea route stripped Venice of its monopoly in the spice trade; and the expansion of the Ottoman Empire deprived Venice of several Aegean and Black Sea islands and ports of call and sharply curtailed its trade with the countries bordering on those waters.

In 1605 a bitter dispute between Venice and Pope Paul V broke out because two monks involved in common-law crimes had been tried in a secular court. This episode dramatized the diametrically opposed attitudes of church and state: on the one hand, the medieval idea of theocratic universality, which had been repressed but was reaffirmed by Pius V in the edition of the papal bull *In Coena Domini* (1568); and on the other, a modern concept affirming complete state sovereignty in temporal matters. This latter concept, product of a long political and juridical tradition in Venice, found a vigorous advocate in the state theologian Paolo Sarpi, and the republic would not give in, even when threatened with a papal interdict. Mediation by Henry IV of France brought a resolution of the controversy, from which Venice emerged with dignity.

Venice stubbornly defended its Near Eastern possessions against the Turks with its fleet, the chief element in its remaining power on the international scene, and its considerable financial resources. At Lepanto the Venetian fleet made a decisive contribution to the victory of the Christian armada. Worn out by 25 years of war, it had to yield the island of Candia (1669); but it acquired the Peloponnese (formally recognized in the Treaty of Carlowitz, 1699) by defeating the Turks beneath the walls of Vienna in its last great triumph in the East.

In the 18th century the ruling class, as in other Italian states, withdrew from commercial pursuits and, as a result, the vitality of public life declined, making Venice easy prey to the invading French Revolutionary armies in 1797.

(E.Po.)

ITALY IN THE 18TH CENTURY

The aftermath of the War of the Spanish Succession

The early 18th century witnessed profound changes in Europe, arising out of what has been defined as the first world conflict in modern history: the War of the Spanish Succession (1701–14). The old political system was no more; from its ashes arose a new pattern of state relations, ratified on the diplomatic level by the treaties of Utrecht (1713) and Rastatt (1714). In Italy the Duchy of Milan and Mantua, the Kingdom of Naples, and Sardinia passed to the Austrian House of Habsburg. Sicily went to Victor Amadeus II of Savoy, who bore the title of king; a succession of surrenders, however, forced even Sicily into the Austrian orbit, with Sardinia transferred to Victor Amadeus II as compensation (Treaty of The Hague, 1720). Meanwhile, Tuscany, when the Medici dynasty became extinct, went to Francis Stephen (Peace of Vienna, 1738), duke of Lorraine and husband of Maria Theresa of Austria. This intricate diplomatic game, punctuated by acute military crises, continued until the Treaty of Aachen in 1748. By its terms Milan was ceded to the Habsburgs in exchange for some slight territorial adjustments in favour of the Piedmontese; Don Carlos, the Bourbon Infante of Spain, was confirmed as king of Naples and Sicily, which had been conquered in 1734; and his brother Philip was accorded the Duchy of Parma and Piacenza. Thus, in approximately half a century, the political situation in the Italian peninsula had completely changed and now settled into a long state of equilibrium. This new situation gave the Italian states undeniable advantages. They found themselves incorporated into a political system that proved more vital and energetic than the previous one and that bore a distinctly European cast. Italy was thrown open to the ideas of the Enlightenment and swept by profound desires for reform in every field. Obviously, a historical process so vast in scope could not be effected swiftly and along predetermined lines; it was conditioned by a variety of factors growing out of specific local situations and the overall political evolution.

Lombardy. The Austrian armies, with the support of Piedmontese troops, entered Milan on September 26, 1706. On March 13, 1707, an armistice with France formally ratified Austrian occupation. Austria sought to impose its own Italian policy firmly and decisively, using its base in Milan to strive for complete control of the peninsula and to expel potential rivals. But it was thwarted by an unfavourable set of circumstances. Internal crises and international complications forced the Habsburgs to extract financial resources from their new possession; and their methods were at times crude and arbitrary. As a result the governors of Milan had to resort to a policy of drastic financial retrenchment, thereby leaving a thin margin for any attempts at governmental reform.

Government reforms. After the storms of the Spanish succession had subsided, the foundation for future reforms was laid. But a terrible economic depression in the 1730s, on top of a renewed outbreak of devastating war in Lombardy, blocked these meagre efforts at reform and paralyzed government activity.

The great powers of Europe concentrated their aims and efforts at preventing Austrian domination over Italy and half of Europe; and they were joined by the House of Savoy, dispassionately opportunistic as always. These were the critical years of the wars of the Polish and the Austrian Succession, during which Austria lost Milan more than once, only to finally regain it in 1747. When things returned to normal with the Treaty of Aachen, the Austrian government readied a plan for a long-overdue and sweeping reorganization of the state. Recent events had laid bare frightening gaps in Milan's economic and social structure. Its technicians and administrators, most of them outrun by events, remained a dangerous source of tensions, so that the government was forced increasingly to intervene in order to safeguard its power. In its interventions it drew on long experience in centralized rule, so that local political elements and their institutions were stifled. Reorganization of Milan was decided upon from 1748 on. Despite obstacles and resistances interposed by the privileged and conservative elements of the population, many varied and weighty problems of the economy were tackled.

Reorganization of Milan

After 1750 important reforms were instituted in the system of labour contracts and tax collections; the administrative code in force under the Spanish was simplified; and a land survey was made. These reforms were technically valuable. But even more important was the logic inherent in the overall reform policy, which was essentially designed to strengthen the central power at the expense of local authorities and of special interests that were particularist and anti-centralist.

The decade 1750–60 ended on a positive note for the Austrians in Lombardy. Their activity there had undoubtedly struck the first serious blows at an antiquated state apparatus and system of living. Once under way, therefore, the process began to develop an independent and irrepressible momentum of its own. Requests and demands of the population grew more and more insistent; the solutions adopted, usually compromises, proved less and less adequate.

A new and more powerful wave of reform hit Milan after 1760, involving again the system of labour contracts, customs, and tolls. This led in 1765 to the creation of a higher body of study and verification, the Supreme Economic Council. The reorganization of the local administrative and political apparatus continued, chiefly at the expense of the Senate, the judiciary, and the highest public offices (even the governorship was reduced to a bureaucratic function). Nor was this done purely by chance: it was in those three areas that the Austrian government had encountered the stiffest resistance. The reforms were always carefully circumscribed so as not to provoke violent reaction. This was also true of church–state relations, with the abolition of the Inquisition and the secularization of censorship while problems of a delicately theological nature were set aside for a time.

The rule of Joseph II. Maria Theresa's death (1780) and the autonomous rule of Joseph II mark a radical turning point in Habsburg policy toward its possessions. The new emperor was, spiritually and ideologically, the product of an age of rationalism and enlightenment, and his behaviour differed markedly from his mother's ponderous and prudent ways. What had been calculation and necessity in Maria Theresa became in Joseph dictate of conscience and adherence to principles; hence the flexible style and comportment of the former and the stubborn, rigid, and unyielding behaviour of the latter. The idea of the state as the organ of absolute power above and against traditional rights, class privileges, and local autonomy was at the root of Joseph's personal and political morality. He promoted centralism and denied any autonomy to persons, institutions, or regions in the state. Like the other Italian possessions, Lombardy was caught up in the fever of the overall reorganization undertaken by the Emperor with utter disregard for consequences. A Government Council made up of six departments suddenly supplanted existing administrative and judicial bodies. A veritable storm blew up over church–state relations, but this time the controversy spilled over into theology as well, including matters of church ritual, in which the government interfered arbitrarily and at times ridiculously. Such passion for reform, though it furnished the Milanese with the tools for modernizing their state and rationalizing its developing economy, did not always find favour with the various segments of the population. Significantly, the intellectuals abandoned the line of collaboration with the Austrian government that they had followed under Maria Theresa and took the lead in opposing what they now called the new despotism. The price the Milanese had to pay for the Emperor's policies in Lombardy must have seemed too high; Joseph II's efforts, massive but lacking flexibility, left the real needs of the state misunderstood or neglected.

Enlightened Despotism

Leopold II, who succeeded his brother in 1790, made an attempt to improve relations by softening Joseph's harshest measures; but Leopold died barely two years later, and his successor, Francis II, was too much absorbed in international affairs to be concerned with Lombardy. Defeated by the French in 1796, the Austrians ended wretchedly, at least for the moment, their political presence in northern Italy.

Tuscany. The Peace of Vienna (1738) ratified on the diplomatic level the results of the War of the Polish Succession. By the terms of the peace, Tuscany was awarded to Francis of Lorraine (reigned 1738–65), and a new phase in the history of the grand duchy began.

The government of Francis of Lorraine. The functions of government were assigned to a Regency Council. With Francis at war with the Turks, two of his representatives presided over the council's work, first Prince Marc de Craon and then Count Emanuel de Richecourt. The ministers were confronted with corruption and malfunction in every branch of the government, both civil and military. These ills were piled on top of the basic problems of Tuscan society inherent in the country's political and economic structure, including the overweening power of an oligarchy over the regency councillor, the pro-Spanish inclinations of a section of public opinion, the disorganization of trade, and the burden of tax-free church property. It quickly became clear, above all to Richecourt, that it was essential to formulate a detailed and thoroughgoing plan of reform that would wipe out abuses and unfair privileges of the elite castes. But the reactions to Richecourt's initial measures and the delicate state of international affairs made it imprudent, for the time being, to proceed along the path of reform. The Spaniards, meanwhile, tried unsuccessfully to install the Bourbon prince Philip in Tuscany. But at the beginning of 1739 Francis of Lorraine entered Florence, visited the principal cities, and departed, leaving in charge a regency again headed by Richecourt. Until 1765 Francis' administration sought to promote general economic progress and welfare. It had some striking successes, particularly in getting rid of institutions that still bore the medieval stamp. Marked improvements were made in administering the economy: trade was liberalized and the public debt reorganized; measures were taken to benefit agriculture. The government moved to rescind feudal legislation and abolish church privileges. In sum, Francis' achievements were far from negligible, even though his political activity in Tuscany had to remain quite limited, since Tuscany was only one of the many provinces in the Habsburg Empire. Francis was crowned Holy Roman Emperor in 1745 (as Francis I), and thenceforth had to immerse himself in far graver and more complicated political problems.

Economic improvements

The government of Peter Leopold. Only with the 25-year rule of Grand Duke Peter Leopold (reigned 1765–90) was Tuscany vouchsafed a genuine revival, with an intense social and political development that no longer benefitted special interests alone. This second son of Francis of Lorraine remained in Florence until 1790, when he ascended the imperial throne as Leopold II. With his advent in Florence, relations between the grand duchy and the court of Vienna were completely altered, though he evoked frequent expressions of reserve and concern in circles close to Empress Maria Theresa by placing in responsible posts Tuscan experts and men of enlightenment, with whose help the most effective reforms were launched. Important steps were taken toward ending internal restrictions on the grain trade; the traditional farming out of taxes was abolished and the tax system overhauled. Immediately thereafter, laws were passed regulating the church's landed property; the enormous holdings of the church were thus broken up and redistributed more equitably.

The decade 1770–80 saw other significant reforms: the abolition of a law defining the jurisdiction of corporations; the suppression of specific statutes and tribunals as well as the high tribunals of Commerce and of the Arts and Professions; and the stimulation of commercial and industrial activity by jettisoning various duties, tolls, and privileges. The entire bureaucratic and administrative state apparatus was fundamentally transformed. The antiquated state structure set up to serve the city-state of Florence was replaced by a modern organization in which the interests of individual institutions coincided with the general interests of the grand duchy. Nor were problems of public safety or those pertaining to military matters ignored; special attention was paid to draining and reclaiming the malaria-infested marshes—an age-old barrier to development of the whole region.

In the decade 1780–90 Peter Leopold included programs of even vaster scope; *e.g.,* church reform, a new constitution, and efforts to increase the peasants' landholdings. He did away with tax exemption for the church, its chief privilege under the old order, and he abolished the Inquisition and suppressed the Society of Jesus (the Jesuits), despite sharp resistance. But when he moved onto jurisdictional ground and sought to reform the church along clearly Jansenist lines he aroused fierce opposition.

A basic factor in restructuring the economy was the grant to small landowners of vast holdings that had formerly been state property. To strengthen and consolidate Austrian rule in Tuscany, Peter Leopold deemed it essential to create a new class of independent small farmers, who would constitute a genuine social base for the government, and he planned to transform the grand duchy from an absolute to a constitutional monarchy, in which the people would have representative bodies—a revolutionary step. Apparently, Peter Leopold was convinced that the monarchy could continue to function solidly only by strengthening its relations on the political and institutional level with the new social strata. The beneficiaries of the agrarian reform thenceforth constituted the pivot of the social structure in Tuscany. His plan for a constitutional monarchy, however, did not materialize because of the still considerable weakness of the new social groups on which the future political structure was to be based, and because Vienna was opposed to any such political orientation on the part of its Italian dominions.

The new Austrian class in Tuscany

In 1787 a new penal code, the Leopoldine code, was promulgated. Calling for the abolition of the death penalty and torture, it broke sharply with the tradition of judicial cruelty and thereby propelled Tuscany, in this respect, into the vanguard of the nations of Europe. In 1790 Peter Leopold was called to Vienna and, on the death of his brother Joseph, named emperor; in March 1791 his son Ferdinand succeeded him as head of the grand duchy (as Ferdinand III).

Leopold's forced departure and the transfer of his powers brought violent disorders in Tuscany in 1790 and again in 1795. They appeared first in an anti-Jansenist guise but quickly turned into a revolt against hunger, poverty, and the high cost of living, which the people attributed to their lack of freedom. The influence of the contemporaneous French Revolution was strong. Faced with such disorders, the regency yielded, and Leopold, in Vienna, was forced to sanction, for the moment, the repeal of free trade in grain. What had in fact occurred was a reaction to the reform policies of 25 years of Leopold's rule, fomented by the clergy and civil servants who had remained hostile to the innovations that had stripped them of prestige and power. In addition, the poorer classes, victims of social inequities and hard hit economically, were exploited and manipulated by the reactionaries. The result was crisis and decline, virtually until 1799.

Naples and Sicily under the Habsburgs. *The mainland.* Naples, too, was acquired by the Habsburgs as a result of the War of the Spanish Succession. Austrian troops entered Naples in 1707 and were warmly received by the people. The most active segments of society—the provincial barons, the urban patricians, and the secular middle classes—each formulated their demands in a different way. But in substance each group demanded a greater measure of autonomy designed to strengthen local powers, a viable economy that could cope with anticipated financial pressures, and a reorganization of the juridical and administrative system. The delicate international situation, to which the Habsburgs had to devote their undivided attention, quickly revealed how unattainable all these demands were. Nevertheless, something was done on behalf of the viceroyalty of Naples when Cardinal Vincenzo Grimani became viceroy (1708–10). While vigorously pursuing a line of stern authority, he drew upon the best energies of the native population and endeavoured to soften social differences and to lighten financial burdens. But in the final years of the War of the Spanish Succession, these needs became more pressing, and the Austrians had to cast about for funds. Thus, the efforts to recognize the state's finances were rendered vain, and Naples once more found itself on the brink of collapse.

Demands for change

When the war ended, the Austrians were able to pay more attention to their new acquisition. The choice of Wierich Lorenz, Graf von Daun, as viceroy (1713–19) signalled a new political course. Daun embarked on the thoroughgoing and long-term project of restoring the economic and financial health of the state. He also engaged in a controversy with the papacy over church government, over which the state traditionally enjoyed extensive control. During these years such institutions as the Collateral Council, an expression of local political power, were strengthened; university reform was attempted; and notable successes were registered in the controversy with the church. A series of international complications, however, forced Charles VI to adopt a financially oppressive policy toward Naples. Vast sums of money went more and more frequently to Vienna, under the name of donations, further impoverishing the land. In exchange, grants and privileges were extended, but almost exclusively to the feudal groups. Thus, Neapolitan society was dealt a setback.

With the return of peace and the arrival in Naples of a new viceroy, Cardinal Michael Friedrich von Althann (1722–28), Habsburg policy took a new turn. Reinforcing the central government, reviving the economy, and easing relations with the church, in line with improved relations between Rome and Vienna, were the aims Althann tenaciously pursued. But his policies offended several groups; these included the nobility, hit hard by the harsh tax measures and forced to relinquish their traditional powers in such areas as the judiciary, and the secular-minded citizenry, resentful of the viceroy's new pro-church attitude. Despite opposition, Althann first planned a new enumeration of the hearths in the country, in order to get a clearer picture of population distribution and thus institute fairer taxation. Next, he sponsored the creation of the Banco di San Carlo, a move that provoked dangerous social tensions. The bank was established with a view to reacquiring large estates for the crown as well as to recoup lost revenues, rather than for serving as a stimulus to revive stagnant economic activities, especially in trade, by distributing large amounts of capital to individual citizens. The nobility saw in the bank an assault on the rentier economy on which it lived and prospered; the middle classes feared it would become the instrument for even more burdensome taxes. The social unrest these measures caused led to Althann's downfall.

The following years brought famine, economic crisis, and an atmosphere of imminent war, which forced the situation to the breaking point. The Austrians could not escape the consequences of the crushing financial burden; since the latter overhung all their policies, every attempt at reform was thwarted. When Don Carlos of Spain ascended the throne of the reborn kingdom, the population and the political authorities welcomed him warmly.

The reign of Charles VI

Sicily. Following the Hague Treaty (1720), the Habsburgs yielded possession of the port of Antwerp, in deference to the requests of England and Holland, and recognized Philip V as king of Spain. As compensation Charles received Sicily, taken from Victor Amadeus II of Savoy, in exchange for Sardinia.

Sicily suddenly proved to be a difficult land to rule, even though Charles VI and his ministers did not stint in their efforts to win over the Sicilians. The first period of the new government was made even more difficult because of the permanent garrisoning of German soldiers on the island, thus giving rise to numerous abuses and clashes. When a functioning civil government was finally set up, Sicily's age-old ills—feudal arrogance, administrative and economic disorder, corruption and chaos in the courts, and municipal particularism—became glaringly evident. An excessive tax load was placed on taxpayers already staggering under their burden, and on many occasions the island parliament was called upon to vote huge levies needed for imperial policy.

But in at least some respects Austria's rule was energetic and mindful of Sicilian interests. Charles VI really tried to lift the island out of the economic and commercial swamp in which it was foundering. He sought to reactivate Sicily's ports, particularly Messina, which was made a free port in 1728, with the aim of reviving the economy in the whole Messina area, especially by attracting foreign commerce and shipping. Measures were taken to cope with a crisis in grain—Sicily's traditional source of wealth—and to improve the declining silk industry and allied activities. The results, however, were far less than anticipated, largely because of a drastic worsening in the general economic situation, and to some extent because Charles' policies promised more than they could achieve. A series of measures adopted around 1730 produced disastrous results, laying the groundwork for economic collapse and probably hastening the political breakdown.

As for relations with the church—especially tricky in Sicily, where the sovereign had the role and functions of apostolic legate—Charles VI proceeded cautiously but firmly. He reassured everyone by his orthodox defense of the faith, even permitting the Inquisition to continue to function, but he was adamant about maintaining the legateship. Aided by loyal and able church ministers, he waged a lively and eventually victorious polemic with popes Innocent XIII and Benedict XIII. The solution, arrived at in 1728, was completely satisfactory to both parties and was given force of law in a papal bull.

The few years of Austrian rule failed to make a dent in many other aspects of Sicilian society. Austria's monetary policy, for example, failed and its cultural policy was weak. Defeated by a Spanish army, the Habsburgs left Sicily in 1734. (G.d'A.)

The first Bourbon period in the south (1734–99). Don Carlos ruled Naples–Sicily as Charles VII, winning popularity by making it an independent kingdom for the first time in two centuries. When he succeeded to the Spanish throne as Charles III in 1759, he left the kingdom to his son Ferdinand, appointing his minister Bernardo Tanucci and a council of regency to rule until Ferdinand came of age in 1767.

In the prevailing spirit of Enlightened Despotism, Tanucci sponsored reforms to modernize the state and increase its power at the expense of traditional institutions. Ferdinand had little aptitude for government, and came to be dominated by his wife, the Austrian archduchess Maria Carolina, whom he had married in 1768. Opposed to Tanucci's pro-Spanish policies, Maria Carolina secured his dismissal in 1776 and, beginning in 1779, replaced him by promoting the rise to power of an English émigré, Sir John Acton.

In the 1790s, Acton and Maria Carolina aligned Naples with Austria and Britain in their struggle against Revolutionary France. When French forces occupied the mainland portion of the kingdom in 1799, the Bourbon government sought refuge in Palermo under British protection. (Ed.)

Revolution, restoration, and unification

THE FRENCH REVOLUTIONARY PERIOD

In the spring of 1796 the French Revolutionary armies burst into Italy. But they had been long preceded by a considerable incursion of revolutionary ideas. As early as the summer of 1788, the Italian newssheets had given priority to "the latest news from France," where a grim struggle between the crown and the Parlement of Paris was taking place. As the Revolution developed, the circulation of these papers increased; they were soon accompanied by a spate of pamphlets, and then, from 1791 onward, by the graphic testimony of émigrés. Despite the vigilant surveillance of the various Italian governments, revolutionary ideas spread widely. Italian public opinion, however, was seldom able before 1796 to distinguish the different forces at war in the political life of France, and the simplistic image of two monolithic fronts—monarchists on the one side, and revolutionaries on the other—remained the prevalent one.

The early years. By 1789 the period of reforms in Italy had come to a close. But those who had hoped much from the work of the enlightened princes had been severely disappointed. The reforms had not widened political power, nor had constitutional steps been taken to confer administrative and governmental responsibilities on the educated classes, the landowners, and the entrepreneurs. Only now, in the light of the French example, could things perhaps be changed.

The rise of the Masons

In the Italy of the old regime, there had been no representative political life. But the increase in the number of Masonic lodges at the end of the 18th century demonstrated the desire for secret discussion of problems different from those that were agitating the academies and the agrarian societies. Not all the Freemasons became supporters of the Revolution and of the French, but many of them did so. The moderate and constitutional demands of the Masonic lodges began to be accompanied by more democratic demands, and there were in Milan, Bologna, Rome, and Naples cells of Illuminati, republican freethinkers, after the pattern recently established in Bavaria by Adam Weishaupt.

The Italian governments were unanimous in opposing France and the ideas of the Revolution. Piedmont actually joined the First Coalition, an alliance made in 1793 of powers opposed to Revolutionary France. Savoy and Nice were invaded as early as the autumn of 1792. Although the King of Naples was forced to yield when the French fleet threatened his capital in December 1792, the other states pursued a policy of stern police repression. Denunciations and trials show how the people of the various Italian states looked to the "French system" as the only effective remedy for their own grievances. In 1792 the Piedmontese tenant farmers, reduced to starvation by the great capitalistic landlords, reminded the King of what was going on in Paris; at Rome the bourgeois entrepreneurs protested against clerical misgovernment and against the temporal power of the papacy; and in the Venetian provinces the nobility and the bourgeoisie brought charges against the aristocratic regime of Venice.

These hopes took concrete form as organized conspiracies in only two states—Piedmont and the Kingdom of Naples. In the south the first pro-revolutionary centres developed in connection with the Masonic lodges; an example is the Celestini Lodge at Naples. But bourgeois elements, with republican and democratic ideals, soon broke away from these and evolved a conspiracy, which was discovered, and the leaders executed in October 1794. The trials, followed by many arrests, sent a stream of emigrants flowing into France, where they later became significantly active.

The Oneglia experiment

Both those who did no more than complain and those who had the courage to conspire hoped to make their countries into modern states, with new, impartial laws and where subjects had a share in politics and government—aims that were sufficiently moderate. But the emigrants, who put themselves at the service of the French government, had a much clearer consciousness of the real aims of revolution. Perhaps the most important among them was Filippo Buonarroti, a Tuscan of an ancient noble family who emigrated to Corsica in 1789 and became a most active revolutionary agent; then, in 1794, he was attached to the French Army of Italy and was appointed National Commissary at Oneglia, a Ligurian town conquered from the King of Sardinia. Here he established a republic based on the views of the French revolutionary leader Robespierre, rallying the Italian exiles, abolishing seigneurial rights, and instituting the deistic cult of the Supreme Being. This extremism was disapproved of by those who gained power in 1794 when Robespierre and the Jacobins fell, and, after less than a year, Buonarroti was recalled to France (March 1795) and sent to prison; after his release he took part in the conspiracy for an armed rising planned by François-Noël (Gracchus) Babeuf and discovered in May 1796. But the example of Oneglia was never forgotten by those Italians who took their ideas from Robespierre and the Jacobins.

French invasion of Italy. The French campaign in Italy, which led to the rise of Napoleon Bonaparte, began in March 1796. In April the Piedmontese army was defeated, and, by the Peace of Paris (May 15, 1796), King Victor Amadeus III of Sardinia was forced to cede the Transalpine provinces and grant the French armed forces passage. On the same day Napoleon entered Austrian-owned Milan; then he pursued the Austrian Army into the territory of the Venetian Republic. During April 1797 the whole Po plain fell into the hands of the French, and the Peace of Tolentino (February 19, 1797) obliged the Pope to surrender Bologna and the northern Papal States. The duchy of Modena was occupied, and the French pushed on into Tuscany as far as Livorno. After defeats in Venetian territory at Arcole and Rivoli (winter 1796–97), the Austrians capitulated at Mantua. With his rear thus protected, Napoleon turned his offensive northward and, crossing the Tagliamento River, entered Austrian territory and by April 1797 was in close reach of Vienna. At Leoben the Austrian plenipotentiaries halted his advance toward the capital with "Preliminaries" (negotiations held on April 18, 1797) that anticipated the partition of Venetia and recognized Napoleon's conquests of Belgium and Lombardy. In the period of peace that followed, the peninsula enjoyed a short period of democracy, which was ended by the Austro-Russian offensive of April 1799.

It is to this brief but decisive period in Italian history that the origins of the Risorgimento, the great Italian national revival of the 19th century, must be traced. Insofar as the Risorgimento involved the formation of political groups affirming the right of the Italian people to achieve a government suited to its desires and its traditions and the growth of a feeling of nationalism and individual responsibility, it certainly began at this time.

The Neapolitan historian Vincenzo Cuoco wrote in 1800 that the Italian Revolution had been a "passive revolution" that, imported from France, had no real roots and had not been the expression of a national governing class. This criticism of the Jacobin rule in Italy between 1796 and 1799 has been continually repeated up to the present day. Historians, feeling the need to distinguish among the several types of republicanism, have characterized the strictly "Jacobin" group as ideologically descended from Robespierrism and the heroic days of the French Terror of 1793–94. But this is to attribute too much ideological rigidity to men who often, from political necessity, shifted their position. It was with no sense of inconsistency that those who had supported the most radical republican democracy later assumed administrative and governmental responsibilities in the Napoleonic Kingdom of Italy from 1805 and in Joachim Murat's (French general; Napoleon's brother-in-law) Kingdom of Naples from 1808.

Yet, among the Italian Francophiles, some distinction needs to be made between the moderates and the extreme democrats. This lay essentially in the different meanings given by each group to the concept of popular sovereignty, to which all alike paid lip service. The doctrine of equality, for instance, could be restricted to a doctrine of equal rights before the law or enlarged to shake the foundations of private property. The differing views of the two groups could also be seen in their attitude to practical details such as taxation, schemes for public education, and for regulating industry and the labour market.

Napoleon's notion of the "sister republics"

The Italian republics of 1796–99. Meanwhile, political initiative was entirely with the French. The Directory, the government set up in France following the adoption of the moderate Thermidorian constitution (known as the Constitution of the Year III), regarded the unexpected conquests in Italy primarily as a bargaining point, but Napoleon, commander of the armies in Italy, strongly favoured the rise there of "sister republics." To organize such new states, which would accept the French hegemony and show promise of financial, political, and administrative stability, Napoleon realized that he must support not the democrats but the moderates, who were in a position to control the economy and public opinion and to crush any possible popular uprisings.

The prevalence of conservative and moderate forces in the cities of Emilia (especially in Bologna) persuaded Napoleon and the agents of the Directory to found the first democratic state there. Thus arose the Cispadane Republic, which, at the Congress of Modena (ended March 1, 1797), adopted a constitution modelled on that of Thermidor but with perhaps greater emphasis on limiting the hegemony of the Catholic Church. Lombardy, where the political struggle was more intense and the democratic party more active, was kept for a longer time under a provisional government, so that the "sister state" in that area, known as the Cisalpine Republic, was not proclaimed until June 29, 1797. The Cispadane Republic was fused with it a month later, Napoleon and the Directory regarding the danger of setting up an overstrong state as offset by its anti-Austrian function and by the weakening of the Italian democrats that would follow from their subjection to a central government more easily controllable from Paris. Yet the Cisalpine Republic proved to be the most restless of the states that the French set up in the peninsula—witness the suppression there of newspapers, the temporary detention of journalists and writers, the dissolution of democratic clubs, and the necessity in 1798 of organizing no less than four coups d'etat to exclude from the two legislative assemblies (Consiglio degli Juniori, Consiglio dei Seniori) and the Cisalpine Directory those who most strongly resisted orders from Paris. The moderates now began to emerge as the coming ruling and bureaucratic class, not only because they were protected by the French but also because they had actual capacity and previous administrative experience.

The Ligurian Republic, proclaimed on June 6, 1797, after uprisings there by the pro-French "patriots" against the aristocratic government, had a less troubled history. The moderates, many of them members of the old aristocracy and working hand in glove with the Directory, always retained control and quashed the democrats' hopes of fusion with the Cisalpine Republic.

In northern Italy there were no other Jacobin republics. In Piedmont, the King, after suppressing a series of conspiracies with much bloodshed, was forced by the French to leave the country in December 1798; in February 1799 the kingdom was annexed to France. Venetia, already tampered with by the "Preliminaries" of 1797 drawn up at Leoben, was ceded to Austria by Napoleon by the Peace of Campo Formio (October 17, 1797), which marked yet another stage in the Italian democrats' disillusionment with the "liberators," notably shown in Ugo Foscolo's novel *Le ultime lettere di Jacopo Ortis* (1798; "Last Letters of Jacopo Ortis").

Invasion of the Papal States

Though the First Coalition formed to resist the French had now been dissolved, French penetration into Italy continued, and, as a result of the Pope's hostile attitude and the revolutionary ferment in Rome, the Papal States were invaded in January 1798, and the Roman Republic was proclaimed on March 15; Pope Pius VI withdrew to Tuscany. The French occupation weighed heavily on Rome, as it did elsewhere; and there, too, the balance of power swung to the moderates and conservatives, though perhaps the democratic opposition, consolidated around the Constitutional Club, was freer than elsewhere in Italy. From this milieu came the *Pensieri politici* (1798; "Political Meditations") of the southern exile Vincenzio Russo, one of the most important examples of Italian Jacobin thought.

With Napoleon's departure from Italy in November 1797 for his ill-starred Egyptian expedition, the Italian situation changed. In November 1798 the Bourbon king Ferdinand IV of Naples, yielding to English pressure, crossed the papal border and in a swift campaign occupied Rome to reestablish the pope's dominion there. But the counteroffensive was not long in coming, and, while the Bourbon army was dispersing, the French entered Naples on January 23, 1799, although they were held up for three days by popular resistance. The Bourbon court, protected by the English fleet, prudently retired to safety in Sicily. Thus was born the Parthenopean Republic, which, though its authority extended over only some of the southern provinces—the others remaining either in the throes of anarchy or under Bourbon control—was the most democratic of the Italian states set up between 1796 and 1799. Against the Directory's wishes, the military commander Jean-Étienne Championnet and the commissary Marc-Antoine Jullien (a former Babeufist) set up a revolutionary government, and for a few months the intellectual elite of the south enthusiastically participated in the revolutionary experience.

Collapse of the republics. The political situation was rapidly degenerating, however. The Second Coalition against France had been formed in March 1799, and Austro-Russian troops, after occupying the Cisalpine Republic, reached Turin in less than two months; the whole Po plain was thus lost, and the greater part of the French Army abandoned Naples. But the destruction of the Parthenopean Republic was the work of bands of peasants organized by Cardinal Fabrizio Ruffo, a faithful adherent of the King who had landed in Calabria in February; they quickly disposed of the weak democratic militia. Their Armata della Santa Fede (Army of the Holy Faith) was the most important jacquerie (peasant uprising) in the history of modern Italy; invoking God and the King, they devastated the castles of the aristocracy and occupied the communal land that the barons had usurped; they also massacred the bourgeoisie who had set up provisional municipalities. The struggle against the Jacobins and the French was transformed into a great anti-aristocratic movement, which the monarchy skillfully turned to advantage. Naples surrendered on June 23, 1799, and soon afterward the King returned from Sicily; at the behest of the English admiral Horatio Nelson and of Queen Maria Carolina (sister of Marie-Antoinette of France), and after summary trials, the King ordered the execution of more than 100 patriots, to whom the terms of surrender had granted safe-conduct. Thus the best among the southern administrators were destroyed. King Ferdinand's role as a reformer was now a thing of the past.

Nelson and Naples

Between March and July 1799, the French occupied Tuscany and were driven out of it by a violent peasant uprising (the "Viva Maria"), which developed into a march on the cities, where there were massacres of Jews (at Siena) and of citizens who were, or were presumed to be, Jacobins. The rising re-established the power of the rural clergy and the landowning aristocracy.

In September 1799 the Roman Republic finally fell. All Italy was reconquered, and the French resisted only in Genoa, while a stream of Jacobins took refuge in France. The three years of revolution were ended.

The pro-French "patriots" had completely failed to enlist the support of the masses. From the summer of 1796, the rural districts were in ferment, almost always in opposition to the new rulers; there were peasant marches on cities such as Pavia, Bergamo, Brescia, and those of the Romagna and later of Tuscany, while armed bands, especially in the Marches, Tuscany, and the Kingdom of Naples, controlled or reconquered whole regions. Even in some cities, such as Verona, and especially in Naples, the popular dislike of the French and the Jacobins was clearly apparent. The influence of the clergy and the inordinate taxes levied by the republican regimes do not suffice to explain this reactionary alignment, which was, in fact, much deeper seated; only the gradual formation and development of a grass-roots opposition movement would prove capable of weaning the populace from its innate and instinctive conservatism.

Defeated in the internal political struggle, the Italian Jacobins also suffered disillusion with regard to their French ally. Contributions originally levied for military purposes had everywhere degenerated into pure pillage; the constitutions of the new republics were dictated by the French; members of the opposition were imprisoned or driven out of office by coups d'etat; and, finally, Napoleon had adopted an undisguisedly autocratic policy, shown by his reinstatement of the King in Piedmont in the summer of 1796 and by his arbitrary cession of Venetia to Austria in 1797. But their disillusionment with the French, however severe, was unlikely ever to reconcile the Jacobins with the absolute monarchs; rather, it strengthened their nationalism. In Piedmont there was an anti-French, unionist, and democratic organization (the Raggi), and everywhere the need was felt for strong nationalist governments that would lead the country toward unity and independence.

The French Consulate, 1799–1804. Having become master of France by his coup d'etat of 18 Brumaire (November 9, 1799), Napoleon renewed his Italian conquests. Expected by the Austrians to use the Mont Cenis pass, he crossed the Alps by the Great St. Bernard, almost without opposition, and reoccupied Milan on June 2, 1800. A few days later he inflicted a definitive defeat on the enemy at Marengo, between the Bormida and the Po rivers. Defeated also in Germany, the Second Coalition fell to pieces; by the Treaty of Lunéville (February 9, 1801), Austria returned the Cisalpine territory and some portions of Venetia; and the Ligurian Republic was reestablished. Piedmont was reannexed to France in September 1802, together with Elba and Piombino, as also was the duchy of Parma, although official status was not given to this de facto arrangement until 1808. Austrian influence was ended even in Tuscany, where Louis, son of Ferdinand, the Spanish duke of Parma, was enthroned as king of Etruria. In northern Italy, Austria kept only Venetia, whereas France, directly or indirectly, maintained control from the Alps to the Tuscan Maremma, while the restored papal and Bourbon governments further south had little power.

Melzi and the second Cisalpine Republic

The second Cisalpine Republic, established in June 1800, soon proved to be a transitional regime, since it lacked the necessary joint support of the moderates and landowners. Napoleon's most trusted councillor in Paris for Italian affairs was the Milanese patrician Francesco Melzi d'Eril. This statesman, who in 1796–99 had hoped to see upper Italy united in a constitutional monarchy under a Habsburg or a Bourbon ruler, was the most clear-sighted exponent of the views of the old moderate ruling class, still yearning for absolute and enlightened governments. Napoleon favoured the formation of a large Italian state, provided that he could control it. He wanted an Italian republic with a constitution similar to that then operative in France, with the central authority vested in the president and with the representative structure weak and divided among three "estates," the landed proprietors, the merchants and traders, and the learned men and clerics. In such a state he wanted as president either himself or a member of his family. At Melzi's insistence, the new state was established not by a mere edict issued by the French first consul (Napoleon) but by an Italian Constitutional Assembly that met at Lyon, in France, in January 1802. Napoleon appointed Melzi, who was supported by the majority of the deputies, vice president (he himself had become president only after resistance on the fourth vote) and accepted the change of name from the meaningless Cisalpine Republic to Italian Republic.

Melzi pursued a policy of amalgamation. Though the majority of the prefectures and ministries were in the hands of notables, who were often nobles as well, members of the democratic opposition were also gradually included, being given important administrative, cultural, and military posts. The formation of an Italian army was, during the whole Napoleonic period, one of the major concerns of the government, and enduring nationalist sentiments matured among its ranks. Serving as administrators and politicians, the nobles and the educated bourgeoisie for the first time felt an urge to govern and defend their country. Neither the constant French interference and taking of financial levies nor the absence of an Italian foreign policy diminished their enthusiasm for their new political role.

The Napoleonic Empire, 1804–14. When the first consul became emperor, the Italian Republic became a kingdom (proclaimed on March 17, 1805); King Napoleon appointed as viceroy his stepson Eugène de Beauharnais, and Melzi stepped aside. The more docile Antonio Aldini became secretary of state in his place. Italian autonomy was still further limited; but the Napoleonic victories, constantly increasing the territory of the kingdom, provided some compensation. By the Treaty of Pressburg (December 26, 1805), Venetia was annexed, and, with a separate constitution, Dalmatia and Istria were joined to it; in April 1808 the Marches, too, became part of the kingdom, which, by the Treaty of Schönbrunn (October 14, 1809), lost its nominal sovereignty over Dalmatia and Istria; these, together with Trieste, with other territories taken from Austria, and with Ragusa (modern Dubrovnik, Yugoslavia), became the seven French *départements* of the Illyrian provinces. France directly annexed Liguria (June 4, 1805) and Tuscany (in effect from January 1806, formally from March 2, 1809). And with Napoleon's abolition in 1809 of the temporal power of the papacy, Pope Pius VII, who then excommunicated him, was imprisoned, first in France and later at Savona, in northwestern Italy.

As emperor of France or as king of Italy, Napoleon thus directly controlled all upper and middle Italy. During his rule far-reaching reforms were brought about. Though the new codes of law were almost all translated wholesale from the French, without consideration for Italian traditions, they nevertheless introduced at a stroke, particularly in the field of criminal law, a modern jurisprudence notably sensitive to the rights of the individual. Properties held in mortmain, the old feudal ecclesiastical tenure, specifically those of the regular clergy, were transferred to the state and sold, and the remaining feudal rights and jurisdictions were abolished. Road systems were everywhere improved; and both primary and higher education were widely diffused. The increased pressure of taxation was thus compensated for by a network of new and improved services that were to hasten Italian economic and social progress.

The Continental System, a blockade designed to close the whole of the European continent to British trade, proclaimed on November 21, 1806, was freely broken everywhere, including on the Italian coastline; its true meaning was that of favouring French industry, particularly the silk industry. But the war economy stimulated Italian production and led to the development of industries such as the machine industry and metallurgy and to important public works.

In the south, after the repressions and executions of 1799, the Bourbons experimented with some cautious reforms, mainly fiscal and anti-feudal, in order to further strengthen the loyalty of the rural population, which had already proved so valuable. But the Neapolitan government was desperately weak, both militarily and politically; and, between February and March 1806, the French were able to occupy the whole country, while the court once again took refuge in Sicily. On March 30, 1806, Joseph Bonaparte, brother of Napoleon, was proclaimed king of the Two Sicilies; when he became king of Spain in 1808, he was succeeded as king of Naples by one of the most famous French generals, Joachim Murat. Despite this dynastic change, the nine years of French rule in the south can be considered as a whole and represent the most profound and effective reform movement that the country had hitherto experienced.

King Joachim of Naples

King Joachim was more independent of Paris than King Joseph; in his reign not only were there fewer French ministers and councillors in relation to Neapolitan officials, but he also opposed Napoleon over the application of the Continental System. During the 10 years of French rule, feudal privileges and immunities were finally abolished; but even thus mulcted, the landed aristocracy were still able to retain economic supremacy in the country. Extensively buying up the confiscated property of the church and of other proscribed landowners, they thus subverted Joachim's plan to establish small peasant holdings. Much common land originally usurped by large landowners was, however, recovered, and the position of the *galantuomini*, or bourgeoisie, was definitely strengthened. Fiscal, judicial, educational, and administrative reforms were introduced in line with those already made in the Kingdom of Italy.

Meanwhile, both Sardinia, where the court of Savoy took refuge, and Sicily remained apart from the Napoleonic world. In Sicily, the Bourbons were under a strict English control that, originally purely military and naval, soon became political. When, in 1811–12, the court was in conflict with the Sicilian nobles over fiscal matters and arrested the leaders among them, the British commander, Lord William Bentinck, intervened and enforced the adoption of an extremely moderate constitution that left great power to the nobles, though it markedly limited the absolute power of the king. Sicily then experienced a short and intense period of autonomy and political ferment, which was ended in 1816 when the restored Bourbons abolished the constitution and reunited the island to the kingdom.

The Napoleonic regime fell in Italy as it did in the rest of Europe. Eugène, the viceroy of Italy, and Joachim, the king of Naples, with their respective armies had taken part in the fatal Russian campaign of 1812, but, at the moment of defeat, Joachim deserted Napoleon, returned to Naples, and, after making terms with the Austrians, advanced with his Neapolitan troops as far as the Po (March 1814); Eugène, defeated by the Austrians and Neapolitans, was able, by the terms of the Armistice of Schiarino-Rizzino (April 16), to retain Lombardy; but an insurrection that broke out at Milan on April 20 allowed the Austrians to occupy the entire country.

THE RESTORATION PERIOD

The Vienna settlement. At the Congress of Vienna, held by the victorious allies to resettle Europe, it was decided to restore the Bourbons to Naples. It was for this reason that, seizing the opportunity of Napoleon's return to power during the Hundred Days, King Joachim changed sides yet again and, on March 15, 1815, declared war on Austria and, by the Proclamation of Rimini (March 30, 1815), incited the Italians to a nationalist war. Quickly defeated, he was forced to abdicate in May; after taking refuge in Corsica, he landed at Pizzo di Calabria to reconquer the kingdom but was immediately captured by the Bourbons and executed in October 1815.

The Congress of Vienna established the political condition of Italy that lasted until unification. The emperor of Austria, Francis I, became king of Lombardy-Venetia, which was thus incorporated into the Habsburg states, and the former episcopal principality of Trent was directly annexed to Austria. King Victor Emmanuel I of Savoy, in addition to recovering his dominions, acquired the entire Ligurian territory; the duchy of Parma went to the daughter of Francis I and wife of Napoleon, Marie-Louise of Habsburg, but at her death it was to revert to the House of Bourbon-Parma, which meanwhile was granted the duchy of Lucca; the Habsburg-Estes returned to Modena, further acquiring the duchy of Massa by inheritance in 1825; in Tuscany, the House of Lorraine added to its former domains the State of the Presidi and the reversion of Lucca when the Bourbons should return to Parma (this did not take place until 1847); the Pope recovered his temporal dominions in Italy; and the Bourbon Ferdinand IV of Naples, with the new title of King Ferdinand I of the Two Sicilies, reoccupied his possessions.

Restored hegemony of Austria

Thus the Vienna settlement brought about the disappearance of the three aristocratic republics of Venice, Genoa, and Lucca; it strengthened Piedmont and restored the undisputed hegemony of Austria. Austrian troops garrisoned Ferrara, ready to intervene in case of trouble in the Papal States; Austria was given the right to intervene if necessary in the Kingdom of the Two Sicilies; members of the House of Habsburg reigned in Parma, Modena, and Florence; while Venetia and Lombardy became in practice provinces of the Austrian empire. Only Piedmont remained outside the system that Metternich, the Austrian foreign minister and later chancellor, had imposed on Italy; but, under the secret protection of Russia, its government proved to be equally reactionary.

On April 7, 1815, Francis I proclaimed the formation of the Lombardo-Venetian kingdom; but the new state was a fiction because the two regimes remained separated, each directly subject to the central ministries in Vienna. Thus Milan lost its role of capital city; the majority of the Napoleonic bureaucracy was liquidated, and the centralizing authority of Vienna became all-pervasive; many reforms, especially in jurisprudence, were abolished. Discontent proved general, and Austria reacted by increasingly severe police measures and stricter censorship, suppressing, for example, the best liberal and romantic periodical, *Il Conciliatore* ("The Conciliator"), after only a very brief existence (September 1818–October 1819).

Returned from exile in Sardinia, Victor Emmanuel I of Savoy abolished all the laws promulgated by the French and removed from public office all those who had collaborated with them. He invited the Jesuits back into the kingdom and turned many educational institutions over to the clergy. Hence, a liberal opposition was not long in forming among the nobility and the bourgeoisie.

Francis IV of Modena showed an equally conservative intransigence, whereas in Parma Marie-Louise instituted a mild rule and maintained the principal French reforms. Though many of these were soon abolished in Tuscany by Ferdinand III of Lorraine, the enlightened legislation, economic liberalism, and lax policing and censorship characterizing his rule and that of his successor, Leopold II, made the duchy a haven for liberals. It also stimulated cultural activity: at Florence many Italian writers—such as the poet Giacomo Leopardi, the historian Pietro Colletta, and Niccolò Tommaseo—gathered around the Gabinetto (studio), founded by Gian Pietro Vieusseux; Florence was also the birthplace of a famous periodical, *L'Antologia* (1821–33; "The Anthology").

In the Papal States the Restoration, brought about mainly by the diplomacy of the cautious secretary of state, Cardinal Ercole Consalvi, was characterized by increased centralization of government. And, because the public offices were a monopoly of the clergy, the bourgeoisie and the educated classes, who, under the French and Italian regimes, had held some responsibility, became deeply discontented, especially in the Romagna.

The Carbonari

In the Kingdom of Naples the victorious powers had made sure that the Bourbons would not repeat the reprisals of 1799. The Restoration appeared to begin well, under the balanced policy of the minister Luigi de' Medici, who absorbed the greater part of Murat's capable bureaucracy. Many of the French reforms, both administrative and judicial, were retained, but concessions made to the church by the Concordat of 1818 and strict financial economies hampered the advancement of the bourgeoisie. It was especially among these, the *galantuomini* ("gentlemen"), that discontent found an outlet in an imposing conspiratorial organization, the Carbonari. Already founded during the French period, with a vaguely nationalistic program, it now became more widespread, formulating definitely constitutional aims; the southern bourgeoisie were determined to take part in political life and openly to forward their own interests. From the south the lodges of the Carbonari quickly spread throughout Italy, finding their chief centres of support in the Marches, the Romagna, Piedmont, and Milan.

Events in the 1820s. *Effects of the Spanish Revolution.* The Spanish Revolution of 1820 had repercussions in Italy. In the Kingdom of Naples, former members of Murat's army, connected with the Carbonari, marched on the capital (July 2, 1820) to the cry of "Long live liberty and the constitution" and found immediate support among the bourgeoisie and in the army. The King was forced to yield (July 1820) to the liberals' demand for the introduction of the Spanish constitution of 1812; it not only limited the king's power but also decreased centralization and thus reduced the influence of the capital. But the new regime proved short-lived, for it had too many enemies: the King himself, eager to recover his absolute power; Sicily, attempting a separatist revolution (July 15–17), which was violently suppressed by the Neapolitan constitutional government; and, most serious, Austria, which had been given at Vienna the right of intervening to maintain the restored monarchy. In January 1821 Metternich was able to convoke an international congress at Laibach, attended by representatives of the great powers and of the Italian states, and by King Ferdinand himself. Overcoming the weak Anglo-French opposition, the King obtained approval for military intervention. Accordingly, the Austrian Army descended on the kingdom and, defeating the constitutional troops, occupied Naples on March 23, 1821, re-establishing the King's absolute government.

In Piedmont, the more liberal and cultivated wing of the nobility was hostile to Victor Emmanuel I's reactionary position, and the Carbonarian bourgeoisie, with its constitutional hopes, allied itself with them. In the wake of the events at Naples, a conspiracy was set in motion, supported by the Lombard liberals and receiving the covert approval of Charles Albert of Savoy, prince of Carignano, successor-designate to the throne under Salic law. Between March 9 and 13, the revolt, organized by the military and the bourgeoisie, spread from Alessandria to Turin; the King abdicated in favour of his brother Charles Felix and, in the latter's absence, appointed Charles Albert as regent. On March 14 the Regent proclaimed the Spanish constitution, though its adoption was to be contingent upon the new king's approval. But, from his refuge in Modena, Charles Felix refused to accept it, and, with Austrian assistance and the troops that had remained loyal to him, he rapidly reoccupied the country. Three of the conspirators were executed, and the many prison sentences meted out and the rigorous purge of the army provoked a massive emigration. Charles Albert regained the confidence of the new king, but the reconciliation caused a breach between him and the liberals that persisted after he succeeded to the throne in 1831.

Federati conspiracy

In Lombardy-Venetia there was no revolution, but a complex organization of opponents of the regime was discovered and suppressed. The Carbonarian lodge in Milan was attacked in October 1820, and some of the conspirators were deported. In March 1821 the police found the first evidence of the conspiracy of the *federati,* led by the Milanese nobleman Federico Confalonieri, whose program, though more moderate than that of the Carbonari, was no less anti-Austrian and constitutional; from December 1821 to January 1823, members of the conspiracy were discovered even in the army and the upper bureaucracy. Many received death sentences, all eventually commuted to long terms of imprisonment.

Economic slump and revival. The political reactionism (which was prolonged in the Romagna by executions until 1828) was accompanied during this decade by a general economic recession. After the famine of 1816–17, Russian grain flooded the Italian markets, and there was a crisis of agricultural overproduction; the desperate poverty of the peasantry led to widespread pellagra, brigandage, and grain riots. The slump continued unabated until nearly 1830, when successful mulberry cultivation brought renewed rural prosperity and was sufficient, particularly in

Piedmont and Lombardy, to reestablish agricultural credit and provide capital for the development of the textile and some engineering industries.

Renewed prosperity brought leisure for cultural activities, and the economic and social problems of the country began to be discussed in many periodicals; the most important of these was the Milanese *Annali universali di statistica* ("Universal Annals of Statistics"), whose editor-in-chief for some years was the philosopher Gian Domenico Romagnosi and in which his pupil Carlo Cattaneo began to set forth his ideas. In fact, the ranks of the political and cultural opposition, hitherto comprising only the Lombard and Tuscan moderates, would soon include democrats and Catholics.

The rebellions of 1831 and their aftermath. The failure of the uprisings of 1831 proved that the Carbonari were coming to the end of their usefulness. The hopes raised by the revolution in Paris in July 1830 had set on foot a conspiratorial movement that had spread from Modena to the cities of Emilia, chiefly as a result of the efforts of two Carbonari, Enrico Misley and Ciro Menotti. Unfortunately, they relied on the sympathy of Duke Francis IV of Modena, who was willing to countenance changes that might enlarge his small state. But, when he learned that the plot was known to the Austrian police, he had Menotti and 43 other conspirators arrested. Immediately afterward a revolt overthrew the papal government in Bologna and in a few days spread to the duchies of Modena and Parma, to the Romagna, the Marches, and Umbria, leaving only Lazio under papal rule. But for various reasons the new provisional governments of the rebel cities failed to organize a unified military defense; help they had hoped for from the French was not forthcoming, and thus the Austrian Army was able to re-establish the rule of legitimacy during March 1831.

The rise of Mazzini

The moderate liberal leaders, most of them Carbonari, had shown their readiness to treat with the absolute monarchs and had deeply distrusted those republicans and democrats who sought to achieve unification by force of arms. Another component element of the ultimate unification movement was the Adelfi, the group comprising followers of Filippo Buonarroti, the former Babeufist who had taken part in the events of 1796. Ultimately, the task of organizing the democratic and republican opposition conspiracy was undertaken by a young Genoese, Giuseppe Mazzini. Exiled to France at the age of 25 in 1830, he found himself turning away from both the Carboneria and Buonarrottism. In distinction from the Carboneria, his organization, Giovane Italia (Young Italy), was unionist and republican; but, though it put its trust in the education and participation of the people, it had no egalitarian and Jacobin leanings. The new faction spread, especially in upper Italy, with amazing rapidity, absorbing the Buonarrotian and Carbonarian groups. In 1833–34 the first abortive Mazzinian uprisings took place in Savoy and at Genoa, the latter organized by Giuseppe Garibaldi, who then fled to France; in 1834 the Austrian police identified as many as 2,000 members of Giovane Italia in Lombardy. In 1836 Mazzini, who had established firm relationships with revolutionaries in other countries and had joined them in founding the Giovane Europa (Young Europe), left Switzerland and settled in London.

The repressions they had suffered and witnessed, together with the reinforcement of the conservative status quo in Europe, convinced the moderates that it was useless to organize conspiracies with limited membership, that what was needed was to educate public opinion. Meanwhile, the peace forcibly imposed on the peninsula from 1831 to 1848 favoured economic development, notable everywhere except in the south. The south, in fact, remained backward, and the growth of bourgeois property that resulted from the division of the great feudal holdings did nothing to change this. Thus the imbalance between north and south, which would be felt even more acutely after national unification, continued to grow. Genoa and Milan became two of the chief financial centres of Europe; Piedmontese and Lombard industry expanded rapidly; in Venetia important land-reclamation projects were completed, while, in Tuscany, banks and commercial establishments did a flourishing trade, connected especially with the port of Livorno. Throughout the country the construction of a railway network increased commerce and gave rise to subsidiary industries. This economic revival made it more difficult for the governments to tighten police control; at Milan in 1839, Carlo Cattaneo began publishing his periodical *Politecnico,* in which he argued that the progress of science and technology depended upon government reforms. In the same year Pisa saw the first congress of Italian scientists, which was to reconvene annually down to 1847, assuming a more markedly nationalistic character with each passing year.

Liberal Catholicism

Thus conditions gradually became more favourable for the moderates to realize their aims of increasing education and abolishing censorship and police surveillance. In the cause of unification they sought to standardize tolls and trade practices, increase cultural exchanges throughout the country, and, above all, finally to establish representative institutions suitable to Italian traditions and to Catholicism, the religion of the majority of citizens. Liberal Catholicism found its most important expression in Italy in Vincenzo Gioberti's *Del primato morale e civile degli italiani* (1843; "On the Moral and Civil Primacy of the Italians"), in which he affirmed the idea of progress as a return of the existent to the idea, of man to God, realizable only through the mediation of the church. Gioberti envisioned a new and positive role for the temporal power of the papacy, advocating the development of a federated Italy, free from the Austrian hegemony, under the nominal presidency of the pope. His ideas were influential among important sections of the clergy and among Catholics in general. Under different formulations, the new papalist movement struck root and found its best propagandists and theoreticians in Cesare Balbo, Niccolò Tommaseo, and the Jesuit Antonio Rosmini-Serbati.

The renewal of Mazzinian attempts at armed rebellion (among them the celebrated and ill-fated expedition of the Venetian Bandiera brothers, who landed in Calabria in July 1844 and, with seven companions, were executed by a firing squad), all suppressed with bloodshed, increased the esteem felt for the moderates both by governments and the general public. The election of Pope Pius IX in 1846 augured well for the future of the Papal States; his nomination was the result of anti-Austrian feeling, and at the beginning he showed liberal leanings. His first step was the granting of an amnesty to those who had been sentenced for political reasons; this was followed by a gradual removal from governmental posts of the most reactionary prelates, then by permission to publish political periodicals; finally, in April 1847, he instituted a council of state that, though only on a consultative level, gave to the laity a share in public life. Influenced by this liberalism, rulers elsewhere in Italy granted some reforms; one of the most important was the press law of May 1847, by which Grand Duke Leopold II removed restrictions from the press in Tuscany. But the reforms encouraged extremists, and the reactionary powers of Europe became convinced that the peace of Italy was in danger. In July 1847 Metternich sent Austrian troops to occupy papal Ferrara. This intervention stimulated cordiality and cooperation among the Italian rulers, led by Charles Albert of Piedmont, whose relations with Austria were particularly strained. But, while the sovereigns were discussing the formation of an Italian customs union, rendered more urgent by the famine of 1847, the people began to rise.

The revolutions of 1848. On January 9 the first of the revolutions of 1848 broke out in Sicily, in Palermo. Starting as a popular insurrection, it soon acquired overtones of Sicilian separatism and, supported by the nobility and the bourgeoisie, spread throughout the entire island. Individual reforms were no longer enough to content the revolutionaries, who were determined to have new and more liberal constitutions. Ferdinand II of the Two Sicilies was the first to grant one (January 29, 1848), and the other rulers were compelled to follow his example—Leopold II on February 17, Charles Albert on March 4, and Pius IX on March 14. The only Italian rulers who did not yield were the Austrians, who instead reinforced their garrisons in Lombardy-Venetia, arrested the two famous

Charles Albert's declaration of war

leaders of the opposition at Venice, Daniele Manin and Niccolò Tommaseo, and others at Milan, and suppressed the student demonstrations at Padua and Pavia. But on March 22 and 23, when revolution broke out at Budapest and Vienna, Venice and Milan freed themselves by swift and victorious insurrections. In the course of a few days, almost the whole of Lombardy-Venetia was lost to the Austrians, and their army fell back into the Quadrilateral (the land lying between Mantua, Verona, Peschiera, and Legnago). On March 23 Charles Albert declared war on Austria; it was a risky military decision to take, but chances for a nationalist war seemed good, and he had to assume the lead in order to prevent republican domination of the revolutionary movement. He annexed Parma and Modena, which had already driven out their dukes, and won a few other victories. But then the reverses began. Pius IX, Leopold II, and Ferdinand II, who had at first sent troops to support the Piedmontese army, hastened to withdraw them; the Pope's allocation of April 29 to the cardinals showed his unwillingness to further a nationalist movement and did much to discredit the papacy among patriots. Lombardy and Venetia, though not without opposition, accepted annexation to Piedmont, but the Piedmontese regular army was unable to stand up to the Austrian counteroffensive; after a series of lost battles, Charles Albert was finally defeated at the gates of Milan and on August 6 withdrew behind the Ticino River, leaving the Austrians in possession of the city and the duchy, which a popular insurrection had freed only a few months earlier. The accusation of "royal treachery," which the Lombard democrats then formulated, long survived in Italian political polemics.

By the terms of the Armistice of Salasco (August 9, 1848), the Piedmontese army withdrew from Lombardy. But within Piedmont the new constitution was not abrogated, and revolutionary and democratic ideas remained alive.

The forces of reaction triumphed throughout Europe. At Vienna, Prague, Budapest, and Paris, the revolutions of 1848 were stifled; in Naples the King regained power in a coup on May 15, subsequently reconquering Sicily, while at Rome more conservative policies were followed. But Venice, under the dictatorship of Daniele Manin, refused to accept the Armistice of Salasco and resisted the Austrian siege. In Tuscany, Leopold II, finding that the democrats, led by Giuseppe Montanelli and Francesco Domenico Guerrazzi, were gaining over the moderate ministers and aiming at an Italian Constituent Assembly, fled to Gaeta (February 1849); meanwhile, at Florence, a predominantly democratic provisional government was formed. At Rome, the minister Pellegrino Rossi, a former Carbonaro who had returned from France and embarked on a policy of conciliation, was assassinated (November 15, 1848), and the democrats controlled the situation; in consequence Pius IX fled in disguise (November 24), also taking refuge in Gaeta. At Rome the constitutional government convoked a Constituent Assembly with universal suffrage, which, meeting on February 5, proclaimed the republic. The Italian revolution seemed to have been reborn, and the Piedmontese democrats impelled Charles Albert to renew the war with Austria (March 20, 1849). But on March 23 he was routed at Novara and, on the same day, abdicated and went into exile. He was succeeded by Victor Emmanuel II, to whom the Austrian commander conceded an honourable armistice so as not to weaken the monarchical and moderate forces to the advantage of the democrats. The defeat of Piedmont made the position of the democrats and republicans impossible. In Tuscany the moderates called back the Grand Duke, who returned with Austrian troops that crushed a democratic insurrection at Livorno (May 1849); the reconquest of Brescia in March, after 10 days of fighting, left Venice isolated, though it held out against the Austrian Army until August. The Roman Republic, led by Mazzini and Garibaldi, held out until July 3 against the French corps that the new president of the republic, Louis-Napoleon, had dispatched to repay his clerical supporters. The dispossessed sovereigns everywhere returned, abrogated the constitutions, dissolved the parliaments, and, especially at Naples, filled up the prisons.

UNIFICATION

Cavour enters the Cabinet

The role of Piedmont. The exception to this picture of reaction was Piedmont. There King Victor Emmanuel found himself governing with a Parliament whose democratic majority was unwilling to ratify the treaty of peace with Austria, which was indispensable if the defeated country was to be reorganized. By the skillfully worded Proclamation of Moncalieri (November 20, 1849), which contrasted his policy favourably with that of the other Italian rulers, he inaugurated fresh elections, in which the moderate party emerged victorious. The new ministry was headed by Massimo d'Azeglio, a moderate trusted by the King; the most important of his measures was the Siccardi law curtailing ecclesiastical jurisdiction. In October 1850 Camillo di Cavour entered the Cabinet and, from that time, substantially directed financial policy toward free exchange, arranging international commercial treaties and drawing on foreign credit to reduce the public debt and to develop the railway system. This dynamism was displeasing to the conservatives and to the more cautious moderates, including Azeglio himself, who was displaced in 1852 as a result of Count Cavour's alliance (known as the *connubio*) with the Deputies of the left centre. Despite a series of disagreements between Cavour and the King, who was influenced by the clerical party and had some absolutist leanings, various ecclesiastical, fiscal, and judicial reforms were introduced. The prestige of the Piedmontese government both in Italy and internationally was also reinforced by a variety of factors.

In March 1854 France and England had intervened in support of Turkey against Russia in the Crimea; to obtain the support of Austria, they were prepared to guarantee the status quo in Italy, which only Piedmont at that time was in a position to disrupt. But Cavour, anticipating events, concluded an alliance with the Western powers and sent an expeditionary force to the Crimea, where (May 1855) it performed brilliantly. Thus he was able to sit among the victors at the Congress of Paris (February 1856) and there to affirm that the only threat to the peace of Italy and the only pretext for subversive plots was the burdensome Austrian overlordship. It was a tremendous achievement.

Meanwhile, the democratic and republican movement led by Mazzini was losing ground and crumbling. The failure of the Barraba, an attempt in February 1853 by the population of Milan to overpower the Austrian garrison; the discovery and execution at Belfiore (1852–53) of conspirators concerned in a plot centred at Mantua; and other abortive attempts to launch uprisings in Lunigiana and Cadore all contributed to discredit and discourage the democrats. Mazzini's isolation was completed by his known support of the Sapri expedition (June–July 1857), in which Carlo Pisacane, a Neapolitan Socialist with whom he had been in deep ideological disagreement, landed on the coast of Campania with 300 companions; the expedition ended in a massacre.

The democrats, then, were divided and unable to carry on the revolution. There was nothing to be hoped for in the restored governments. In Lombardy-Venetia, Austria had carried out stern measures of repression, while, in Rome, Pius IX, influenced by the secretary of state Cardinal Giacomo Antonelli, refused to grant any reforms. There was now no room left for the liberal Catholicism of the years following 1848. At Naples as in the duchies, reaction became intractable; in Tuscany the Grand Duke tried vainly to make the country forget that he had recovered his throne only by the help of Austria.

Società Nazionale Italiana

The War of 1859. So only in Piedmont was there any hope left for the reformers. There Cavour succeeded in establishing in 1857 a monarchist–unionist party, the Società Nazionale Italiana (Italian National Society), to which the presidency of Manin and the vice presidency of Garibaldi gave a wider appeal than if it had been staffed only by moderates. Though not outlawing conspiratorial movements, Cavour wanted to solve the Italian question by international politics rather than by revolution. At a secret conference held at Plombières, France (July 1858), he arranged with the emperor Napoleon III that the French would intervene in Italy should Piedmont be invaded by Austria; he was clearly planning the complete expulsion of Austria from the peninsula. The price he was to pay for this help was the cession of Nice and Savoy to France and the suppression of the Mazzinian party, erroneously thought by Napoleon to be responsible for the dynamite attack made on him at Paris on January 14, which was, in fact, planned by the Romagno Felice Orsini. The Franco-Piedmontese alliance was finally concluded in January 1859, and, with Napoleon's approval, Victor Emmanuel delivered a speech from the throne in which he declared himself ready to hear the "cry of woe" against Austria that was rising in every part of Italy. Meanwhile, the military party in Vienna was urging the emperor Francis Joseph to declare war. On April 23 an insulting ultimatum demanding that Piedmont demobilize was rejected, and three days later the Austrian Army took the offensive. Thus, as Cavour had hoped, the proviso for French intervention became operative. The allies were victorious in the battles of Magenta, San Martino, and Solferino (June 1859), but, while the routed Austrian Army was in retreat, Napoleon III suddenly signed the Armistice of Villafranca with the Austrians. This sudden change of policy was caused by events in Italy, where unification seemed about to become a reality. At Florence on April 27, Leopold II was overthrown by an insurrection and fled, while the government passed to the moderates, with Baron Bettino Ricasoli as the emerging leader; in June the duchies of Parma and Modena revolted, followed by the Legations (the northern Papal States). The Marches and Umbria also rebelled but were quickly reconquered by papal troops. The liberated provinces declared their wish to be united to Piedmont, but France then, no more than at any time in the past, did not want a united Italy. At Villafranca it was arranged that Napoleon III should receive Lombardy from Austria and cede it to Piedmont; the sovereigns of Modena and Tuscany were to be restored, and plans were made to establish a federation among the rulers of the Italian states. This was a serious defeat for Cavour's policy, and he resigned in July 1859, the King replacing him in the government by Urbano Rattazzi. But England was particularly opposed to the forceful restoration of the rulers of Emilia and Tuscany, and even Napoleon III, who meanwhile had increased his prestige in France by the annexation of Nice and Savoy, was unenthusiastic; in this climate of international opinion, Cavour's policy soon returned to favour, and he resumed office on January 21, 1860. A series of plebiscites were then held in Tuscany and Emilia, all of which declared for union with Piedmont. The fear of a democratic revolution, the need to weaken Austria, and Great Britain's desire to set up a strong Italian state to balance France all induced the European powers to assist the Piedmontese monarchy in obtaining this great success.

Garibaldi and the Thousand. But the democratic party refused to admit that the national revolution was in any way complete, when so many parts of Italy remained under their old sovereigns. The most suitable place for a democratic revival was Sicily, where autonomous opposition to the Bourbon government was endemic and extremist. In April 1860 a popular insurrection broke out in Palermo (the *rivolta della Gancia*), which, though it was quickly quelled, spread through the cities and the countryside under the unmistakable influence of Mazzinian agents. It was then, at the beginning of May 1860, that the democrats showed that they were able to overcome the deep differences that had divided them during the previous decade. The Expedition of the Thousand, which, with the tacit approval of Cavour, set sail from Quarto, near Genoa, under the command of Garibaldi, had been principally recruited among the bourgeoisie of Lombardy and Venetia but also contained volunteers from all the old states and represented the most divergent forces.

Despite the scant preparation, the expedition, which was almost entirely without arms, after disembarking at Marsala on May 11 conquered nearly the whole of Sicily in less than three months. The factors that made this possible were the revolutionary ferment already existing there and Garibaldi's military skill. But the attitude of the Sicilian peasants proved ambivalent; at first welcoming the invaders, they later became disappointed by the failure to partition the feudal estates, and they even fought the Garibaldians. Though on May 14 he had proclaimed himself "Dictator in the name of Victor Emmanuel, king of Italy," Garibaldi had set up a provisional Sicilian government, actually directed by his associate Francesco Crispi, which came into serious conflict with Cavour's emissaries to the island; Cavour was afraid of a turn toward republicanism. Meanwhile, the European powers attempted to mediate, and the new king of the Two Sicilies, Francis II, granted a constitution and promised autonomy and pardon to the Sicilian rebels. But without the consent—and apparently even against the wishes—of Victor Emmanuel, Garibaldi crossed to Calabria on August 19, 1860, and on September 7 made a triumphant entry into Naples, which the King had abandoned, fleeing to Gaeta. On October 1 the last serious Bourbon resistance was overcome at the Battle of the Volturno, near Caserta. But the prestige of Garibaldi and of the democrats had grown too great, and it was time Cavour resumed the initiative. Persuading Napoleon III to make only a formal protest, he occupied the Marches and Umbria (September 1860); so as not to offend Napoleon's clericalism, it was agreed that Rome and Lazio should remain under the pope, while the rest of Italy was to become a moderate constitutional kingdom. On October 26, 1860, Victor Emmanuel, having entered Neapolitan territory with his army, met Garibaldi, who hailed him as "king of Italy"; during October and November, the formerly papal and Bourbon provinces voted by plebiscites to be annexed to the kingdom of Italy. The kingdom's inauguration was formally proclaimed on March 17, 1861, by a Parliament meeting in Turin, and soon afterward (March 25 and 27) Cavour affirmed the necessity that, with the approval of France and with the famous formula "a free church in a free state," Rome should become the Italian capital; but, when he died, on June 6, 1861, the "Roman question," with many others, remained unsolved.

Garibaldi enters Naples

The kingdom proclaimed

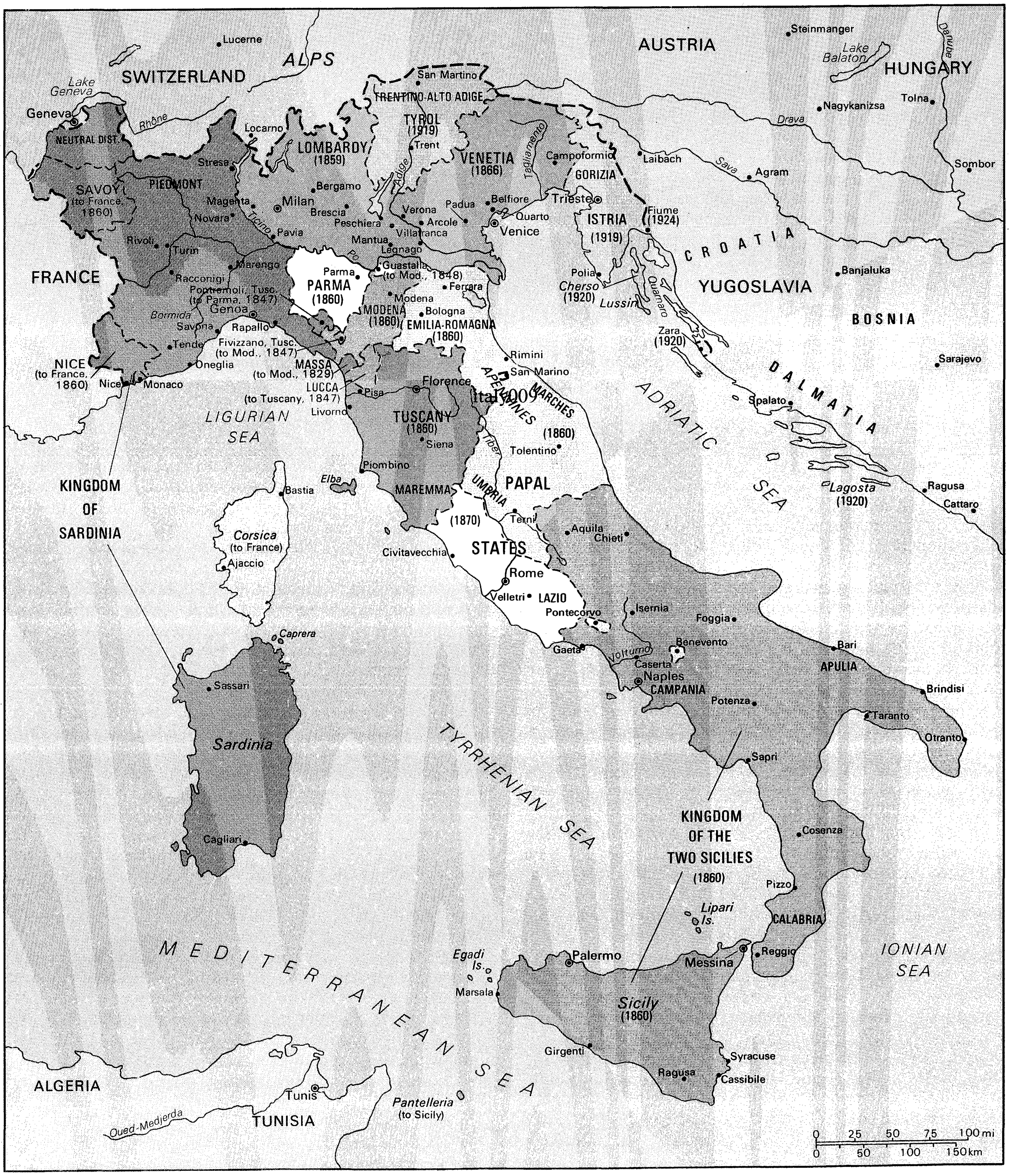
Condition of the Italian kingdom. In 1861 the kingdom had 26,000,000 inhabitants, 78 percent of whom were illiterate, while 70 percent of its active population were engaged in agriculture. Thus it seemed unlikely that Italy could make the economic progress shown by other European countries in that period. The group that had gained the majority in Parliament in 1861 was the moderate-conservative right, based principally on an alliance between the Piedmontese group, headed by Giovanni Lanza and Quintino Sella, which controlled industries and banks, and a Tuscan group, led by Bettino Ricasoli, which was interested in commerce and transportation. This political class wanted a centralized governmental structure that would allow the Parliament and hence the executive power to control the local administrations, especially where democratic forces or autonomistic aspirations might otherwise become preponderant. By a series of laws in 1865, they effected legislative unification and established firm central control over the provinces and their communes through the appointment of regional prefects. The democratic opposition, entirely preoccupied with the problems of freeing Rome and Venetia, made little resistance to this authoritarian construction of the state.

Centralization certainly provided no remedy for the serious economic imbalance between north and south. The free-trade policy of the right-wing politicians ruined the weak industries of the south, especially the woolen industry in the Salerno district, which had hitherto been protected. Moreover, the south had few railways, which were built under contracts that suggested government corruption; and the systems of poor relief and education were, and remained, miserably inadequate. Naples, which in 1861 had 447,000 inhabitants and was the most populous city in Italy (Turin came next, with only 205,000), was afflicted with pauperism and viewed with jealousy by the smaller cities of the south. Yet the most intense wretchedness was in the rural districts, where the peasants had totally failed to acquire any proportion of the expropriated estates. Consequently, many of them took to an especially violent form of brigandage, which, though it was organized by former Bourbon officials and even by Bourbon emissaries loyal to the exiled Francis II, was primarily a peasant war directed especially against the agrarian bourgeoisie. The movement was harshly suppressed by troops, and at least 5,000 peasants were executed; but it was not brought to an end until 1865. Public opinion, however, remained largely dominated by the political problem of completing the country's unification. The democrats wanted above all to solve the Roman question, and, when Ricasoli's ministry (June 1861–March 1862) was succeeded by that of Urbano Rattazzi, a Piedmontese lawyer apparently more liberal than Ricasoli, it seemed that the time had come. In July and August, Garibaldi, despite government prohibition, raised armed bands in Sicily and began moving up the peninsula again to march on Rome; but Rattazzi, preoccupied with the attitude of France and Austria, and realizing that French troops were garrisoning Rome, had the army disperse the Garibaldian troops rounded up at Aspromonte, in Calabria, where (August 29, 1862) Garibaldi was wounded and put under arrest for two months. The scandal that followed brought about the fall of the government. An attempt at a partial solution of the Roman question was made by the ministry (March 1863–September 1864) of Marco Minghetti. By the Convention of September (signed at Paris on September 15, 1864), Napoleon III agreed to gradually withdraw French troops from papal territory in the course of the next two years; in return, Italy undertook to respect papal rule and—by a secret clause—to transfer the capital from Turin to Florence. When this condition became public, there was an uprising in Turin, with 30 dead (September 21–22), and Minghetti was forced to resign.

The Roman question

The adherence of Venetia and Rome. Two years later, attention was diverted from Rome to Venice by the outbreak of war (June 1866) between Austria and Prussia. Italy, governed by the ministry of the Piedmontese Alfonso La Marmora, took the opportunity to attack the Austrian-held lands in Italy but was defeated on land at Custoza on June 24 and at sea off Lissa on July 20. Only

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision Copyright © 1964 by Barnes & Noble, Inc.



The unification of Italy. The dates are those of annexation, first to the Kingdom of Sardinia and, after 1861, to the Kingdom of Italy.

the Garibaldian volunteers in Trentino had some success. By the Treaty of Vienna (October 3, 1866), Italy, through the mediation of Napoleon III, obtained the cession of Venetia. After a short-lived Minghetti ministry, Rattazzi returned to power, giving tacit consent to the stationing of Garibaldian bands along the papal boundary; but Rattazzi resigned, and meanwhile Garibaldi went into action and was defeated by French troops at Mentana (November 3); upon re-entering Italian territory, he was arrested and sent to the island of Caprera, between Corsica and Sardinia, where he had property.

Both diplomatically and militarily Italy had suffered a marked loss of prestige. Nor was the internal situation happy. There were separatist revolts in Palermo (1866) and others around Parma (1869) because of the tax on milling grain; furthermore, financial restrictions, necessary in order to effect reorganizations and to support the army, made the government unpopular.

The ministry of Giovanni Lanza and Quintino Sella, formed in December 1869, was perhaps the most typical among those of the right wing—it repressed the Mazzinians, advocated free trade, and was prudent in foreign affairs, with a pro-French bias. But, despite its lack of brilliance and its subservience to France—shown when it almost yielded to King Victor Emmanuel's desire to intervene in the Franco-Prussian war of 1870—it achieved the solution of the Roman question. This was made possible because the defeat and abdication of Napoleon III ended French protection of the papacy. On September 20, 1870, after a symbolic resistance by the papal army, Italian troops entered Rome, opening a breach in the wall of Porta Pia. The Pope, refusing to accept the situation, forthwith withdrew inside the Vatican.

With the taking of Rome, the geographical unification of Italy was completed. But an equally important popular unification remained totally unachieved. A vast gulf continued to divide the small proportion, perhaps 2 percent, of the population that, on the basis of a property suffrage, enjoyed electoral rights and was represented in moderate and anticlerical chambers and governments from the mainly illiterate masses, who, preponderantly peasant and Catholic, were beginning to be stirred by working class and Socialist doctrines. (Ma.B.)

DEVELOPMENTS FROM 1870 TO 1914

Once unification was achieved with the capture of Rome (even though Trent and Trieste, the *terre irredente,* or "unredeemed areas," still under Austrian rule, remained outside the boundaries), and once the Roman question was provisionally, albeit unilaterally, settled with the enactment of the Law of Guarantees (May 13, 1871), which guaranteed the pope full ecclesiastical freedom, the Giovanni Lanza–Quintino Sella government focussed its attention on domestic problems. The most crucial of these was the need to improve the financial situation, and this was accomplished in 1876 by holding back public expenditure and increasing revenue by taxation.

Financial problems

Minghetti's last ministry. When Lanza fell in June 1873, Marco Minghetti, who succeeded him in July, attacked the problem of the inflationary paper currency circulation that grew substantially after the introduction of forced currency under the Act of April 20, 1874, which stipulated the volume of currency to be put into circulation for about 20 years; this act set up a consortium of six major issuing private banks (with the result that no state central bank was established). But the Minghetti Cabinet failed to secure adoption of the fiscal measures it had submitted to the Chamber, which was therefore dissolved (September 20, 1874). The ensuing elections resulted in a heavy gain for the left. Next to the traditional, moderately progressive Piedmontese left, headed by Agostino Depretis, and the more consistently progressive left of the other north central regions (both representing the middle class groups of the north), the southern left, the champion of the interests of the landed bourgeoisie of the south and more moderate than the traditional left, showed a remarkable gain.

In spite of these differences, the left, after its success at the polls in 1874, presented a more solid parliamentary front than did the right: indeed, the disagreement within the right on the question of "saving" the railroads by making their operation a state responsibility was responsible in March 1876 for the downfall of the last government of the traditional right, brought about by the Tuscan rightist group's aligning itself with the left against state operation of the railroads. The right, the "moderate" heirs of Cavour, who had built up the unitary state, unified the national market, and set the country's financial house in order, was now succeeded by the left, a party whose leadership consisted mainly of men of the Risorgimento democracy who, unlike the more strictly orthodox Mazzinian republicans, had accepted the monarchist–liberal solution. More resolutely secular and anticlerical than the right, the left had a broader social base extending to the upper levels of the urban working classes, with whose support it was prepared to widen the narrow bases of the unitary state: its program's chief elements were extension of the franchise, compulsory elementary education, tax reform, abolition of the forced currency, and appointment of mayors by election.

Disagreement within the parliamentary right

Depretis and the parliamentary left, 1876–87. In the decade after 1876, Italian public life was dominated by Depretis, the leader of the left, who was president of the Council almost uninterruptedly from March 25, 1876, until his death on July 29, 1887, and who headed eight administrations in which he reserved almost invariably the internal-affairs portfolio and often the foreign-affairs portfolio for himself. A skilled parliamentarian, a talented administrator, and a realistic and flexible statesman, Depretis proceeded with caution in carrying out the program of the left, partly because of the internal divergences within the left itself, which comprised—from left to right—the extreme group of Agostino Bertani, still open to republican promptings; Francesco Crispi's group, clinging to the democratic traditions of the Risorgimento; the Lombard progressive left of Giuseppe Zanardelli and Benedetto Cairoli; and Giovanni Nicotera's group, predominantly southern, bestowing political patronage and more conservative. Nonetheless, on July 15, 1877, the Coppino Act was passed, making the first two years of elementary schooling compulsory, a provision that, though inadequate, helped to lower the level of illiteracy. In 1882 a reform increased the electorate to 6.9 percent of the total population, including a substantial number of workers in the north. The forced currency was abolished between 1881 and 1883. On January 1, 1884, the unpopular grist tax was finally repealed. The Railroad Agreements Act (April 27, 1885) vested responsibility for operating the railroads in private companies and strengthened the hand of the southerners. Public expenditure, however, rose noticeably from 1881 because of increased allocations for the armed forces, and the deficit created was aggravated by the minister Agostino Magliani's "exuberant" financial policy.

Finally, the free-trade tradition was abandoned, and a protectionist measure in favour of industry was enacted in 1878; in 1887 a general tariff was introduced to provide much greater protection for some branches of industry. Because of the widespread farming crisis of the 1880s, the import duty on grains was raised substantially, jeopardizing the already depressed living standard of the lower income groups. The tariffs of 1887 strengthened a political block within the mainly southern group of absentee owners of large estates and those entrepreneurial groups of the upper middle class and of the bourgeois-minded nobility who had directed the Risorgimento process.

"Transformism"

One basic feature of Depretis' politico-parliamentary system much criticized as a source of political corruption and degradation was "transformism," whereby moderate or conservative deputies moved over to the benches of the leftist majority because they had had more than their fill of the wrangling over the ideals and programs of the rightist era and because priority was given to economic and administrative matters, on which agreement in Parliament was now made easier.

Not even under the personal influence of the new king, Umberto I, who succeeded Victor Emmanuel II in 1878, did the left change the foreign-policy aims pursued by the right from 1870 to 1876, namely, to terminate the alliance with France and to seek closer relationships with Germany and Austria-Hungary, as promoted by Marquis Emilio Visconti-Venosta. The policy of cautious reflection required for the consolidation of the young Italian state was, however, offset by the isolation of Italy, which became obvious during the eastern crisis of 1877–78 and at the Congress of Berlin and especially at the tension over the French occupation of Tunisia in 1881.

The Triple Alliance

This isolation ended in 1882, when Italy concluded the Triple Alliance with Germany and Austria-Hungary. The treaty, which had a stabilizing effect on domestic policy, provided, among other things, that Germany, Austria-Hungary, and Italy would undertake to support each other if ever attacked by other powers and that the other two contracting parties would remain neutral if the third was forced into declaring war on another power. The Triple Alliance, from which Italy derived no great advantages, was nevertheless renewed in 1887 on rather more favourable terms that, in addition to giving Italy anti-French assurances that the status quo in the Mediterranean would be maintained, allowed it in substance to request that Trentino and Trieste be ceded to it in the event of any encroachment by Austria in the Balkans.

Colonial expansionism became more marked as imperialistic interests developed. The government acquired Aseb (Assab) in 1882, previously in the possession of the Rubbatino shipping company from 1869 to 1870, and Italian forces occupied Mesewa (1885), a bridgehead for subsequent penetration into Eritrea that was interrupted by the defeat at Dogali (January 26, 1887).

Crispi, to 1891. *"Strong" foreign policy.* Depretis was succeeded by Francesco Crispi, a former Mazzinian who had accepted the constitutional monarchy and whose initial democratic fervour had been steadily waning. His coming to power (August 1887) ushered in a new phase of Italian policy characterized by motives of prestige and colonial expansionist trends at the international level and by a "strong" line at home. An admirer of Bismarck and a fervent nationalist, Crispi adopted an intransigent attitude toward France both in the matter of trade relations and in dealing with the problems created by spheres of influence in the Mediterranean, while forging closer links with the Central Powers, especially Germany.

Expansion in East Africa

Crispi then emphasized the imperialistic aspects of colonial expansion in East Africa. The outcome was the protectorate established over the Somali sultanates of Obbia and Migiuritinia (1889) and later extended to the coast of Benadir; the occupation of the interior of Eritrea (1889), recognized by Menelik II, emperor of Ethiopia, in the Treaty of Ucciali (May 2, 1889); and the attempt to impose an Italian protectorate on Ethiopia.

Domestic policy. In domestic policy, Crispi, the guiding spirit of the industrial–agricultural block consolidated in 1887, encouraged the trends toward authoritarianism that emerged on several occasions from 1861 onward. The first phase of Crispi's decade was also characterized by legislative activity that, while seeking to strengthen the executive by enhancing the authority of the president of the council, began to broaden the basis of public life and to provide a more honest and efficient administration: the extended franchise in local government elections; the elective appointment of mayors by commune councils; reform of the system of administrative law; and promulgation of a new criminal code, the Zanardelli code, which abolished the death penalty and granted some freedom to strike. In addition, Crispi's minister of finance (1889–90), Giovanni Giolitti, reduced the deficit by strict economy.

The Rudinì government and Giolitti's first administration. In February 1891 Crispi resigned, and, after a brief administration (February 1891–May 1892) headed by Marquis Antonio Starabba di Rudinì, who renewed the Triple Alliance for 12 years, Giolitti became prime minister on May 25, 1892. Giolitti, a Piedmontese who had played no part in the Risorgimento, followed a more progressive line.

The years 1870–90 witnessed the emergence and early stages of two movements, one Socialist and the other Catholic, that were radically to alter the nature of the political struggle. The first organizational framework of incipient Italian Socialism on a nationwide scale was the Italian Federation of the International Workingmen's Association (Rimini Conference, August 4, 1872), whose constitution—partly inspired by Mikhail Bakunin, the Russian anarchist—was made possible by the radicalization of the younger republican generation, which had sympathized with the Paris Commune of 1871. The First International, whose orientation was anti-authoritarian, anarchistic, and collectivist in Italy, spread chiefly to Naples, Sicily, and central Italy, recruiting its supporters from among the urban workers and young intellectuals. But it failed to enlist support among the peasantry, in spite of the rural insurrectionist movement that the internationalists started in 1877 and that they thought should have been supported by the peasants of the south. The futility of the insurrectionist method brought the International, after 1877, to a crisis aggravated by the abandonment of their positions by some of its leaders, including Andrea Costa, founder of the Revolutionary Socialist Party of Romagna in 1881 (as from 1884 the Italian Revolutionary Socialist Party). Moreover, in 1882, the Italian Labour Party (Partito Operaio Italiano) came into existence in Milan and rallied Lombardy workers to a labour program based on trade union opposition. The same years witnessed completion of the transition from radical democracy to Socialism of a group of Lombardy intellectuals, among them Filippo Turati and Leonida Bissolati, who assimilated Marxist ideas; another intellectual, Antonio Labriola, advocated the establishment in Italy of a Marxist Socialist culture free of Positivist contamination from which, in his opinion, the Lombardian Socialist group was not immune.

Birth of organized Italian Socialism

Among the Catholics, the "intransigent" element, defenders of the rights of the papacy against the "usurping" Italian state and advocates of political abstentionism, prevailed over the clerico-moderates, who favoured an accommodation with the state. The "intransigents" acquired an effective organizational instrument in the *Opera dei congressi* (1875), promoting—especially from 1885 onward—a range of economic and social activities, mainly in north central rural areas.

With these developments in mind, Giolitti tried to introduce a financial policy that would lighten the burden on the lower income classes but also would permit more freedom in political and trade union organization. It is not by pure chance that, during Giolitti's first administration, the Italian Socialist Party (PSI) was founded (August 1892) as an entity separate from the Anarchists. The continuation of this policy was nevertheless impeded by Giolitti's resignation (November 24, 1893) because of the Banca Romana scandal, in which the financial aspects were compounded with political corruption.

The Banca Romana scandal

Crispi, 1893–96. Giolitti was again succeeded by Crispi at a time when the economic and bank crisis had become more acute and in Sicily the Fasci dei Lavoratori movement (unions of labour, especially of peasant groups, linked with the PSI) had become more extensive. In order to deal with the tense situation, Crispi made use of the law enforcement authorities and exercised emergency powers: disbandment of the Fasci and declaration of a state of siege in Sicily and Lunigiana, where an uprising fomented by the Anarchists had been attempted (January 1894); adoption of "anti-Anarchist" laws; and disbandment of the PSI (October 1894).

Crispi later succeeded in establishing equilibrium in the credit sector by reorganizing the banks, whereby the Banca d'Italia came out on top as the leading issuing bank; and he strengthened his parliamentary position through the elections of May 1895, held after a revision of the electoral rolls had reduced the number of voters to the detriment of leftist opposition parties.

The discontent engendered by Crispi's policy, particularly noticeable in the north, expressed itself over the African crisis. The annexation of Tigre as part of Eritrea (October 1895) actually forced the emperor Menelik II, who had already denounced the Treaty of Ucciali in 1893, into war. The campaign ended in an Italian defeat (Aduwa), forcing Crispi to resign on March 5, 1896. Thus there came to a close, with a heavy deficit, an era of Italian foreign policy characterized by an ill-prepared imperialistic drive.

Rudinì, 1896–98. Rudinì, who succeeded Crispi for the

Treaty of Addis Ababa

second time (May 1896), settled the African problem by the Treaty of Addis Ababa (October 1896), which recognized the independence of Ethiopia and left Italy in possession of Eritrea up to the Mareb line and the Somali Protectorate. His domestic policy, however, was vague, wavering between the pressures of the conservative right—which sought the replacement of the parliamentary type of government by the "constitutional" type, with ministers responsible only to the sovereign—and the pressures of the radical left. Rudinì began by adopting a tolerant attitude (amnesty for political prisoners) and introduced some social welfare measures, but a Cabinet reshuffle (July 1896) shifted the balance to the right, thrusting the radicals into opposition.

Finally, Rudinì's inadequacy became apparent in the crisis of 1898, when, as a result of the rise in grain prices, heightened tension convulsed the whole country and culminated in mass demonstrations in Milan. He called the army into the Lombard capital, then in a state of siege, and resorted to severe repression.

When Rudinì resigned in view of the Chamber's hostility (June 18, 1898), Luigi Pelloux, who succeeded him, girded himself to continue repression; and, when—after he had prepared a bill highly detrimental to constitutional freedoms—the extreme left resorted to parliamentary obstruction, he promulgated the repressive laws by royal decree and adjourned the Chamber. This authoritarian gambit backfired: the elections of June 1900 were a triumph for the extreme left, forcing Pelloux to resign.

Agriculture and industry, 1870–90. From 1870 to 1900, Italy remained a predominantly farming country whose agriculture was characterized by major discrepancies and striking contrasts. In the Po Valley of the north, with the help of development projects, the large estates, with their capitalistic owner-managers, continued to gain ground, as did a class of capitalist entrepreneurs ("tenant farmers") employing hordes of farmhands. In the central regions (Tuscany, the Marches, Umbria, and part of Emilia), the predominant system was sharecropping. The south was both the realm of the large estates and also of peasant holdings often too small to offer their owners a bare livelihood and therefore compelling them to work as labourers or tenants on the large estates.

Crisis in agriculture

The unification of the market and the construction of the railroad network spurred production, especially of specialized crops intended for export (olive oil, wine, citrus fruits), but, except for the capitalistic estates of the north, the increase in output was generally from the steady expansion of the area under cultivation rather than from the investment of capital. Agriculture suffered severely from the effects of the crisis of the 1880s. The general drop in prices and output had been of particular significance for grains. The state's endeavour to cope with the situation by taxing grains (1887), while it enabled the big southern farms to survive, was detrimental to the small- and medium-scale wine-growing peasants of the south, who found themselves—because of the trade war with France—excluded from one of their principal markets.

In 1870 Italy did not yet possess an industrial base of significant proportions, even though, in the northern regions, there were signs of a developing manufacturing industry, especially in the textile sector (silk, cotton, and wool). The industrial "takeoff" was held back not so much by the low level of capital formation as by the smallness of the domestic market and by the tendency to invest in speculative rather than directly productive ventures. This being so, state intervention, particularly the policy of protectionism (tariffs of 1878 and 1887), enabled steady progress to be made in the textile sector, now assured of the domestic market, while the establishment of Terni and other complexes launched the iron and steel industry in the direction of modern production.

The three decades under discussion witnessed an initial phase of expansion, although it was slowed down by recurrent crises, as a result of which the gross product of industry at the end of the century still amounted to only around 20 percent of the overall gross national product. Lastly, per capita income remained essentially at a standstill, and the standard of living remained low.

Accession of Victor Emmanuel III

Saracco, 1900–01, and Zanardelli, 1901–03. Giuseppe Saracco, who succeeded Pelloux (June 1900–February 1901), was a "caretaker" minister who withdrew the drafts of illiberal laws and launched a policy of détente that was not even interrupted by the assassination of Umberto I (July 29, 1900) by an Anarchist, because the new king, Victor Emmanuel III, showed his desire to guarantee constitutional freedoms. On Saracco's resignation, a government was formed by Giuseppe Zanardelli, whose minister of the interior, Giolitti, inspired his policy. The Zanardelli–Giolitti administration (February 1901–October 1903) marked a turning point in Italian life, ushering in a period of renunciation of conservative authoritarianism and of liberalization of domestic policy.

The Giolitti era. When he became president of the Council in November 1903, Giolitti dominated the scene until the outbreak of World War I, heading three administrations (November 1903–March 1905; May 1906–December 1909; March 1911–March 1914) apart from a few intervening caretaker cabinets (Alessandro Fortis, March 1905–February 1906; Sidney Sonnino, February–May 1906 and December 1909–March 1910; Luigi Luzzatti, March 1910-March 1911).

Social reform and the growth of organized labour. Giolitti sought to bring the radical and Socialist left into the constitutional picture and to introduce a policy of reform to meet the demands of the upper strata of the working classes. He abandoned the posture of repressing the labour movement, recognized the freedom to organize and to strike, and remained largely neutral in labour disputes. A series of accompanying legislative innovations, while of limited scope, helped to better the lot of the working classes (restraints on the employment of child and female labour; 24 consecutive hours of rest per week in industry; measures to benefit workers in rice fields, etc.). Lastly, the Credaro Act (1911) made state bodies responsible for operating elementary schools.

Giolitti's policy encouraged the rapid growth of the organized labour movement. Hence, in 1901 there was a noteworthy expansion of the trade federations and the trade-union councils (town-based territorial organs), which became increasingly socialistic and which also engaged in educational and assistance activities. Concurrently there was a gradual spread of the peasants' associations in regions with big capitalistic farms (Po Valley and part of Apulia). In this way masses of farm labourers were organized and, in 1900–01, set off a wild wave of strikes designed to raise low wages and improve harsh working conditions. The farm labourers' organization found its rallying point in the constitution of the Federazione Nazionale dei Lavoratori della Terra (National Federation of Agricultural Workers), or Federterra (157,000 members in 1910). Federterra, however, mainly reflected the aspirations of the labourers (more farmhands and fewer peasants) and made "socialization" of the land the central theme of its program. This limited its penetration into the regions of small peasant holdings and sharecroppers, where the peasants' main desire was to own their piece of land. National unification of the Socialist-oriented trade unions occurred in 1906 with the establishment of the Confederazione Generale del Lavoro (CGL), or General Confederation of Labour. The CGL soon became an efficient organ (384,000 registered members in 1911), but its cautious reformism benefitted the more advanced labour categories, concentrated in the north, and neglected the strata of less skilled and unorganized workers, chiefly in the south.

Catholic trade unions

Even the Catholics, who had entered the field of trade union organization after the publication of *Rerum Novarum* (1891), stepped up their activity, especially among the peasants of the north central region and the textile workers' branch, where women predominated (104,000 registered members in the Catholic, or "white," trade unions in 1910).

Domestic policy. Giolitti succeeded in moving farther ahead with his domestic policy also, thanks to the support of the reformist wings of the PSI (Turati, Claudio Treves, Ivanoe Bonomi, Leonida Bissolati), which predominated in the parliamentary group and the CGL and often lent Giolitti support in Parliament but refused to join his administrations. But the "Socialist revolutionary" movement, which had no specific ideology, and the "revolutionary trade unionism" splinter group, which was influenced by the ideas of the French revolutionary syndicalist Georges Sorel and which quit the PSI in 1907, were fiercely hostile to Giolitti.

During the Giolitti period, Parliament and government were active in the economic and financial sectors, even in relation to the economic expansion of those years. When the railroad agreements of 1885 expired in 1905, responsibility for running the railroads was taken over by the state; in June 1906 legislation was enacted to convert the public debt from 5 to 3.75 percent (and later to 3.5 percent), an operation that was successful because of the sound state of the national financial situation, which showed a surplus up to 1909–10; in 1912 it was the turn of the state life insurance monopoly, and, at the same time, a great impetus was given to public works.

Against this, one unworthy aspect of "Giolittism" was the use made by Giolitti of political patronage to maintain his power, which he did with an unscrupulousness that reached its climax during the elections.

Foreign policy. In foreign policy Giolitti tried to find Italy an alternative to the Triple Alliance and subordination to Germany. In addition to improved relations with England, the rapprochement with France, started by Rudinì, who had settled the Tunisian question in 1896 and ended the trade war with France (by means of a trade treaty of November 21, 1897), was encouraged. This led to the Italo-French agreements of 1902, which delimited the spheres of influence of the two countries in North Africa and pledged them to mutual neutrality in the event of aggression by third parties. As a corollary, relations with Austria grew worse because of the friction created by recurrent irredentism, Italy's commercial penetration of the Balkans, and Italy's interest in Albania, as did relations with Germany, especially after the Moroccan crisis of 1905, during which Italy maintained an impartial position between France and Germany. The tension with Austria became acute in 1908, when Austria's occupation of Bosnia without compensation for Italy placed the Triple Alliance in jeopardy.

The conquest of Libya

Giolitti strengthened the Italian position in the Mediterranean with the conquest of Libya (Italo-Turkish War of 1911–12), which was opposed by the nationalists, who were starting to grow aggressive and were supported by sectors of the Catholic world. But the Libyan war had important domestic repercussions. The conflict upset the delicate internal balance of the PSI and caused it to break away from the reformist movement of the right wing (Bissolati, Bonomi, Angelo Cabrini) in favour of the war, while the reformists of the left, with Turati, remained true to the anti-colonialist tradition of Italian Socialism. Moreover, there emerged a new left, nurtured on idealistic volunteerism, whose most typical representative was Benito Mussolini, the creator of an eclectic platform founded on criticism of parliamentarianism, on antimilitarism, and on the advocacy of revolutionary violence.

The situation having changed thus, Giolitti could no longer count on the PSI for support and therefore turned to the Catholics. The latter agreed to give their support in the 1913 elections—the first with virtually universal suffrage—to the governmental candidates (the Gentiloni Pact). The liberal–Catholic alliance, concluded as an anti-Socialist measure, did not, however, prevent the government's relative defeat, the Catholics, PSI Socialists, and the Reformist Socialist Party of Bonomi and Bissolati all gaining seats at the expense of the liberal groups.

The economy. The turn of the century and the Giolitti period were years of growth for the Italian economy. Farm production actually rose, thanks to internal and export demand, to mounting investments in capitalistically structured estates, and to the completion of land-improvement projects in Emilia-Romagna. But, above all, the industrialization process gathered speed, with a growth rate of industrial output of 6.7 percent per annum for the period 1896–1908. Both the spread of "mixed" banking and the increase of capital stock in industry contributed to this growth. In addition to the textile industry, the electric power (Edison) and the metallurgical industries became increasingly important. And per capita income in 1910–14 was 28.8 percent higher than in 1896–1900.

Industrial concentration in the north

The industrial concerns, however, were concentrated in the northwestern part of the country, which in 1911, with 27 percent of the total population, accounted for about 58 percent of the total industrial workers. Industrialization was practically nonexistent in the rural and backward south, aggravating the so-called southern problem. Emigration, mainly from the southern regions, started as a large-scale exodus in 1880 and grew to an annual average of 600,000 emigrants from the start of the 20th century to 1914, mostly headed for the Americas.

WORLD WAR I AND THE RISE OF FASCISM

The Salandra government. Giolitti, who resigned in March 1914 because of the radicals' switchover to the opposition, was succeeded by Antonio Salandra, heading a government that was expected to leave the way open again for Giolitti. But the outbreak of World War I drastically altered the nature of the political struggle. Salandra, after weathering a series of storms culminating in Red Week (serious popular uprisings that occurred in the Marches and the Romagna in June), coolly faced up to the war crisis. He ruled that the provisions of the Triple Alliance did not apply to Austria's attack on Serbia and therefore decided on neutrality (August 2, 1914). In the following months, however, Salandra considered intervening on the side of the Allies, particularly after the disappointing outcome of the attempt to bargain for the *terre irredente* with Austria. Meanwhile, the neutralists in Italy consisted of Giolitti supporters, Socialists, and Catholics. The interventionists included republicans, reformists of the Bissolati–Bonomi school, some splinter groups of liberal democrats, who regarded the conflict as the last war of independence and the final achievement of national unity, and expansionist and imperialistic-minded elements; these were later joined by revolutionary-Socialist and anarcho–trade-unionist groups and by Mussolini, therefore expelled from the PSI.

The Treaty of London

When negotiations with Austria broke down, Salandra sought contacts with the Allies that finally led to the Treaty of London (April 26, 1915), which remained secret; in return for intervention on the Allies' side, Italy was to receive Trentino, Trieste, Alto Adige (South Tirol) up to the Brenner Pass, Gorizia and Istria up to the Quarnaro, and northern Dalmatia. The Prime Minister, denouncing the Triple Alliance on May 4, was, however, forced to resign (May 13) by the neutralist majority in the Chamber. The interventionists stepped up their pressure with a well-orchestrated propaganda campaign and aggressive street demonstrations. The King refused the Prime Minister's resignation (May 16th) and took the decision to intervene, a decision that, while not formally exceeding his constitutional authority, nevertheless ran counter to the feelings of the majority in the country. The Salandra government thus took over full powers and, on May 24, declared war on Austria.

Entrance into the war and the peace settlement. Salandra resigned in June 1916 after Austrian success in Trentino, and a national coalition government was formed under Paolo Boselli, who declared war on Germany (August 28, 1916). His government respected constitutional freedom and tolerated the pacific-neutral PSI, even after violent antiwar demonstrations in Turin in August 1917. Boselli was toppled after the Battle of Caporetto disaster (October 1917), and the new government was formed in November by Vittorio Emanuele Orlando, who, at one of the most critical junctures in the history of united Italy, led the country on to victory.

The "mutilated victory"

Having got out of the war, Italy was shaken by a crisis in which the political, economic, and social elements were intermeshed. Nationalists and imperialists, dissatisfied with the peace conference at which the Allies, while granting Trent, Trieste, and the Brenner frontier, did not recognize Italy's claims to Fiume (modern Rijeka, Yugoslavia) and Dalmatia and its expansionist aims in Albania and Asia Minor, began to bandy about the myth of the "mutilated victory." The middle classes felt the effects of inflation acutely and were terrified by the atmosphere of social tension; the industrialists and farmers were anxious about the turmoil stirred up by the workers, which, on a scale hitherto unknown, was assuming the political character of a struggle to take over the state; the peasant masses pressed that the prospect of "land for the peasants;" glibly conjured up in time of need by the leading groups, should be realized.

Nitti, 1919–20. Orlando, depressed by the course the Paris negotiations were taking, resigned on June 19, 1919; and Francesco Saverio Nitti, a democrat and an expert on the problems of the south, formed a government of the left and Giolitti supporters. Nitti did indeed promote the passage of progressive social legislation (compulsory unemployment, sickness, and old-age insurance) and tried, unsuccessfully, to introduce the eight-hour working day. But the increase in social strife obliged him to call often on the police to quell the turmoil.

In the elections of November 16, 1919, the first held on the basis of universal suffrage and proportional representation, the PSI scored a triumph (30 percent of the votes and 156 seats in the Chamber). The results were likewise favourable for the Italian Popular Party (PPI), the Catholic but formally nondenominational party founded in January 1919 by Don Luigi Sturzo and campaigning for improvement of peasant small holdings and for administrative decentralization (20.5 percent of the votes and 100 seats). But the elections disappointed the liberals, democrats, and rightists. Thus it became difficult to form stable governments, which would henceforth require the support either of the rigidly intransigent Socialists or the Populists.

As regards foreign policy, Nitti renounced the claims to Dalmatia and to the colonies, thus incurring the accusation of "capitulator" from the nationalist groups of the right. Against this background the question of Fiume became more embittered. Fiume, claimed by both Italy and Yugoslavia, was occupied in a piratical *coup de main* led by Gabriele D'Annunzio (September 12, 1919).

Fiume seized

Giolitti's last ministry: growth of Fascism. Internal tension and the difficult parliamentary situation led to Nitti's resignation (June 9, 1920). The 78-year-old Giolitti now served his last term as prime minister, supported by the liberal democrats and, conditionally, by the Populists. Giolitti concluded with Yugoslavia the Treaty of Rapallo (November 12, 1920), which made Fiume a free state and gave Italy some Dalmatian islands.

Giolitti then tried to deal with the domestic situation by his old method of staying neutral in labour disputes and using the police to maintain law and order. In this way he weathered the crisis of the September 1920 "occupation of the factories" by workers in Turin and Milan. He cut the budgetary deficit by increasing tax revenue (*e.g.,* special capital levy, heavier death duties, and taxes on higher income brackets) and reducing expenditure.

Still, Giolitti's experiment turned out to be not enough to control the critical events in which the country was embroiled. He could not count on the PSI's support. The PSI's "maximalist" left (victorious in the party congress of October 1919) harped on revolution and the dictatorship of the proletariat, although its verbal extremism was not matched by capacity to guide the masses. Moreover, the anti-electoral Communist splinter group of Amadeo Bordiga and that of Antonio Gramsci's "New Order" organized themselves on the left of the maximalists. These splinter groups, having broken away from the PSI in January 1921, formed the Italian Communist Party (PCI).

Birth of the Italian Communist Party and of Fascism

Meanwhile, Fascism was developing menacingly. Launched in Milan on March 23, 1919, by Mussolini, the Fascist movement did not at once take a precise ideological stance and combined pragmatist, revolutionary–trade-unionist, nationalist, imperialist, and irrationalist elements in its platform; its program laid stress on direct opposition to Socialism and "Bolshevism" and on collaboration between management and labour in production to further the higher interests of the "nation." Fascism availed itself flexibly of the elements of disillusionment and discontent prevailing in the country: nationalistic hysteria intensified by the "mutilated victory"; the operational difficulties of the parliamentary state; the effects of the economic situation; management's anxiety about the growth of trade unionism (CGL's registered membership rose from 249,000 at the end of 1918 to 2,320,000 at the end of 1920); and the middle classes' fear of Socialism.

After a difficult beginning (in the 1919 elections the Fascist vote was insignificant), Fascism gathered strength after D'Annunzio's seizure of Fiume (supported by Mussolini) and especially after the failure of the occupation of the factories, the turning point of Socialist fortunes. Allying itself with the traditionally more reactionary groups of the Italian ruling classes, Fascism became a kind of armed reaction against the working classes. Deploying "punitive expeditions" by paramilitary "squads," Fascist strategy succeeded, from the late 1920s onward, in throwing the Socialist and trade union movement into disorder, first in the Po Valley and then in the northern cities, not sparing even the Catholic organizations.

The victory of Fascism. The victory of Fascism, a predominantly urban movement with a broad lower middle class base, reached rural areas in 1921–22; it was encouraged by the indecision of the Socialists, who were torn with strife—in October 1922 a further schism saw the expulsion of the reformists, who formed the Unitary Socialist Party (PSU). But the movement was cold-shouldered by Giolitti and other liberal democrats, who believed that they could "constitutionalize" Fascism and absorb it within the liberal state. When the Chamber was dissolved, Giolitti announced new elections for May 15, 1921, and formed a governmental National Bloc comprising, besides the liberals, democrats, nationalists, and Fascist candidates. Nevertheless, out of 535 seats, the left gained 123 seats, plus 15 to the Communists, and the Populists obtained 108 seats, but, within the National Bloc (275 seats), the Fascists succeeded in securing 35.

Having changed from a movement into a party and dropping republicanism at the Congress of Rome (November 1921, 300,000 registered members), Fascism, under the weak governments that followed Giolitti (Bonomi, July 1921–February 1922; Luigi Facta, February–October 1922), set off to conquer power. When the last attempt to block his way, the general strike proclaimed on July 31, 1922, misfired, Mussolini carried out the March on Rome (October 28, 1922), aided by the surrender of the King, who refused to sanction a state of siege and entrusted Mussolini with a mandate to form a new government, which took office on October 31.

The March on Rome

The kingdom and Fascism

BEFORE WORLD WAR II

First years of Fascist government. Fascism did not work immediately for a complete breakdown of the Italian political and constitutional system; and Mussolini's first government included liberal and populist ministers. But the establishment of a totalitarian regime, with power vested in the Fascist Party and, through it, in its leader, or *duce,* was the goal. Accordingly, as one act of violence against its opponents followed another, Fascism institutionalized its own armed force by setting up the Voluntary Militia for National Security (January 1923) and transformed the electoral system in its own favour (the Acerbo Act, which made the kingdom a single national constituency and assigned an absolute majority—51 percent—of the seats in the Chamber to the party holding the relative majority).

At the elections of April 6, 1924, held in a suffocating atmosphere, 64 percent of the votes went to the "national" government lists (374 deputies, 275 of them officially Fascists), except in northern Italy, where more votes went to the opposition.

Assassination of Matteotti

The assassination by Fascists of the PSU deputy, Giacomo Matteotti (June 10, 1924), who had denounced in the Chamber the violence of the electoral campaign, created tension. Fascism, morally isolated, went through a phase of disbandment, while the opposition groups (Aventinians, named after the ancient Roman *plebs'* secession from the Roman assembly on Aventine hill—signifying withdrawal from the political scene) stopped their parliamentary work in protest. But the indecision of the Aventinians and the attitude of the King, who did not revoke Mussolini's mandate, allowed Fascism to climb back again. After having assumed full responsibility for what had happened (speech in the Chamber, January 3, 1925), Mussolini abandoned collaboration with supporting groups and set up a total dictatorship.

Between 1925 and 1926, Italy was thus transformed into a police state: the activity of Parliament, the parties, and trade unions was increasingly curtailed, and freedom of the press and association became a sham. An act of December 24, 1925, made the president of the Council "head of the government" with no responsibility to Parliament, removable only by the sovereign, and the sole person competent to lay down the agenda of the two chambers. Further provisions ended any autonomy in local government. Representatives of the Confederazione Generale dell'Industria and the Confederazione delle Corporazioni (*i.e.,* the Fascist-oriented trade unions) signed an agreement in October 1925 asserting that they alone represented the industrialists and the workers, respectively. A few months later other laws (April 3 and July, 1926) confirmed this trade union monopoly and empowered the Fascist unions to draw up collective labour contracts; in addition, strikes and lockouts were prohibited, and the authority to settle collective-bargaining disputes was vested in the labour judiciary. Lastly, provision was made for setting up national bodies to maintain liaison among the trade union organizations of the various factors of production by branches of production; they were called corporations and regarded as the cornerstone of the corporate state, which was to be founded on cooperation of the classes and the integrated organization of production to further the development of "national power."

Provision for the Fascist corporation

The dictatorship completed. An attempt on his life (October 31, 1926) gave Mussolini the opportunity to abolish all remaining freedoms. Provisions were introduced to disband all the parties, end the parliamentary mandate of 120 Aventinians, restore the death penalty for some political prisoners and introduce special police measures ("forced residence" and "admonishment"), and set up the Special Tribunal for the Defense of the State. A new criminal code and code of criminal procedure (limiting rights of defense and abolishing the jury) and a new Public Security Consolidation Act took effect in 1931.

Mussolini continued to build up a totalitarian state, in which everything was subordinated to the personal will of the *duce.* A new electoral law (March 16, 1928) finally subordinated the Chamber, with the introduction of a single list of candidates selected by the Fascist Grand Council and submitted as a whole for approval by plebiscite. The Grand Council, set up as a party organ early in 1923, was "constitutionalized" (December 1928) and became a state organ in the service of Mussolini.

The regime then embarked on a campaign to make the country Fascist, using radio, press, and school, dragooning the young into paramilitary organizations, and making party membership obligatory for admission to government service, including the judiciary. The corporate structure, however, remained a facade. The corporations, numbering 22, were not established until 1934, although the Ministry of Corporations had been set up as early as July 1926. Even though granted wide powers of overall economic control, the corporations confined themselves to advisory activities and had little impact on economic and social life.

Lastly, the constitutional reform of January 1939, replacing the Chamber of Deputies by the Chamber of Fasces and Corporations, whose membership was automatic by virtue of membership of other Fascist organs, was of virtually no consequence.

The Lateran Treaty

In the matter of state–church relations, Fascism concluded the Lateran Treaty with the Holy See (February 11, 1929). These pacts, which enhanced Mussolini's prestige and enabled him to use the church hierarchy in certain cases to support the regime, repealed the Law of Guarantees of 1871, created the State of Vatican City, and affirmed the catholicity of the Italian state, recognizing religious marriage as valid in civil law and introducing religious teaching in intermediate-level schools.

As from 1926, the opposition to Fascism decided to act under cover or to emigrate. The Communists chose the first alternative and focussed their program on internal affairs: they fomented strikes and disturbances and infiltrated the very mass organizations of Fascism. The price they paid was high (the party leader, Gramsci, arrested in 1926, died in penal servitude in 1937; and over 4,000 of the approximately 4,700 people sentenced by the Special Tribunal were Communists); but the uninterrupted existence of the PCI created the conditions for its growth as a major party after the fall of Fascism. The same alternative was chosen by "Justice and Liberty," an intellectual liberal-Socialist movement founded by Gaetano Salvemini, Carlo Rosselli, Emilio Lussu, Ernesto Rossi, and others in 1929 and destined to amalgamate with the Action Party (PdA) in 1949. Catholic opposition was expressed within traditional organizations, such as Catholic Action, a training ground for many cadres of the future Christian Democratic Party; liberal opposition was exemplified by Benedetto Croce and the group associated with him in the periodical *La Critica.*

Socialists in exile

The two Socialist parties (PSI and PSU, reunited in 1930) and the Republican party (PRI) were reconstituted in exile; in 1927 they organized the Anti-Fascist Action Coalition. With the disbandment of the coalition (1934), while the Socialists in Italy became reorganized in Rodolfo Morandi's Internal Centre, Socialists and Communists in exile (under the "popular front" policy) signed a united action agreement in August 1934.

Fascist foreign policy. In foreign policy, after an initial phase of moderation (Italian–Yugoslav Agreement of Rome, January 27, 1924, which gave Italy Fiume; good relations with Great Britain; accession to the Treaty of Locarno, October 1925), Mussolini pursued an expansionist policy, based on revision of the peace treaties and the Italian presence in the Mediterranean and Danube–Balkan areas. The often inconsistent dynamism of Fascist foreign policy grew more intense after 1930; thus there was the Four Power Pact of June 7, 1933, between Italy, Germany, France, and Great Britain on the revision of treaties, which Mussolini hoped would allow him to lay the groundwork for Italian hegemony in central Europe but which was rendered meaningless by the French attitude of suspicion; the Austrian-Hungarian agreement of March 1934, designed to protect Austria's independence from Hitler's threats; and the Stresa Conference (April 1935), at which the Italians, French, and British adopted a position in favour of Austria's territorial integrity and against the rearming of Nazi Germany.

Soon after Stresa, Mussolini, erroneously believing that these agreements had given him a free hand, turned his attention to colonial expansion, namely to Ethiopia, which he attacked on October 2, 1935. The war ended victoriously for Italy, and, in May 1936, Victor Emmanuel III donned the ephemeral crown of the Empire of Ethiopia.

The "Rome–Berlin Axis"

Meanwhile, the ties between Fascist Italy and Nazi Germany led to the Berlin agreements of October 23, 1938—the "Rome–Berlin Axis," which provided for a joint military effort to support Gen. Francisco Franco in the Spanish Civil War. Directly after the occupation of Albania (April 1939), the Italian–German alliance was finally consolidated by the Pact of Steel (May 22, 1939), which was of both a defensive and offensive nature.

The economy under Fascism. *Economic advance in the mid-1920s.* As regards the economy, Fascist policy followed a liberal line favouring private enterprise up to 1925. Traditional customs protectionism, embodied in the tariff of 1921, had been reduced by trade agreements; but previous governments' progressive measures, such as state life insurance and compulsory registration of share certificates, were abolished. The year 1925 ushered in a period of state intervention, with measures to reduce the balance-of-payments deficit, including the attempt to obtain from domestic sources the grain required to meet national needs regardless of cost; higher customs tariffs; and the vesting of power in the minister of finance to prohibit certain imports and to set quotas. But, since the lira continued to drop in value, the government was obliged in 1926 to declare the Banca d'Italia the only issuing bank and to stabilize the currency by revaluation (92.46 lire to the pound sterling, the so-called "90 quota"), which was, however, too high and harmful to exports. Thus began a period of deflation, with a drop in prices damaging to agriculture, wage and salary cuts, and a rise in unemployment.

Nevertheless, in spite of the negative effects of the "90 quota," the Italian economy followed a rising trend up to 1929. The national income climbed from 95,000,000,000 lire (at 1938 prices) in 1921 to 124,600,000,000 in 1929, thanks to the manufacturing boom (from 54 to 90 between 1921 and 1929, base year 1938 = 100). In particular, the electric power, chemical, and metallurgical industries made headway, with streamlining that strengthened monopolistic complexes such as Edison, Snia Viscosa, Fiat, Montecatini, and Pirelli.

The Institute for Industrial Reconstruction

Effect of the world depression. The world depression of 1929 had serious repercussions on industrial output and on the banking system. As a result, there was a further increase in state intervention, namely, the establishment in 1933 of a state agency, the Institute for Industrial Reconstruction (IRI), to rescue some major banks encumbered by locked-up capital, which meant that a substantial part of the credit system was immobilized. Moreover, by increasing the share participations of the rescued banks, IRI—and hence the state—became a shareholder in a complex of companies that, in 1939, represented more than 44 percent of Italy's share capital and acquired a controlling interest in a group of them equivalent to 18 percent of the total capital. Thus there emerged a striking feature of the modern Italian economy; namely, the coexistence of frequently interlinked private and public sectors.

The measures introduced as a result of the 1929 depression, especially those of 1935, made it easier to launch the policy of "autarky" designed to make the Italian economy self-sufficient. In addition to the impetus given to grain production—which expanded, however, at the expense of specialized crops—there was exploration for minerals, expansion of hydroelectric plants, reorganization of the iron and steel industry by the establishment of a holding company of IRI (Finsider), and promotion of the armaments industries.

Industry revived, but the recovery was moderate, so that on the eve of World War II Italy was still primarily an agricultural country. As before, industrial enterprises, especially the major ones, continued at the close of the 1930s to be concentrated in the north, while the south remained an agricultural and depressed area with a population of poor peasants and labourers.

ITALY IN WORLD WAR II

Mussolini's fall from power. On the outbreak of World War II in 1939, Italy first adopted a position of nonbelligerence, entering on Germany's side (June 10, 1940) only when German successes in France misled Mussolini into believing that Germany's victory was a certainty. But soon the unfavourable course of the military operations, revealing the country's lack of preparedness and the adventurousness of Mussolini's policy, widened the gap between the majority of the country, sorely pressed by the war, and the regime—a gap illustrated by strikes in many factories of the north in March 1943.

The landing of the Allies in Sicily (July 10, 1943) was the death knell of Fascism. Mussolini, in the minority at a meeting of the Grand Council (July 25) at which it was proposed to deprive him of authority and to restore power to the crown, was placed under arrest by the King, belatedly anxious to dissociate his responsibilities from those of the *Duce,* while the unsuccessful reaction of the Fascist Party and the wave of anti-Fascist demonstrations following Mussolini's fall brought about the regime's collapse.

Armistice with the Allies

The Badoglio government. The government of Marshal Pietro Badoglio, appointed by the King and consisting of military men and technicians, negotiated with the Allies. On September 3 the armistice of Cassibilie was signed, providing for the Italian armed forces' unconditional surrender and the establishment of an Allied administration. The announcement of the armistice, made by Badoglio over the radio on September 8, caused the disbandment or transfer to the Germans of almost all the Italian units in the peninsula, France, and the Balkans; and the country, thrown into confusion, came to be divided in two by the front line. In the south, controlled by the Allies, who were fighting the Wehrmacht up the peninsula, the Badoglio government declared war on Germany (October 13, 1943), thus securing the status of "co-belligerent" for Italy; in the north, under German occupation, after Mussolini's rescue by German parachutists (September 12, 1943) and the reconstitution of the Fascist Party (now called the Republican Fascist Party), the Italian Social Republic came into existence (September 17) as an ally of Germany.

The Committee of National Liberation and the Resistance

Emergence of anti-Fascist parties: the liberation of Italy. Meanwhile, emerging anti-Fascist parties were active. Among those with the largest following were the Christian Democrats (DC), a Christian-based party formed of various Catholic-oriented groups; the Socialist Party of Proletarian Unity (PSIUP), the result of the merger of the PSI and the Proletarian Unity Movement (founded by Lelio Basso in January 1943); and the PCI. These three parties, together with the smaller parties (Liberal, Action, Republican, and Labour Democracy), set up at Rome on September 9, 1943, the Committee of National Liberation (CLN), which called on Italians to fight against Nazism–Fascism. Thus the Resistance began. The armed struggle grew with the emergence of large partisan formations that went to make up, as from June 1944, the military command of the Corps of Volunteers for Liberty. The partisans' activities, which continually harassed the Nazi–Fascist troops at great sacrifice of life (about 36,000 killed in Resistance ranks and about 10,000 civilians killed in reprisals), were supported by the strikes of March 1944 and the spring of 1945 in northern Italy.

The anti-Fascist parties in the south, who had refused representation in the government as being prejudiced in favour of the abdication of King Victor Emmanuel III and his son Umberto, changed their attitude after the U.S.S.R. recognized the Badoglio government (March 1944) and after the concurrent about-face of the PCI, whose secretary, Palmiro Togliatti, on March 31 proposed the formation of a provisional government of national unity. Thus the CLN parties were admitted to the Badoglio administration (which came to power on April 22), and royal powers were transferred to Prince Umberto. When Rome was liberated (June 4–5, 1944), Umberto was proclaimed lieutenant general, and a coalition government of the CLN parties headed by Bonomi was formed, but under virtual Allied control. A Constituent Assembly, once territorial liberation was completed, was to be convened to solve the constitutional question, and the government entered a crisis (December 1944) because of disputes over purging Fascist elements from the administration (which the left wished to have carried out more decisively) and over the powers of the CLN (which the moderates wished to be limited). A second edition of the Bonomi administration was characterized by a move to the right.

The "northern wind"

Parri's coalition government, June–December 1945. After the liberation (to which the general partisan uprising of April 1945 contributed), the long negotiations to form a government that would fulfill the desire for political and social renewal, expressed with much greater forcefulness in the north than in the south (the so-called "northern wind"), led to the formation of a coalition government of the CLN parties (June 1945) headed by Ferrucio Parri (of the PdA) as a compromise between the opposing candidacies of the Socialist Pietro Nenni and the Christian Democrat Alcide De Gasperi. In the Parri administration, profound differences soon ended the unity achieved during the Resistance and set the pattern of political deployment that characterized the following years. On the one side stood the PSIUP and the PCI, whose mass support came from the industrial and labourer proletariat of the north and the sharecroppers of the centre. The extreme left wished to set up an advanced popular democracy and to transform the social structure through antimonopolist measures (nationalization, worker-based management councils, etc.) and the limination of ownership of large estates through land reform. The left was opposed by the Christian Democrats (backed by the liberals), an interclass and moderate party worried about the "Communist" danger, an anxiety shared by the rightist Uomo Qualunque (UQ) movement with its Fascist nostalgia.

The Parri administration tackled the problem of Sicilian separatism and reopened the question of purging, which Parri tried to extend to private industry, but his leftist leanings led, in November 1945, to the withdrawal from the Cabinet of the liberal ministers, soon to be followed by the Christian Democrats, and to the government's subsequent downfall. This was the beginning—with a coalition Cabinet of the CLN parties, except for PdA—of the long series of administrations of De Gasperi, who turned out to be the ablest Italian political leader since Cavour. De Gasperi, supported by the Allies, soon formulated Italian policy along moderate lines.

Abdication of Victor Emmanuel III

After the local government elections of March–April 1946 had strengthened the majority parties of the DC, PSIUP, and PCI, and after the abdication (May 9–10) of Victor Emmanuel III in favour of his less compromised son Umberto (Umberto II), there were held simultaneously on June 2—with universal suffrage including women—a referendum on the constitution and elections for a Constituent Assembly whose powers were limited to the drafting of a constitution and the ratification of treaties. In the constitutional referendum, 12,717,923 votes (54 percent) were cast for a republic; in the elections to the Constituent Assembly, held on the basis of proportional representation, the Christian Democrats (DC) scored a clear victory (8,101,004, or 35.2 percent), followed by the PSIUP (4,758,129, or 20.7 percent) and the PCI (4,356,686, or 18.9 percent). The liberals, on the other hand, saw their ranks sorely depleted as compared with the pre-Fascist period (6.8 percent), while the UQ vote (5.3 percent) was greater than expected, above those of the PRI, monarchists, and PdA.

The republic

THE DE GASPERI ERA, 1945–53

In the new De Gasperi administration, set up with the participation of the DC, PSIUP, PCI, and PRI, the differences between the DC and the leftists became stronger because De Gasperi took conservative steps to strengthen the powers of the state in the maintenance of law and order. In January 1947, moreover, the conflicting trends within the PSIUP led to the breaking away of the rightist wing (Giuseppe Saragat, Ludovico D'Aragona, and others), which favoured the Western democracies and was "autonomist" vis-à-vis the PCI, while the leftist elements (Nenni, Morandi, Basso) favoured close cooperation with the Communists. The seceders thus founded the Socialist Party of Italian Workers (PSLI), while the PSIUP changed its name to the PSI, which was joined in October 1947 by the majority of the dissolved PdA.

The peace treaty

Foreign policy aligned with the West. Meanwhile, after De Gasperi's visit to the United States in January 1947, which, in the climate of the Cold War, implied a choice of side, the Italian government signed the peace treaty on February 10, 1947. The treaty, which seemed to the public to be extremely severe, not only limited the armed forces and laid down the scale of reparations but also provided for some frontier adjustments with France (Tende-Brigue, Mont Cenis, etc.), the surrender of the Dodecanese Islands—occupied during the Italian–Turkish War—to Greece, and renunciation of the colonies (Eritrea eventually went to Ethiopia, Libya became independent, and a ten-year trusteeship over Somalia was assigned to Italy). Trieste was temporarily constituted a Free Territory subdivided into two zones, under Anglo-American military and Yugoslav military administration, respectively. As for the Alto Adige, the De Gasperi–Grüber agreement of September 5, 1946, whereby Austria recognized the Brenner frontier and Italy undertook to grant the region a broad measure of self-government, was included in the treaty.

Politics at home. After signature of the peace treaty, the "imposed coexistence" of the DC and the leftist groups in the administration became more difficult. De Gasperi resigned on May 12, 1947, and formed a one-party Christian Democratic (plus a few independents) government. Strengthened by the success of anti-inflationary measures and by the failure of leftist disturbances to bring down the government, as well as by economic aid from the United States (Marshall Plan and OECD), De Gasperi broadened his administration in December by including PSLI and PRI representatives.

The republican constitution

The Constituent Assembly had completed its work by approving on December 22, 1947, the text of the new constitution, which entered into force on January 1, 1948. The republican constitution, the result of compromise (the case of article 7 is typical: it provided for the inclusion of the Lateran Treaty in the new constitution), embodied the innovative ideas of the Resistance period, laid down the "right to work" as a basic human right, and repudiated war. The separation of powers was upheld: the executive (entrusted to the president of the republic, elected for a term of seven years by the two branches of the legislature in joint session), the legislative (assigned to the Chamber of Deputies and the Senate, both elected on the basis of universal suffrage), and the judicial; and locally autonomous regions were established, although their coming into operation was long delayed.

The elections of April 18, 1948, were characterized by awareness that a decisive choice was involved. The result was an unexpectedly overwhelming victory of the DC (48.5 percent of the votes, with an absolute majority of the seats in the Chamber, 304 out of 574), which secured the support of the clergy and of a larger cross section of the conservative electorate. The Communists and Socialists, united in the Popular Democratic Front, did not manage, with 31 percent, to improve on their positions of 1946. The DC's gains, however, were mainly at the expense of the right: liberals and UQ, standing as the National Bloc (3.8 percent), and the Italian Social Movement (MSI), a neo-Fascist group (2 percent). De Gasperi, who, with an absolute majority in the Chamber, could have formed a one-party government, preferred instead to build a broader base, initiating the period of "centrism" (fifth De Gasperi administration, with the DC, PSLI, PRI, and PLI [the Partito Liberale Italiano], May 1948–January 1950; sixth De Gasperi administration, tripartite without the PLI, January 1950–July 1951; seventh De Gasperi administration, the DC and PRI, July 1951–July 1953). During the years of centrism, Italy entered the Western bloc, acceding to the North Atlantic Treaty Organization and joining the Council of Europe and the European Coal and Steel Community.

Domestic policy and the economy. In domestic policy, there were developments in trade unionism: the unity achieved in June 1944, when the Italian General Confederation of Labour (CGIL) was reconstituted, came to an end. The Catholic trade unions broke away from the CGIL (a member of the Soviet-oriented World Federation of Trade Unions) and organized the Free General Italian Confederation of Labour (LCGIL). In June 1949 the Social Democrat and Republican trade unions constituted the Italian Federation of Labour (FIL). In 1950 the LCGIL and the FIL merged into the Italian Confederation of Workers' Trade Unions (CISL), while in the same year two new central trade union organs came into being, the Italian Union of Labour (UIL), a coalition of some of the Social Democrat and Republican trade unions, and the neo-Fascist Italian Confederation of National Workers' Trade Unions (CISNAL).

Industrial progress

In the same years, after the prewar level of production had been restored in 1949, the industrial sector made rapid progress, assisted by the abundance of cheap labour. This dynamism led to the establishment of a competitive iron and steel industry, to the beginning of the radical transformation of the chemical sector with the discovery of large methane deposits in the Po Valley and the entry of the Edison company on the scene, and to the founding of the National Hydrocarbons Authority (ENI), a public corporation set up in 1953; the automobile industry also began to expand. Even though there was large-scale state intervention in such branches as iron and steel and chemicals, management of the economy remained largely in the hands of private enterprise. Nevertheless, Italy remained an agricultural country as regards employment, more than 8,000,000 people being employed in farming as against some 6,000,000 in industry in 1953.

Under the pressure of peasant unrest (squatting in 1949–50), the centrist administrations sought to introduce reforms into the rural areas, particularly in the south. Thus a special act (May 1950) for the Sila, a zone in Calabria, and the so-called *stralcio* law ("provisional order") on land reform (October 1950) were passed, providing for the expropriation of about 1,700,000 acres (700,000 hectares) for distribution among designated families (approximately 270,000 acres [110,000 hectares] at the end of 1960). But these measures did not go far enough. In an effort to narrow the widening gap between the north and south, the Southern Fund was set up in 1951 and helped to build infrastructures but did not succeed in stimulating widespread industrialization. In spite of the economic recovery, Italy continued to suffer from mass unemployment, which led once again to emigration.

Shortly before the election of 1953, the DC induced Parliament to pass—amid bitter attacks of the opposition parties—an act that would have given 65 percent of the seats in the Chamber to the party, or group of allied parties, polling 50.01 percent of the votes. But the results of the June 7, 1953, elections prevented this "legislative swindle" from going through. The four "allied" centre parties obtained only 49.85 percent of the votes, showing a general drop as compared with 1948, particularly for the DC and the PSDI (Italian Social Democratic Party, the name adopted from January 1952 by the PSLI). (F.d.Pe.)

SUCCESSORS OF DE GASPERI

Years of instability. *Government.* The second republican administration saw six successive Christian Democrat one-party administrations between 1953 and 1958. The limited capacity of these governments to manoeuvre severely curtailed their ability to pass important legislation. Thus laws for instituting the regions (except for those with special statutes), for establishing the Constitutional Court, and for replacing the Fascist codes were postponed. On the credit side, the governments between 1953 and 1958 could claim only the Ten-Year Plan for Growth and Development (Vanoni Plan), which, however, did not become operational; the strengthening of ENI, which was given exclusive rights to explore for oil and methane in the Po Valley; and the largely unsuccessful effort to stimulate the industrialization of the south. As for foreign policy, the Trieste question was settled (October 1954), administration of the two zones being assigned to Italy and Yugoslavia, respectively.

Settlement of the Trieste question

The government's relative lack of mobility, however, was offset by movement within the parties. In the DC, Amintore Fanfani, leader of the party left, who became secretary general in July 1954 following De Gasperi's death that year, sought to reorganize the party and to make it more independent of the parallel organs. In the PSI, with the atmosphere of international détente that had set in, the autonomist current (Nenni) favoured breaking away from the PCI and collaborating with the government on a reform program. In the PCI, Togliatti, after the 20th congress of the Communist Party of the Soviet Union and the events in Hungary (1956), based the PCI's position on the noninevitability of war, on a democratic advance toward Socialism, on "national roads to Socialism," and on the "polycentrism" of the international Communist movement. The PLI, under its new secretary, Giovanni Malagodi (1954), moved to the right, establishing closer links with the business world.

The elections of May 1958 revealed the substantial stability of the electorate, with slight gains for the DC (42.2 percent) at the expense of the rightists. Following the elections, the DC and the PSI moved toward a rapprochement. The progressive attitudes of the church under Pope John XXIII (1958) also favoured this trend. The result—after considerable intraparty manoeuvring and a series of unstable governments, including a right-wing administration (Fernando Tambroni, March–July 1960) that aroused fears of a Fascist-style revival—was the "opening to the left" (*apertura a sinistra*). The formation of a left-of-centre government came during Fanfani's fourth administration (February 1962–May 1963). The coalition included the DC, the PSDI, and the PRI. The PSI, as agreed upon, maintained its abstention, which was interpreted as a show of confidence in the government. In the elections of May 1962, Antonio Segni was chosen president of the republic.

Opening to the left

Industrial development. The years 1952–62 saw the doubling of the national income and a 62 percent increase in per capita income. This "economic miracle" was largely the result of the development of manufacturing, which made Italy a predominantly industrial country. Industrial production rose from 27 to 44 percent of total output, while employment in this sector grew from 29.6 to 38.6 percent of the working population. At the same time, employment in agriculture fell from 39.6 to 27 percent (and still further in subsequent years) of the people, with a mass migration from the countryside. The population shifts from the countryside to the towns and from south to north, involving millions of people, created serious overcrowding, urban sprawl, and substandard housing in many parts of the northern "industrial triangle" (Milan, Turin, Genoa). The increase in investments, the technological modernization and rationalization of plants, the creation of the European Economic Community (EEC), and the successes achieved by semipublic enterprises in the iron and steel sector were matched by the surplus of manpower, which helped to depress wage levels and encouraged further emigration.

The parliamentary shift to centre–left. The elections for the fourth republican administration (April 1963) disappointed the two left-of-centre protagonists. The DC (38.3 percent) lost part of its conservative support to the PLI while the PSI remained at a standstill. The PCI and the PSDI both gained. After a short-lived (June–November 1963) Christian Democrat government under Giovanni Leone, Aldo Moro finally formed (December 5) a left-of-centre "organic government" (one with direct Socialist [PSI] participation) committed to the adoption of economic programming, establishment of locally autonomous regions, and reform in the urban, school, and agricultural sectors. The PSI, however, faced internal difficulties. The leftist splinter group, opposed to collaboration with the DC, left the party to set up (December 1963) the Italian Socialist Party of Proletarian Unity (Partito Socialista Italiano di Unità Proletaria, or PSIUP), which was joined by some 20 percent of the PSI parliamentarians. President Segni resigned (because of illness) in December 1964 and was replaced on December 28 by Giuseppe Saragat.

Presidency of Saragat

The Moro government, faced with inflation (closely connected with an expansion in consumption caused by higher wage levels) and a distressing balance-of-payments deficit, introduced measures to hold back consumption. It also contracted for a loan of $225,000,000 with the U.S. Treasury and the International Monetary Fund. The economic situation brought the reform policy to a halt and led to a new series of crises. In March 1966, for the third time, Moro formed a centre-left government, which was finally able to introduce reforms such as the act on economic programming, a prerequisite for instituting the locally autonomous regions. Measures dealing with university reform and family rights, however, were delayed.

The Christian Democrats, meanwhile, were torn between leftist elements, which favoured détente in relations with the Communists, and right-of-centre elements opposed to any such rapprochement. The warring Socialist factions managed to reunite and at a joint congress (October 1966) formed the Unified Socialist Party (Partito Socialista Unificato, or PSU).

In the elections of May 19–20, 1968, the DC held its ground (39.1 percent), but the PSU suffered heavy losses and some of the Socialist voters switched to the PCI and, especially, the PSIUP. The Monarchists practically vanished, and the PLI and the MSI also fell behind. Giovanni Leone headed a Christian Democrat one-party caretaker administration from June to November 1968, when Mariano Rumor formed (December 13) a coalition of the DC, PSI, and PRI. Rumor's coalition lasted until July 1969, when the PSI's social democrat wing split off. Rumor, therefore, formed a new one-party government (August 1969) that had to face acute political and social tensions. The three major national trade-union confederations struck for the renewal of their contracts and for a new domestic policy

Rumor coalition

(November 19). This critical situation was exacerbated by an attempt to blow up a Milan bank (December 12, 1969) in which several persons were killed.

Faced with this situation, the left-of-centre parties formed a government to offer better guarantees of stability and efficiency. After a long crisis, which began on February 7, 1970, a third Rumor left-of-centre coalition was set up in April. It introduced some highly significant legislation, including the Workers' Statute (offering more effective guarantees of the workers' freedom and dignity and trade-union liberty in business enterprises), the referendum to repeal legislation, and the regional finance act. On July 6 Rumor unexpectedly resigned. He was replaced (August 6) by another Christian Democrat, Emilio Colombo, under whose leadership the centre–left coalition passed two measures that had aroused great controversy, the Divorce Bill and the Finance Bill. Colombo's government, however, was unable to halt the deterioration of the economy, and he resigned in January 1972.

Elections in May failed to resolve the political deadlock. Giulio Andreotti headed a centre–right government until he was succeeded in July 1973 by a Rumor-led centre–left coalition of Christian Democrats, Republicans, Socialists, and Social Democrats. The Communists and the trade unions adopted a tolerant attitude. In February 1974, however, the Republicans withdrew over economic policy, and Rumor formed a new administration in March, excluding the Republicans but dependent upon their parliamentary support. The right wing of the Christian Democrats forced the government to hold a referendum on the Divorce Law in May 1974. Despite strong opposition from the Roman Catholic Church, a large majority favoured the law, and the referendum was considered a severe setback for the Christian Democrats. In October, after a prolonged Cabinet crisis, the Rumor administration resigned. Italy remained without a government until the end of November, when Moro formed a coalition of Christian Democrats and Republicans. Depending on the parliamentary support of parties outside the coalition, this government was extremely weak.

Referendum on the Divorce Law

Successive governments had failed to cope with the decline of the economy and public services, corruption in high places, and the growth of lawlessness. In the regional elections of June 1975, the voters registered their discontent: the Communists attracted 33 percent of the vote. The Christian Democrats still led, with 35 percent, but their domination of political life was clearly threatened. In reaction, the party replaced Amintore Fanfani—the long-time secretary general and a stout opponent of any rapprochement with the Communists—with Benigno Zaccagnini. The Communists, meanwhile, rather than favouring the formation of a left-wing coalition government, continued to press for the "historic compromise" (*compromesso storico*), a program for Italy's future based on an alliance of Communists and Christian Democrats.

Moro's coalition collapsed in January 1976 after the withdrawal by the Socialists. To stave off a general election, regarded as futile by the major parties, Moro formed a minority Christian Democrat administration a month later. Without adequate support in Parliament, however, the government was forced to hold a general election in June. The Communists showed increased strength, receiving more than 34 percent of the vote. The Christian Democrats continued to reject the "historic compromise" and insisted on excluding the Communists from power. They were forced to seek assurances of Communist abstention, however, before a new government could be formed. At the end of July, Giulio Andreotti formed a new government, which introduced severe austerity measures to deal with the continuing economic crisis.

In July 1977, after four months of negotiations, the Communists at last received a measure of participation in the government. The opposition parties gained a significant voice in policy making, but they still had no direct role in government. The initial program included measures to strengthen the economy and to uphold law and order. Communist support for the Christian Democrats alienated the extreme left, some of whom resorted to the violent political tactics already associated with the extreme right. The arrangement, however, enhanced the authority of the government, which no longer feared defeat in Parliament, and suited both Christian Democrats and Communists.

Despite a number of parliamentary crises and increasing political violence during the autumn of 1977 and the first half of 1978, the agreement held firm. The most sensational episode in the long series of political kidnappings, shootings, and "kneecappings" of prominent businessmen, intellectuals, and members of the judiciary was the abduction of the Christian Democratic Party leader and former premier, Aldo Moro, on March 16 by members of the Red Brigades (Brigate Rosse), an extreme left-wing terrorist group. The attack took place shortly before Parliament began its debate on a vote of confidence for the latest Andreotti Cabinet, which had taken office March 13. The kidnappers attempted to bargain with the government for release of brigade members then on trial. After long and agonized debate, the government refused to negotiate with the Red Brigades.

Moro murder

Moro's bullet-riddled body was found near the centre of Rome on May 9. A week later, on May 14–15, in local elections involving about 10 percent of the electorate, the Communist Party obtained 16 percent less of the popular vote than the Christian Democrats, compared with a difference of only 3.3 percent between the two parties in 1976. The vote was widely interpreted as a sympathy vote for the Christian Democrats. (C.G.Se.)

On June 15 Pres. Giovanni Leone resigned because of allegations connecting him to the Lockheed bribery scandal in which the American aircraft company was said to have bribed high military officials and politicians to facilitate purchase of Lockheed aircraft. Alessandro Pertini, a Socialist, succeeded Leone in July as Italy's seventh president, to serve a seven-year term. When his term expired, the nearly 90-year-old Pertini was succeeded in 1985 by Francesco Cossiga, a Christian Democrat.

The rapid turnover of centre–left coalition governments, usually under Christian Democratic leadership, continued into the 1980s. The election of 1983 led to the republic's first Socialist-led coalition, the premiership being taken over by the leader of the Socialist Party, Bettino Craxi. This coalition nearly collapsed in 1985, in connection with events following the hijacking of the Italian cruise ship *Achille Lauro* by four Palestinians claiming to be members of the Palestine Liberation Front, a splinter group of the PLO, but Craxi was able to win a vote of confidence in the Chamber of Deputies by a secure margin.

Despite the history of rapid government turnovers—well over 40 since the establishment of the republic in 1946—the Italian political scene has remained remarkably stable. Since the election of 1963 government coalitions have been centre–left. Although the Communists have never been officially part of the government, they have, through various agreements, been able to exert influence on governmental decisions.

Economic troubles. Even though inflation and economic stagnation have troubled the Italian economy since about 1969, it, nonetheless, has shown remarkable resiliency. For the five-year period ending in 1976, Italy ranked seventh among the West's main industrial powers in its gross domestic product. In its economic growth rate, only Canada and Japan exceeded the Italian achievement during that period.

The government made numerous efforts to stabilize the economy. The most serious problem besetting the economy was the increasingly high level of public spending. To ease the problem, the Craxi government passed an austerity budget in 1983, which included measures such as tax increases, cuts in social services and pensions, and a proposal to modify the highly controversial *scala mobile,* the inflation-linked quarterly wage adjustment. Italy's economic problems, however, are substantially alleviated by a vigorous "submerged economy," which operates outside of government control and which is estimated to amount to as much as one-third of the national income.

For later developments in the political history of Italy, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (C.G.Se./Ed.)

TRADITIONAL REGIONS

The regions in this section are arranged according to geographical location. First will be found the regions in northern Italy (Liguria, Lombardy, Piedmont, Sardinia, and Venetia); next, those in central Italy (Tuscany, Emilia-Romagna, and the Papal States); lastly, those in southern Italy (Mezzogiorno and Sicily).

Liguria

PHYSICAL AND HUMAN GEOGRAPHY

Liguria, the smallest of the regions of Italy, is composed of the provinces of Genoa, Imperia, La Spezia, and Savona. With an area of 2,089 square miles (5,410 square kilometres), Liguria is shaped like a semicircle, reaching from the mouth of the Roia River to that of the Magra and from the French-Italian frontier to Tuscany. The region is dominated by the Maritime Alps as far as the Cadibona Pass and by the Ligurian Apennines east of that point. The narrow, picturesquely indented coastal fringe, the Italian Riviera, is customarily divided into a western section, the Riviera di Ponente, and an eastern section, the Riviera di Levante, the point of division being the apex of the Ligurian arc at Voltri. The mountains rise steeply behind the coast, and small rivers that cut deeply into the mountains form narrow valleys. The port city of Genoa (Genova) is the capital of Genoa (Genova) Province and figures prominently in the history of Liguria. It stands at the head of a large gulf, close to the Lombard Plain and the Alpine passes leading north.

Because of the shelter from the winter winds afforded by the Alps and the Apennines, Liguria is particularly favoured in the production of early vegetables, flowers (especially in the westernmost section), olives, and wine. Some livestock is raised in the mountains. Industries are concentrated in and around the city of Genoa, around Savona, and along the shores of the Gulf of La Spezia. Genoa and La Spezia contain the leading shipyards of Italy, and La Spezia is Italy's major naval base. Iron and steel and machinery are produced in Savona, Imperia, and Genoa; chemicals and petrochemicals at Genoa and at Vado outside Savona. Textiles and food industries are located in nearly all the major cities. Many forms of handicrafts, including ceramics, weaving, and ivory and filigree work, are also produced. Not least among the region's sources of income is the tourist trade, and there are numerous resorts scattered all along the coast.

The main railroad lines connect Genoa with Nice and Marseille in France to the west, with Pisa and Rome to the southeast, and with Milan, Turin, and Switzerland to the north. An automobile toll road runs from Genoa to the Po Valley.

HISTORY

Ancient history. In the earliest historical period, Liguria referred to an area where in the Early Iron Age Hallstatt culture there was great similarity in the Urnfield pottery and metal types (including a peculiar razor type) between Catalonia, Languedoc, Roussillon, southern Switzerland, and northern Italy, through all of which burial in flat graves was prevalent. It has been argued that it was to this Early Iron Age cultural unity that the Greeks applied the ethnic name "Liguri," although the objection can be raised that this identity in Urnfield types no longer applied at the period of the first Greek settlements.

The affinities that exist between the peoples stretching from the Pyrenees to the Arno in prehistoric times are Neolithic rather than Iron Age, and it is probable that the Urnfield elements of the Early Iron Age left the indigenous ethnic stratum undisturbed. This stratum was basically a Neolithic one, and it is to this or rather to types of Neolithic culture in this area that archaeologists are now apt to apply the ethnic term Ligurian. This Neolithic mode of life continued in village settlements in remote places despite intrusive Celtic elements of later Urnfield invasion, and it was probably to loose political groupings of these people that classical authors attach the name. Rough handmade pottery with finger-impressed decoration of basically Neolithic type has been found in the lowest levels of the Greek city of Ampurias in Spain and of the southern French Celtic fortresses of Ensérune and Cayla de Mailhac, at both of which Iberian influence was strong from the 5th century BC onward.

No texts speak of Ligurians (or of Iberians for that matter) in southern Gaul as nations or attribute definite racial characteristics to them. They were said in classical sources to have made up the indigenous population of the northwestern Mediterranian coast from the mouth of the Ebro River in Sapin to the mouth of the Arno River in Italy. Scholars have had difficulty differentiating the Liguri from Iberians. Ancient authors (Strabo, Diodorus Siculus) describe them as a rough and strong people whose piracy the Romans deplored. These, however, are late texts and refer to the Celticized Ligurians (Celtoligures) between the Rhône and the Arno rivers. Strabo tells us distinctly that they were of a different race from the Gauls or Celts, and Diodorus mentions that they lived in villages and made a difficult living from the rocky mountainous soil.

Their boldness caused them to be in great demand as mercenaries. They served Hamilcar in 480 BC and the Sicilian Greek colonies in the time of Agathocles and openly sided with Carthage in the Second Punic War. Not until 180 BC were steps taken for their final reduction by Rome, when under the proconsuls P. Cornelius Cethegus and M. Baebius 40,000 Ligurians were deported to Samnium and settled near Beneventum (Benevento).

Stretching from Gaul to Etruria, Liguria was made the ninth division of Italy by Augustus and contained the following tribes: the Friniates on the northern slopes of the Apennines, the Briniates and the Apuani in the Vara and Magra valleys respectively, the Genuates around Genoa, and the Veturii to the west of these, the Vediantiti around modern Vence, the Intemelii whose capital was Albium Intemelium (modern Ventimiglia), and the Ingauni whose capital was Albium Ingaunum (modern Albenga). North of the Apennines there were lesser tribes, the Vagienni around Augusta Vagiennorum (modern Bene), the Statielli around Aquae Statiellae (modern Acqui). The Taurini near modern Turin and many other tribes listed by Pliny and Livy were also considered Ligurians.

Origins and early history (to the 10th century) of Genoa. In ancient times, Genoa was first the site of a Ligurian fort, and in the 5th and 4th centuries BC its fine natural harbour made it a commercial emporium with Etruscan, Phoenician, and Greek contacts. From the 3rd century BC it was a major Roman station on the coastal route to Provence; another road led into Lombardy. The barbarians who overran Italy when the Roman Empire dissolved were defeated by the Byzantine general Belisarius in the mid-6th century AD, and Genoa remained under Byzantine rule until its conquest in about 643 by the Lombards, who destroyed its walls and allowed its trade to decay. Yet even in the 10th century, when Genoa was repeatedly sacked by the Saracens, a bare minimum of urban life and overseas trade still survived.

Genoa in the 11th–19th centuries. Genoa began to emerge as a leading power as early as the 11th century. The Muslim raids had provoked the local nobles to lead Genoa's fishermen and farmers, often in alliance with Pisa, in retaliatory attacks against the Muslims in Corsica, Sardinia, Sicily, and the Balearics. The spoils were invested in new ventures in Naples and Amalfi, in Sicily, Spain, and North Africa. After 1097 participation in the Crusades brought the Genoese profitable opportunities to hire out shipping and lend money, and Genoa secured trading quarters and privileges in Syrian and Byzantine ports. The Genoese extended their dominion eastward and westward along the coast, developed business activities inland, and supplemented meagre local produce by importing foodstuffs and raw materials; they became experts and innovators in shipbuilding, navigation, and cartography,

Genoese expansion

in industrial and banking techniques, and in types of contracts that enabled even poor men to form partnerships and invest capital in lucrative overseas trade.

The commune of Genoa. The gradual evolution of the commune of Genoa, whose autonomy received recognition from the Holy Roman Emperor in 1162, provided a form of government that alleviated social tension and assisted the landed nobility and the flourishing bourgeoisie to collaborate in foreign enterprises. Perpetual strife among the magnates, however, led in 1191 to the appointment of a foreign podesta (chief magistrate). Despite further innovations, such as a local *capitano del popolo* ("captain of the people") in 1257, internal dissensions continued to encourage foreign intervention. In 1339 the Genoese Simone Boccanegra became the first of a series of local rulers, called "doges," who were chosen for life but could seldom diminish faction and disorder.

Trade rivalries. Genoa's remarkable expansion involved constant struggles with trade rivals. In 1204 the Venetians manipulated the Fourth Crusade to secure predominance in Byzantium; but the Genoese reversed the position by assisting the exiled emperor Michael Palaeologus to recapture Constantinople in 1261. In 1284 Genoa's fleet destroyed Pisan sea power at the Battle of Meloria. Corsica and Sardinia were long-established spheres of influence and colonization, but Genoese merchants were able to penetrate to distant India and England; to set up colonies in the Crimea, at Phocaea with its alum monopoly, and on Chios with its mastic; to bring back slaves and gold from North Africa; and to establish commercial communities in Cyprus and Castile.

Genoa became one of Europe's largest cities; by 1300 its population possibly approached 100,000. It had strong walls, grand patrician palaces, and churches built in black and white marble stripes. Luxurious living standards were based, in part, on imported domestic slaves.

In the 14th century, European colonial rivalries were heightened by a general economic recession and, from 1348 onward, by the disastrous visitations of the Black Death (plague). A series of bitter wars against the Aragonese, who secured control of Sardinia, and against the Venetians came to a climax in 1380, when the Genoese narrowly failed to capture Venice itself.

Genoa was as successful overseas as Venice even though the Genoese lacked the disciplined constitution and public spirit that enabled Venice to run a rigidly state-controlled colonial empire, and despite Genoa's weakness in the face of public bankruptcy and private individualism. From 1396, however, internal disorder resulted in repeated submission to foreign rule by the French and by Milan. The amalgamation in 1405 of many government creditors to form the independent Banco di San Giorgio saved the Genoese government by taking over the disastrous national debt; but the bank secured so many privileges that it came to rival the state itself. After 1453 the Genoese lost all their Black Sea and Levantine colonies except Chios to the Turks, yet they found the resilience to diversify their activities and shift them westward, discovering new commodities and markets. Genoese entrepreneurs intensified their initiatives in Aragon, Castile, and Portugal and participated in new ventures along the African coast and in the Atlantic isles. Christopher Columbus, the discoveror of America in 1492, was a Genoese.

Foreign influence

Alliance with Spain. In 1528, in the midst of a great European struggle for supremacy, the famous Genoese admiral Andrea Doria shifted his service from France to the Holy Roman Emperor Charles V (Charles I of Spain); Genoa escaped French domination through a Spanish alliance that allowed Genoese financiers, the great specialists in exchange operations, to handle huge sums for the Spanish crown. The Genoese controlled Spanish and Neapolitan trade, and Peruvian silver poured into their banks; by about 1570 they were the principal bankers of Catholic Europe. Also in 1528 Andrea Doria introduced a new constitution giving power to the magnates, with doges elected for two-year periods. This regime provided an era of more stable though increasingly oligarchic government.

Loss of independence. Real decadence came as Genoa was increasingly excluded from the prosperous Northern and Atlantic economy. Its fortunes were dictated by the great European succession wars. The city was ruinously bombarded by the French in 1684; was occupied in 1746 by the Austrians, against whom there was a popular uprising; and lost its last Mediterranean colony, Corsica, to France in 1768. In 1796 Napoleonic troops occupied Genoa, which sustained a terrible Austrian siege in 1800. The Genoese expelled the French in 1814, but the next year's peace treaty gave Genoa to Piedmont. The resulting discontent, both republican and anti-Piedmontese, shaped Genoa-born Giuseppe Mazzini, the great prophet of the Italian Risorgimento. In 1860 Giuseppe Garibaldi sailed for Sicily from Genoa with his army of liberation.

Genoa since 1860. Genoa became the major port of the new unified Italy, rivalling Marseille in France. Railway building, industrial development, and shipbuilding yards all inserted Genoa into the great industrial complex of northern Italy, while the Simplon and other Alpine tunnels greatly enlarged its hinterland. Genoa has an important university and the other attributes of a great modern city. Its revolutionary traditions erupted in a successful insurrection against the Germans in April 1945. The heavy damage sustained during World War II was repaired, and Genoa has remained one of Italy's greatest towns.

(A.T.L./G.Kh.)

Lombardy

PHYSICAL AND HUMAN GEOGRAPHY

The northern region of Lombardy (Italian Lombardia) is composed of the provinces of Bergamo, Brescia, Como, Cremona, Mantova, Milano, Pavia, Sondrio, and Varese. With an area of 9,191 square miles (23,804 square kilometres), it is the leading industrial and commercial region of Italy and one of the country's best farming areas.

Physically, Lombardy may be divided into three zones: a northern, mountainous zone; a median, hilly zone; and a southern, flat zone. The northern zone is divided into the Alpine and the pre-Alpine zones. The Alpine zone, where crystalline rocks prevail, comprises part of the Leopontine and Rhaetian Alps, the Orobie Alps, the Ortles, and the Adamello. In the Bernina it reaches a height of 13,304 feet (4,058 metres) and has many glaciers. The pre-Alpine zone, mostly calcareous, though also dolomitic, attains less elevated heights but occasionally rises above the 8,000-foot line. It is particularly beautiful at some of its massifs, such as the Grigne. The hilly zone, partly composed of a morainic material with some morainic amphitheatres, is gently undulating. The alluvial plain, sloping northwest–southeast, is divided into high and low areas; the former has a gravelly soil, poor in superficial water; the latter, with plentiful moisture content, is separated from the former by the spring line where the waters, hidden in the subsoil at the higher level, gush out.

Lombardy has many rivers, all tributaries of the Po. The principal ones are the Ticino; the Adda, with its affluents the Brembo and Serio; the Oglio, with its affluents the Mella and the Chiese; and the Mincio. The Valtellina and the Brembana, Seriaria, Camonica, Trompia, and Sabbia valleys are among the most beautiful in Italy. The region also has many Alpine, pre-Alpine, and infra-morainic lakes, containing all or part of Lake Garda, the largest lake in Italy, Lake Maggiore, the Lake of Lugano, Lake Como, Lake Iseo, Lake Idro, Lake Varese, and the lakes of the Brianza (Pusiano, Annone, Alserio, and Segrino).

The climate, though in the main continental, is variable, because of the great differences of height and the presence of large water areas; it is most continental on the lower plain at Milan, Pavia, and Cremona. The rainfall, not less than 24 inches (610 millimetres) annually in the area near the Po, reaches 80 inches in the mountainous regions.

Lombardy has a higher than average population density. The density is greatest in the pre-Alpine zone and on the plains, the areas of intense economic, agricultural, and industrial development.

Agriculture, especially on the plains, is industrialized, and high productivity is achieved by scientific use of fertilizers and by irrigation. Grasslands, where grass is mowed up to eight times a year in the *marcite* ("flooded meadows"), cereal growing (rice, wheat, and corn [maize]), and sugar-beet cultivation are characteristic of the low plains; the higher plains grow cereals, green vegetables, fruit trees, and mulberries. The hilly zone has fruit and chestnut trees; the climate and soil around the lakes are especially suitable for olive trees and limes. In the pre-Alpine zone vines grow at altitudes as high as 2,400 feet above sea level, and on the Alps there is excellent grazing. In the Alpine hamlets cattle breeding is scientifically practiced; about half of the cattle are milk producing. Pigs are also raised, and sheep are bred for both wool and meat. Honey is another important product. Lombardy ranks high in the production of silk cocoons.

A national park, of about 350 square miles, was set up at Stelvio in 1935 for the preservation of indigenous animal life.

Lombardy is part of the industrial triangle of northern Italy, marked by the cities of Genoa, Turin, and Milan. The Milan metropolitan area is known for steel and iron, automobiles and trucks, machine tools and machines, chemicals and pharmaceuticals made in Milan, Monza, Sesto San Giovanni, and other centres. Brescia, Pavia, and Cremona manufacture trucks, engines, and machinery; Como, Legnano, and Gallarate, textiles; the Brianza district, furniture; Vigevano and Varese, leather. Food industries are located in many of the smaller towns. Milan is the leading industrial city of Italy, its principal banking centre, and its leader in wholesale and retail trade. The region's railway network radiates from Milan, which has direct rail communications with Switzerland, France, and West Germany through the Simplon and St. Gotthard passes; with Turin and Paris via the Mont Cenis rail line; with Venice, Trieste, and Yugoslavia to the east; with Genoa, Bologna, Rome, and Bari to the south and southeast. Automobile expressways connect Milan with Turin, Genoa, Venice, and Florence, and there are excellent highways throughout Lombardy.

HISTORY

The Lombards appear in classical writings of the 1st century AD as one of a number of tribes who formed the Suebi. Their home was then evidently in northwestern Germany, on the left bank of the lower Elbe River. If distinctive types of late Iron Age pottery and brooches are rightly associated with them, their settlements extended both west and east of the medieval Bardengau, which, with its principal community, Bardowiek (near Lüneburg), preserves the second element of their tribal name (Langobardas; classical Latin Langobardi, medieval Latin Longobardi). According to a tradition written down in the 7th century they arrived there from a more northerly home. The same source asserts that they got their name from their long, uncut beards; but a different explanation is possible.

Before AD 6, Roman soldiers had fought against them; in AD 17 they supported the Cherusci against the Marcomanni; and shortly after AD 47 they helped reestablish a deposed ruler of the Cherusci. About AD 166 Lombards were among the barbarians who attacked the Danube frontier. The main body of the Lombards, however, is generally thought to have pursued a settled pastoral existence in northern Germany until the beginnings of the great migrations late in the 4th century, and even after this others remained to be absorbed by the expanding Saxons. The tradition written down in the 7th century that the migration southward was associated with a change from leadership by a duke to rule by a king, in the person of Agelmund, may be authentic. The same source's account of the stages by which the Lombards moved to the middle Danube is extremely obscure. There is some archaeological evidence, however, that they established themselves successively along the upper Elbe and in the later Moravia, where in the mid-5th century they were temporarily part of the Hunnish empire of Attila.

4th-century migrations

At the end of 487 the kingdom of the Rugii, which roughly coincided with Austria north of the Danube, was destroyed, and the Lombards occupied the area. Contact with the Ostrogoths and Franks brought them into touch with late Roman culture. The earliest examples of elaborately decorated brooches, in which the styles and techniques current in these societies have in turn been adapted to Lombard taste, appear in their graves shortly after this. In the early 6th century the Lombards temporarily expanded toward the Tisa River, which may have been one reason for the war fought *c.* 508 under their king Tato against the Heruli and their king Rudolf, which ended in the almost complete destruction of the Heruli. The Lombards then, according to their 8th-century historian Paulus Diaconus, "began of their own accord to seek occasions of war." At this period also they probably received their first Christian missionaries. When, later, they entered Italy they were certainly Arians; but there is evidence that at one time they had favoured orthodoxy and subsequently changed, possibly under Gothic influence. Tato was killed *c.* 510 by his nephew Waccho, who ruled for 30 years, his kingdom extending across the Danube as far as Lake Balaton. Shortly after 536 the Byzantine emperor Justinian I made a treaty with him directed against the Gepidae who held the country east of the Danube. But for nearly a decade the Lombards gained little or no advantage from it, since Waccho's successor, from *c.* 540 to 546, was a minor. In 546 Audoin began a new royal dynasty; and at the beginning of his reign he and his subjects were allowed to establish themselves in the formerly Ostrogothic lands as far as and even beyond the Save. At that time, it seems, they began to adapt their tribal organization and institutions to the imperial military system of the period, in which a hierarchy of dukes, counts, and others commanded warrior bands formed from related families or kin groups. For two decades intermittent wars with the Gepidae alternated with truces and the attempts of both sides to gain imperial support for their own ends. About 565 Audoin died. In the same year the Avars appeared in the west; with them Audoin's son and successor Alboin made a compact to destroy the Gepidae. The decisive battle (*c.* 567), however, was fought between the Lombards and the Gepidae only. The latter were destroyed and their king killed, and Alboin married his daughter Rosamund.

Even before this the Lombards seem to have had their eyes on Italy, where the Byzantine armies had recently overthrown the Ostrogothic kingdom. In the spring of 568, having arranged with the Avars that they could return to and reclaim their lands in Pannonia if they did not like Italy, they crossed the Julian Alps.

The Lombards' invasion of northern Italy was almost unopposed, and by September 569 they had conquered all the principal cities north of the Po except Pavia. Subsequently Alboin sent armies across the Apennines; in 571 he may have threatened Rome itself, though only western Emilia and part of Tuscany were permanently conquered at this time. Simultaneously two other armies respectively occupied Spoleto with the land eastward to the Adriatic and Benevento with adjacent areas of southern Italy. Pavia fell in 572. Shortly afterward Alboin was murdered in revenge for having forced his wife to drink from her dead father's skull. The 18-month rule of his successor, Cleph, was marked by the ruthless treatment of the Italian landowners.

On his death the Lombards chose no successor. Instead the dukes, who with their followings had been or were now associated with one of the cities in the occupied area, exercised authority in their particular city-territories. Place-names and archaeological evidence indicate that the war bands that settled south of the Apennines were fewer than those that settled in the Po Valley and the northeast. But everywhere the Lombards were outnumbered by the native population, and they seem to have established themselves in groups in easily defended positions on the outskirts of the cities or in similar "castles" in the countryside. For a time they may have lived largely on tribute, though quite early many of them must have acquired lands of their own.

The 10-year "rule of the dukes" was later viewed as one of violence and disorder, in which the church suffered at least as much as the native landowners. In 584, however, threatened by a Frankish invasion that the dukes had provoked, the Lombards made Cleph's son Authari king; and when he died in 590 he was succeeded by Agilulf, duke of Turin, who married his widow Theodelinda, a Bavarian and a Catholic. A Franco-Byzantine alliance, which this marriage was partly intended to counter, had inflicted serious defeats on the Lombards by 590, but Agilulf was subsequently able to recover most of what had been lost. In the next decade Rome itself was threatened, first by the Duke of Spoleto and subsequently by Agilulf, being saved only by the efforts of Pope Gregory I. By 605 the Lombards were in complete control of the Po Valley and Emilia east of the Panaro; the Byzantines retained only the Venetian coastal strip, the Ligurian coast, the Po basin and the Adriatic coast southward, Perugia and the adjacent Apennine crossings, Rome and its neighbourhood, and Naples and small areas in the southeast and southwest. In 643 a Lombard army under King Rothari (reigned 636–652) occupied Liguria.

When Authari became king, the dukes surrendered half their estates for his maintenance and his court, and royal officials knowns as gastaldi were appointed to administer them and (eventually) to act as a check on the power of the dukes. Border duchies such as Friuli and Trento were always difficult to control, and except for relatively brief periods the dukes of Spoleto and Benevento were able to act independently of the king; but the internal weaknesses of the Lombard kingdom must not be exaggerated. By the 620s Pavia was emerging as something like a capital. The royal palace (built by Theodoric the Ostrogoth) was the centre of an administrative organization whose officials and techniques, like the documents they wrote, owed much to Roman traditions as transmitted by the Byzantines. Rothari's "Edict" of 643, in which the laws of his people were recorded for the first time, illustrates many aspects of this "Romanization"; but it also shows how tenaciously the Lombards maintained many of their Germanic customs—though some, such as vendetta, had been modified by Christian and royal influence.

Arianism

The Lombards' Arianism helped to maintain their separateness from the king's Roman subjects. Theodelinda's Catholicism had little permanent influence except for her support of the influential monastery of Bobbio. Arian bishops continued in parts of northern Italy, and Rothari was actively anti-orthodox; it was only in King Perctarit's reign (671–688) and after the intervention of Roman missionaries that the Lombards abandoned Arianism.

Theodelinda's line died out with the brutal Aripert II (reigned 700–712), and a new dynasty was raised to the throne. Its second representative, Liudprand, who reigned from 712 to 744, was probably the greatest of the Lombard kings. Until 726 he seems to have been concerned exclusively with the internal condition of his kingdom. Subsequently, helped by the internal dissensions resulting from imperial policies, he steadily reduced the area still under Byzantine rule; and he also made his authority effective in Spoleto and Benevento. The laws which he issued in 15 of the 31 years of his reign reveal an increase in royal power, a growing opposition to violent revenge, and the greater importance of property transactions. Coins and documents from his court confirm the impression of a strong and effective monarch.

In 751 Ravenna and the remaining central Italian Byzantine territories were occupied by Aistulf (king 749–756), who then invaded the territory around Rome. On the appeal of Pope Stephen II (III), the Frankish king Pepin led two expeditions into Italy in 754/755 and 756, and compelled Aistulf to surrender to the papacy most of his conquests since 751. When in 772 Aistulf's successor Desiderius invaded these papal territories, Pope Adrian I sought the help of the Frankish king Charles (Charlemagne). The latter entered Italy in 773 and after a year's siege Pavia finally fell to his armies. Desiderius was captured, the great men of the kingdom made their submission, and Charles became king of the Lombards as well as king of the Franks. The region was later ruled by Spain (1535–1713), Austria (1713–96), and France (1796–1814). In 1859 Lombardy came under Italian rule.

The traditions and achievements of the Lombards were not forgotten with the disappearance of an independent kingdom. In the 780s Paulus Diaconus, himself of an old Lombard family, wrote their history from the beginning to the death of Liudprand; and in the 9th century the princes of Benevento, whom the Franks failed to subject, regarded themselves as maintaining the ancient traditions of the Lombard people. The Italian language includes a number of words of Lombard origin; and for at least two centuries the judicial institutions of northern Italy showed traces of Lombard influence. The Lombards also played a part in the development of later Germanic ornament (in the 7th century), though their contribution to the art of Italy was negligible. Finally many features of later Italian life which have Roman origins, such as the notariate, were transmitted via the Lombards. (A.C./G.Kh.)

Piedmont

PHYSICAL AND HUMAN GEOGRAPHY

The northwestern region of Piedmont (Piemonte) is composed of the provinces of Alessandria, Asti, Cuneo, Novara, Turin (Torino), and Vercelli. It has an area of 9,807 square miles (25,400 square kilometres). To the south, west, and north it is surrounded by the vast arc of the Ligurian Apennines and the Maritime, Cottian, Graian, and Pennine Alps. The core of the region is the Po Valley, open to the east and consisting, especially in its eastern portion, of some of the best farmlands in Italy. South of the Po are the low and intensively cultivated hills of Monferrato and of the Langhe. In the foothills of the Alps are Lakes Maggiore and Orta. The Po and its tributaries, the Dora Baltea, Dora Riparia, Sesia, Tanaro, Ticino, and Scrivia, provide ample water for agriculture.

The alpine arc of Piedmont plays a vital part in the power production of the region and of north Italy as a whole; its hydroelectric plants provide energy for industry, transportation, and domestic use. The forests provide lumber; the alpine and sub-alpine meadows afford excellent pasture for cattle as the base of a prosperous dairy industry. The lowlands produce wheat and rice, vegetables and fruit, milk and cheese; the hills south of the Po are noted for some of Italy's highest quality wines, both of the sparkling (Asti) and still (Barbera) variety. Piedmont is part of the great industrial triangle of north Italy (Turin–Genoa–Milan), and its industries are characterized by their variety as well as by their important output. Turin, the largest city and the leading industrial centre, is the location of one of Europe's largest automobile plants, as well as of printing, textile, and machine industries. Ivrea, northeast of Turin, is the headquarters of one of the leading makers of office machinery in Europe. Textiles, chemicals, paper, rubber, nonmetallic minerals, and glass are among the other important Piedmontese industries.

Through Piedmont passes the principal rail connection between France and Italy, the Turin–Mt. Cenis Tunnel–Paris line, while to the north the Simplon Tunnel leads to Switzerland. A network of roads and expressways ties all parts of the region closely together; Genoa, easily reached from Piedmont, serves as its port. An all-weather road between France and Italy, passing through a long tunnel under Mont Blanc, thence through the Valle d'Aosta region to Turin and Milan, and a tunnel under the St. Bernard Pass between Switzerland and Italy were opened in the 1960s.

HISTORY

In Roman times Piedmont played a role of importance because its passes served as connections between Italy and the transalpine provinces of the empire. After periods of Lombard and Frankish rule, the House of Savoy emerged as the most important feudatory of northwest Italy. This dynasty first became powerful as successor to the marquesses of Ivrea and of Turin, but after 1400 its control of both slopes of the Alps, ruling over what is now French Savoy (Savoie) and over Piedmont, gave it undisputed sovereignty over much of the region. After 1700 practically all of Piedmont passed under Savoyard domination, while the addition of Sardinia to its territories provided it with still wider interests. During the Italian Risorgimento it was Piedmont that led the attempts of 1848, 1859, and 1866 to unite all Italy, and Victor Emmanuel II, originally king of Piedmont, became modern Italy's first king in 1861. (G.Kh.)

Sardinia

Character and location

The second largest island in the Mediterranean Sea, second in size only to Sicily, Sardinia (Italian Sardegna) was for centuries ravaged by invaders, malaria ridden, backward, and poverty stricken. Almost ignored by the modern world until the 1950s, this autonomous region within the Italian republic was later included in an ambitious industrial development scheme. It also became a major tourist centre.

Ichnusa or Sandaliotis, the names given to the island by the Greeks, derived from its shape, which resembles a sandal or foot, the later name, Sardinia, stemming from Latin.

Divided into the provinces of Sassari (in the north), Nuoro and Oristano (in the centre), and Cagliari (in the south), Sardinia measures some 150 miles (240 kilometres) from north to south and about 75 miles at its widest point, with a surface area of 9,194 square miles (23,813 square kilometres). It lies 112 sea miles from the Italian mainland and 120 from Africa, which is just visible on a clear winter's day from Sardinia's most southerly point, Capo Teulada. Its island neighbour, Corsica, lies 7½ miles to the north, across the Strait of Bonifacio.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Among the natural features of the island, the mountains are outstanding, particularly the Gennargentu in Nuoro, the Limbara heights in Gallura, and the famous *macchia,* the grasslands mingled with scrub of cistus, lentisk, myrtle, prickly pear, and dwarf oaks, which covers most of the uncultivated countryside. Yet man, too, has left an impressive imprint: among a notable range of architectural monuments, generally Romanesque or Pisan in style, the Paleo-Christian Church of San Gavino at Porto Torres is exemplary. Rising, like many of its kind, among solitary, rolling hills, the beautiful Pisan Saccargia di Santa Trinità (located near Sassari), of alternate limestone and basalt, lends its own character to the landscape, as do the ancient citadel walls rising high above the modern city of Cagliari.

Topographically, a pattern of rolling uplands predominates. It leads in the eastern centre to the Gennargentu, with La Marmora (6,017 feet [1,834 metres]), the highest point in Sardinia, just above Bruncu Spina (6,001 feet). The Punta Balestieri (4,468 feet) tops the granite range of the Limbara, surrounded by ancient cork forests. Only in the extreme northwest and in the west coast near Oristano are there large areas of low-lying land. The plain of the Campidano, 60 miles long, running diagonally from Cagliari to Oristano, the granary of imperial Rome, is an exception. Today, its fertile soil produces much of the island's corn, as well as fruit and vegetables.

Climatic patterns

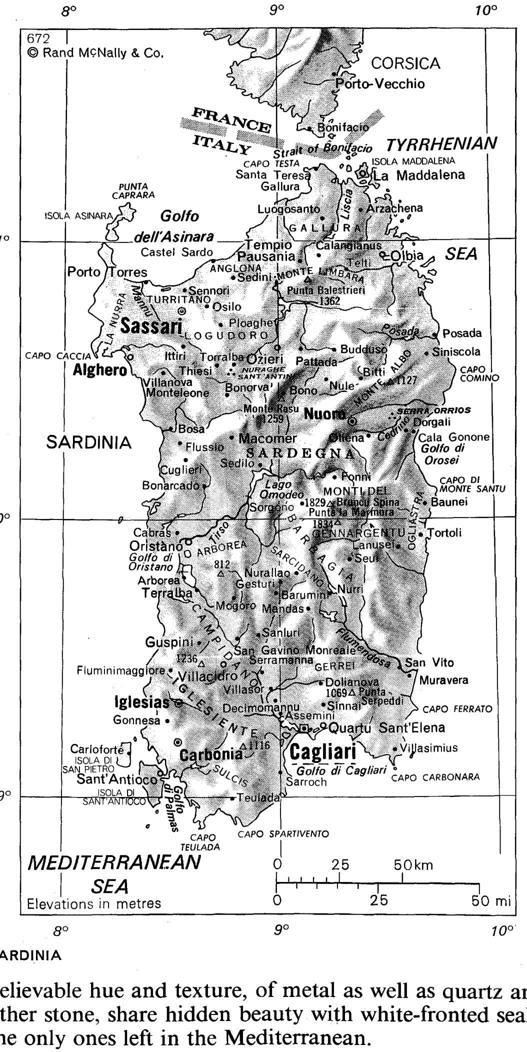
With some nine months each year of sunshine, Sardinia has a predominantly mild climate; there are only slight variations of temperature—between 43° and 55° F (6° and 13° C) in winter and from 61° to 79° F (16° to 26° C) in summer, except in the mountains, where snow may lie on the topmost peaks for as long as six months. The name Gennargentu—"the silver gate"—refers to the spectacular effect of the sun on the gleaming snow crystals. There is generally a breeze, the prevailing wind being the *maestrale,* from the northwest; the *greco* is from the east, and only when the *sirocco* blows from the southeast is the air damp and oppressive.

Rainfall is low, the mean in Cagliari being only 18.9 inches (480 millimetres), in Sassari, 23.1 inches, though this increases in the higher lands. The longest rivers, the Tirso (93 miles) and the Flumendosa (79 miles), both rise in Nuoro, fed by innumerable mountain streams that dry up in summer. There are more than 26,000 springs on the island, and some 200 lakes have been formed, with major reservoirs on the Tirso and its tributaries. For months from the end of January onward the island is a paradise of fruit blossoms and of wild flowers that yield their fragrance and colour to the thyme and other herbs mingling in the *macchia.* Oleander blooms abound in every shade from white to vivid reds. In addition to cork oaks, ilex, chestnuts, and pines throng the forests.

Cattle, goats, and pigs are allowed to wander freely in the uplands, with only sheep being herded as protection against foxes and other predators. In early summer the shepherds lead their flocks to the highlands, the men taking only a little wine and their *carta da musica* (wafer-thin bread), finding meat and water as they need it, milking their ewes and making *peccorino* cheese.

Donkeys abound; the only albinos in the world are found in Asinara, and at Castel Sardo there is a diminutive breed hardly larger than a big dog. The Sard's horse is prized above all, and its well-being was protected by laws in medieval times. In the heart of the island, in the plain of Gesturi, there is still a breed of wild horse, swift and beautiful, of unknown origin, also protected by law. Conservation measures also apply to the mufflon (horned sheep), the rare red and fallow deer, and, in the marshes of Cagliari, a species of heron and also the fleeting spring flocks of flamingos. Other wild birds include eagles, quail, and woodcock. Wild-boar hunting is allowed only at certain seasons, and there are no poisonous snakes.

The 500 miles of coastline, fringed with small bays and endless white sands, are riddled with deep caves, many unexplored. The two most famous are the Grotta di Nettuno (Cave of Neptune), among the coral beds of Alghero, and the Grotta del Bue Marino (the Grotto of the Sea Ox), stretching more than two miles long at Cala Gonone on the east coast. Here, stalactites and stalagmites of unbelievable hue and texture, of metal as well as quartz and other stone, share hidden beauty with white-fronted seals, the only ones left in the Mediterranean.



SARDINIA

Sardinia has also three first-class natural harbours: at Cagliari; at Asinara on the northwest; and, on the extreme northeast, among an archipelago of seven small islands, the harbour of La Maddalena, where Lord Nelson held his fleet for nine months before setting out on the voyage that led to Trafalgar. He declared it to be the finest harbour in the Mediterranean and—unavailingly—besought his country to take it.

The people. A people whose origin has remained unknown, the Sards (and their similarly mysterious language) have inevitably been influenced by the successive nations that occupied the island. Some of the linguistic results are not without humour: at Alghero, half the people speak Catalan but the others Sardo, so communication has to be in Italian. Similarly, the islanders of San Pietro still speak Genoese, inherited from refugees from Tabarca who settled there in the 18th century; on the adjacent San'Antioco, Sardo is the language, so once more Italian must be used. In general, Tuscan has a definite sphere of influence north of a line from Olbia, through Tempio, to Castel Sardo; Genoese predominates around Sassari, with the exception of Alghero, Spanish, and Arabic in the south; and the purest native strains, both of people and of language, are found in the isolated mountains of the Gennargentu. The strongest foreign associations are Spanish, contacts with the Iberian Peninsula having been dated from as early as the 2nd millennium BC. The legendary foundation of Nora by an Iberian chief, Norax, from Tartessus, before the Phoenician arrival, is perhaps associated with this link. Certainly, modern Sardo includes many Spanish words.

Linguistic complexities

The Sardinians are a devout Catholic people. The population has increased in recent years, in spite of considerable migration. Distribution is uneven, the only considerable concentrations being in Cagliari and Sassari, and then, much further down the scale, at Nuoro, Iglesias, and Oristano.

Profound material and psychological changes are besetting contemporary Sard life. All serious observers agree that the Sards are a dignified, brave, and loyal people, whose backward, almost tribal way of life, especially in isolated villages and mountain areas, is suddenly being catapulted, as it were, from the flint to the jet age, without the centuries of gradual transformation experienced on the mainland of Europe.

In the past, robbery or violence among the Sards arose generally from either hunger or vendetta; the former is often tempered by generosity from others hardly better off, and from it stems, in part, the fiercely held tradition of hospitality to the stranger. Loyalty to friend or family has, for centuries, facilitated survival in the face of savage butchery. It is a natural habit as ingrained as sleeping and eating, the results of failure to observe it in the past often being so dire that only blood could wash away the stain. It is said, with pride, that if a Sard is your friend he is your friend for life and that no Sard will ever betray another to a "foreigner." That includes Italians, who, after all, were invaders, too.

The economy. The two most important developments in the economy in recent decades have been the inclusion of Sardinia in the Italian government-sponsored scheme of aid for the developing areas of southern Italy, Sardinia, and Sicily and the extermination of the malaria mosquito. Known as the Cassa per il Mezzogiorno (Fund for the South), the former project gave financial help and technical advice toward the establishment of new businesses, assistance in the training of workers, temporary remission of taxes, and other incentives. The plan was based on the idea that in the scheduled areas lies a potentially enormous supply of untapped labour that could help to give Italy an important lead in export markets and increase its value within the EEC. Petrochemical plants in Sardinia at Porto Torres and at Sarroch near Cagliari were extended, and other far-reaching plans involved foreign as well as Italian investors.

The campaign against the mosquito

The extermination of the malaria mosquito, begun in 1943 by the United Nations Relief and Rehabilitation Administration, UNRRA, with a pilot scheme at Oristano, was followed in 1947 by a campaign that covered the whole island and that was financed by the Rockefeller Foundation of the U.S. For centuries the ubiquitous disease, known as *intemperie,* was thought to be caused by the climate, and its elimination, apart from the obvious improvement in health of the people, helped to treble the number of foreign tourists over the next decades. An outstanding example of tourist development was the construction of a luxury resort, the Costa Smeralda, at Arzachena, by an international consortium headed by the Aga Khan IV.

Sardinian mining for gold, silver, lead, copper, and other ores has been known since ancient times, and the island is the most important source of fluorspar in Europe. Poor-quality coal is mined at Carbonia and is mainland Italy's only source of natural supply, as is the sea salt from the lagoons of Carloforte and Cagliari. Hard wheat is the important crop; olives, citrus fruits, and grapes are also grown. The centuries-old cork forests of Gallura produce most of Italy's total output. Other industries include the making of *peccorino* cheese and the processing of such wines as the *vernaccia* and *Malvasia,* from the Ogliastra, and the celebrated red wine of Olbia.

At Stintino, on the Golfo dell'Asinara (Gulf of Asinara), and also at Carloforte, fishing for *tonno* (tunny) is skilled, well organized, and profitable for all. It is financed by Genoese. The only difficulty with this industry is that the course of these great fish, storming their way through to spawning grounds in the Black Sea, remains uncertain until the last moment.

The building of good motor roads linking the main centres of Sardinia and improvements of lesser routes, in conjunction with considerable government concessions to tourists bringing cars to the island, greatly increased motor traffic. Rail communications between the main centres are now fairly good, but other services are slow and limited. The two international airports at Cagliari and Alghero were joined in the early 1970s by a third, at Olbia. A private airline connects the Costa Smeralda with the mainland. Maritime services are good and include car ferries. Connections are frequent between the Sardinian towns of Cagliari, Olbia, Golfo degli Aranci, and Porto Torres and the mainland towns of Civitavecchia, Genoa, Naples, Palermo, Toulon, and Tunis.

Administrative and social conditions. Article 1 of the special statute for Sardinia, dated February 26, 1948, states: "Sardinia and its islands shall be considered an autonomous region with its own legal status, within a united Republic of Italy." Government is administered by a small executive body, the *giunta,* appointed from among members of the Regional Council, themselves elected on the basis of proportional representation in the ratio of one councillor for every 20,000 inhabitants. Members of the Council elect a president from among themselves; he has the right to be present at Cabinet meetings in Rome when matters concerning Sardinia are involved. Only in rare cases—usually where expert knowledge is required—are other than regional councillors appointed to administrative posts within the *giunta.*

The constitutional background

Elections are held every four years. Every citizen reaching the age of 21 has the right to vote but cannot become a councillor until the age of 25. There are no postal votes; although, in theory, absentee voters abroad can claim free second-class transport "to the border," this privilege is not much exercised.

The wide legislative powers held by the *giunta* affect nearly all spheres of life: social services, labour, industrial conditions, hotel and tourist trade, and public works, including building and urban planning. The region retains the bulk of the income collected from the many tax sources, as well as from a state tobacco monopoly. The central government nevertheless retains control of customs and excise, the armed forces, the *carabinieri* and the finance police (who watch the entry of goods and may examine corporate records in tax matters), the urban and rural police having only limited functions. The railways are also state property, but other means of transport within the island, as well as local sea and air traffic, are the responsibility of the *giunta.* Two years of military service are compulsory, as on the mainland.

Under Italian law every citizen has the right to free education, and in Sardinia "subscription" is available to assist talented children from poor families to attend universities. Taxes include contributions toward medical care and pensions, the latter being payable at the age of 55 for women and 65 for men.

Within the framework of the Regional Council the four provinces have considerable administrative powers over their domestic affairs, all of which are jealously guarded. Isolation of small villages is one of the problems but is, to a growing extent, being overcome by the organization of communes, whereby two or more communities join for mutual benefit and protection. Each commune has at least one policeman, while in other, solitary villages, the local mayor has authority to enlist up to 15 volunteers, known as *baracelli,* whose powers are limited but who perform a useful function. The origin of this force goes back to the days of the Spanish vigilantes.

Cultural life. As might be expected in an island with so long and varied a history, folklore and craftwork abound in Sardinia, in many cases the origins and inspiration for both going back to forgotten centuries and pagan beginnings. Witchcraft and both black and white magic are still part of life. No one realized this better than the Nuorese authoress Grazia Deledda, who was given the Nobel Prize for literature in 1926 for her understanding portrayal of the power and passions of life in the primitive communities around her.

Magical survivals

Every town and village has its own festival, but the most important is the Sagra di Sant'Efisio at Cagliari in early May, which commemorates the martyrdom of a Roman general who was converted to Christianity. All Sardinia flocks to the capital for this spectacle, which affords an opportunity to witness the exquisitely embroidered (and often valuable) costumes of the women. The men's national costume includes, together with the stocking cap seen on the nuraghic bronzes, a tunic said to be similar to that worn under their armour by Roman soldiers.

Most of the festivals involve feats of horsemanship, as at the Sartiglia of Oristano, the dangerous Ardia at Sedilo, and the Cavalcata at Sassari, where there is also a Feast of the Candles. At Mamoiada, in Nuoro, a remarkable feature is the Feast of the Mamuthones, where men, wearing masks and sheepskins on their backs loaded with bells, perform a ritual dance, ending with a symbolic "killing" of the scapegoat. The variety of songs and dances is endless; they are often performed to the accompaniment of the *launeddas,* ancient triple pipes.

Much of the craftwork, still unspoiled by the intrusion of modernity, is skilled and beautifully made, some being decorated with designs based on Punic and even earlier symbolic patterns. Gold filigree, often allied to coral, is always worn on special occasions. Wood carving exhibits a great variety and includes the making and decorating of the cassapanca, the bridal chest. Other examples include leatherwork from Dorgali, basket making of palm or asphodel from Castel Sardo or Flussio, as well as finely woven carpets from Nule and Ploaghe, filet lace from Bosa, ceramics and terra-cotta from, among other places, Tortoli, Assemini, and Oristano, and the finely woven *arazzi,* or wall hangings, made in Mogoro, some of which show Arab influence.

HISTORY

Ancient origins and the classical period. The dominating feature of the island (7,000 examples are said to exist) is the nuraghi: strange, truncated cone structures of huge blocks of basalt taken from extinct volcanoes, built without any bonding. Most are quite small, a few obviously fortresses, with wells and other defensive measures; two of these latter, Sant'Antine at Torralba and Su Nuraxi at Barumini, are three stories, and about 50 feet, high. One nuraghe is always within sight of another, the greatest concentrations being in the northwest and south centre. There is also an important nuraghic village at Serra Orrios, near Dorgali, with traces of nearly 80 buildings identified, including temples, wells, and a theatre. The rock-cut tombs of Anghelu Ruju, the Dolmen, Tombe dei Giganti (Tombs of the Giants, for mass burial), and Domus de Janas (Witches' Homes) are also yielding much interesting information to archaeologists.

But though modern science has helped to identify and approximately date weapons, jewelry, utensils, and votive objects of metal, pottery, obsidian, and other stones, almost nothing is known of the nuraghic people themselves. The expression of *il vero Sardo* "the true Sard," referring to a certain type of man or woman, rarely met with, obviously different, may hint at an ethnic survival, and there are a few place-names that have no origin in Greek, Punic, or Latin tongues. Expert opinion now gives the dates of the nuraghi from about 1500 to 400 BC, but the mystery remains how such a people, well organized, with remarkable engineering skill and an equal talent in the creation of the beautiful and often witty bronzes to be seen in the Cagliari and Sassari museums, apparently left no trace of a written word.

Phoenicians were the island's first recorded settlers, at about 800 BC. They traded in metals and founded colonies in the south at Nora, Sulcis, Bithia, Tharros, and Karalis (or Cagliari). The Greeks soon raided the north and sacked the town of Olbia, to be followed in turn by the Carthaginians, against whom the Sards revolted repeatedly. In 238 BC the long and brutal Roman occupation began. It was to last nearly 700 years, with Sardinia becoming the first Roman province and Cagliari a port for the Roman fleet in AD 46.

Phoenician, Roman, and Vandal imprint

The medieval and modern period. After the Romans there were the Vandals, about AD 477, followed briefly by the Romans again, then Byzantines, and later still the Saracens, who in 711 sacked and occupied Cagliari, forcing the luckless Sards to pay heavy ransom.

Ninth-century records mention the *giúdice* (governing judge) for the first time, and at the beginning of the 11th century the four divisions of the island are described: that of Cagliari (for the south), Arborea (the centre), Logudoro (Torres), and Gallura (northeast), each with its own *giúdice.* By this time the Italian cities of Pisa and Genoa, with papal support, were struggling against each other and the Sards for domination of the islands. Vatican influence shifted, however, to Alfonso IV of Aragon, who in 1326 defeated the Pisans, taking Cagliari by force. Later in the same century, the famous warrior queen Eleanor of Arborea unsuccessfully rallied the island against the invaders. Her outstanding work, the codifying of the laws begun by her father, was nevertheless completed, and in 1421, after her death, her Carta de Logu was accepted by the Sard Parliament as valid for the whole island and remained so until the Treaty of Utrecht in 1713. In the 15th century, meanwhile, under Spanish occupation, Sardinia was raised to the status of a dominion, with its own viceroy. In 1720 Victor Amadeus II of Savoy was proclaimed first king of all Sardinia, and in 1861 the accession of Victor Emmanuel II finally united Sardinia with Italy, although it was not until 1948 that Sardinia was given a degree of autonomous government. (M.De.)

Venetia

PHYSICAL AND HUMAN GEOGRAPHY

The name Venetia is used in English to denote that part of northern Italy that lies east of Lombardy, with various senses according to the historical period in question. In classical Latin, Venetia meant the territory of the Veneti. In modern Italian, however, Venezia by itself is simply the name of Venice; but if some qualifying term is added, a larger area can be designated.

The Republic of Venice extended its power over the Italian mainland northward, westward, and eastward in the Middle Ages. The name Venetia may thus be applied in a general sense to cover these old Venetian possessions, which finally included all the modern region of Veneto, sometimes also called Venezia Euganea, and a large part of the modern region of Friuli-Venezia Giulia.

The republic's territory west of Lake Garda and the Mincio River (Brescia, Bergamo, and Crema, held from the late 15th century) was, however, regarded as part of Lombardy rather than of Venetia; nor did the latter name include Dalmatia, likewise a possession of the republic.

The modern region of Veneto (also called Venezia Euganea) comprises the provinces of Venice, Padua, Rovigo, Verona, Vicenza, Treviso, and Belluno. Its area is 7,095 square miles (18,377 square kilometres). It is surrounded by Lombardy on the west, Trentino-Alto Adige to the north, Emilia-Romagna to the south, and the Adriatic Sea, Friuli-Venezia Giulia, and Austria to the east and northeast. The northern limit of Veneto is marked by a mountainous area, including the Dolomites, between Lake Garda at the southwestern extremity and the Carnic Alps

(Alpi Carniche) at the northeastern end. The southern portion consists of a fertile plain extending to the Gulf of Venice and watered chiefly by the Po, Adige, Brenta, Piave, and Livenza rivers. The mouths of these rivers, especially the several mouths of the Po, form an extensive delta area with shore lagoons. The climate is maritime in the south and turns Alpine toward the north.

Agriculture

Veneto is a chief producer of wheat, sugar beet, and hemp; while corn, dairy cattle fodder, and fruit (apples, pears, peaches, cherries), as well as vines on the sunny slopes of the mountains, are also grown. Cattle are extensively raised on the plain and fishing is practiced in the Venetian Lagoon. Much use is made of irrigation, and there has been a good deal of land reclamation, especially in the Po delta. After World War II large estates were expropriated for distribution to smallholders. Sources of energy for the region are methane gas from the Miocene rocks of the Po plain and hydroelectric power from the Alpine area. The larger towns of the plain have textile, silk, lace, hemp, paper, founding, and shipbuilding industries. Marghera, the port of Venice, has zinc and aluminum smelting plants and oil refineries, and it also produces sulfuric acid and synthetic ammonia for the chemical industry. Venice, besides being a tourist centre, carries on its traditional glass-, lace-, and tapestry-making crafts. The adjacent island of Murano specializes in glassware and that of Burano in lace. The only other notable port is Chioggia, south of Venice, a centre of fisheries and lace making. Verona is a chief communications centre between Italy and central Europe and an agricultural market, with engineering, chemical, and wine industries besides artistic handicrafts; Padua is an agricultural centre and has engineering, silk, textile, and chemical industries; Vicenza is also an agricultural market and manufactures textiles, fertilizers, and luxury goods. Treviso has textile mills and food-processing and other industrial plants; Rovigo is the agricultural centre for the Po delta and has a sugar-refining industry; Valdagno is a model industrial city and textile centre.

The region is well served by a dense road network, notably the Venice–Milan–Turin motorway (*autostrada*). Venice is connected to the mainland by a road and rail bridge. The chief local traffic route in Venice is the Grand Canal. (G.Kh./Ed.)

HISTORY

Origins. A people called the Veneti arrived in this territory about 1000 BC. They became allies of the Romans, who founded in 181 BC the colony of Aquileia, later famous as a Christian patriarchate. The region suffered severely from the barbarian invasions. Venice came into existence after the fall of the Roman Empire in the West. The Lombard invasion of northern Italy in AD 568 caused inhabitants of Altino and Aquileia, to take refuge in the lagoons, where only a few fishermen and salt workers lived without fixed abodes. The first islands to be occupied were Torcello, in the Venetian lagoon north of Burano, and Grado near Aquileia in the Marano lagoon. When the Exarchate of Ravenna was created *c.* 584, the region formed part of it. In 607 the patriarchate of Aquileia was transferred to Grado. When Oderzo, the last remaining mainland city of the Byzantines, fell to the Lombards in 641, political authority was transferred to an unnamed island later called Cittanova Eracliana in the Venetian lagoon. The Veneto-Byzantine area was then restricted to the lagoons and isolated politically, though not for trading purposes, from the other Byzantine possessions in Italy. The refugees formed new communities on islands of the lagoon and set up administration units.

The first elected duke was Orso, chosen in an anti-Byzantine military declaration in 727, but he was succeeded by Byzantine officials until *c.* 751, when the Exarchate of Ravenna came to an end. At the same time there was internal crisis between Cittanova and Malamocco, a settlement on the Lido in command of the main communication routes that had been developed by exiles from Cittanova. In the second half of the 8th century, Malamocco wrested from Cittanova the prerogatives of government but was not strong enough to unite the local administrations in a centralized regime.

Internal political strife

This internal political strife was complicated by the attempts of the patriarchs of Grado, who had extended their temporal influence along the coast as far as Equilo, to support the defeated inhabitants of Cittanova and Torcello against Malamocco. The doge Maurizio set up the bishopric of Olivolo in the Rialto island group, in order to check the southward advance of Grado, having already allied himself with the surviving Lombards and the Byzantines of Istria to oppose the Frankish expansion under the Carolingians. The new doge Obelerio and his brother Beato formed an alliance with the Franks of Italy and brought Venice under the subjection of the young king Pepin (died 810) in order to free themselves from Byzantine overlordship.

Rialto. Pro-Byzantine reaction to this event under the doges of the Partecipazi family led to the transfer of the seat of government from Malamocco to the Rialto group of islands, by then the natural centre for the exiles of the factional fighting. The move both centralized political activity and assured territorial and political independence. These were accompanied by a parallel development of the social and economic life of the lagoon islands, particularly the Rialto itself. Though a Franco-Byzantine treaty of 814 guaranteed to Venice political and juridical independence from the rule of the Western Empire, it did not confirm any effective dependence on the Byzantine Empire, and by 840–841 the doge was negotiating international agreements in his own name with Western governments without involving the Byzantine authorities. This freedom of Venice from Byzantine control was never sanctioned by any diplomatic or juridical document, but it became hallowed by custom and by centuries of uncontested power. The unusual legal and political position of a small duchy situated in territorial isolation between two great empires, in friendly relations with the West, and in theoretical dependence on but effective independence of the Byzantines, contributed greatly to its function as a trading middleman.

A succession of serious internal crises concerning the office of doge, from the time of the Partecipazi to that of the Badoer, Condiano, and Orseolo families, did not halt the rapid development of trade, which increased the national wealth and extended Venice's political influence from the Adriatic to the Mediterranean. The increase in private wealth led to the gradual achievement of internal stability by creating a broader ruling class that was capable of putting a limit to the power of the doge. Gradually a national consciousness developed, of which the first signs reflected in the daily life of the citizens were the reforms of Doge Orso Partecipazio I in the late 9th century. The national church, ruled by the patriarch of Grado, was reorganized by the institution of five new bishoprics. From the time of Giovanni Partecipazio's successor (887), the doge was chosen by popular election, though without destroying the monarchic system or the custom of co-regency, which assured the continuity of power. Finally the group of Rialto islands was solemnly transformed into the city of Venice (*civitas Venetiarum*).

The new order. The final collapse of family faction rule under the Orseolo regime led to a change in the system of government, inaugurated by Doge Domenico Flabanico (1032–42). He restored to the people the sovereign right (obscured during civil strife) to elect the doge, but the term *populus* was in practice restricted to the residents of the Rialto and, more narrowly, to the group of nobles who regularly frequented the doge's palace. The executive organ was the ducal curia (*dux* and *judices*), and the legislative assembly was summoned to approve the doge's acts. The temporal authority was preserved from ecclesiastical intervention, while the national church, centred on the patriarchate of Grado, was strengthened by the abandonment of metropolitan jurisdiction (Istria) outside the boundaries of the duchy. A new church was built for St. Mark, symbol of the Venetian spirit, under Doge Domenico Contarini (1043–70), an energetic defender of Grado's metropolitan rights and of the religious independence of the duchy.

Expansion of trade. In external affairs Contarini and his successors remained neutral (despite the complaints of Pope Gregory VII) in the conflict between papacy and empire, while safeguarding Venetian economic interests in the Adriatic when the conflict began to be reflected on the Dalmatian coast. But the greatest danger to Venetian economic interests was the 11th-century Norman expansion under Robert Guiscard, which threatened to cut Venetian communications to the south. The successful action taken against the Normans by Doge Domenico Selvo and his successor Vitale Falier was intended more to assure Venetian freedom on the sea than to aid the Byzantine Empire, and it made clear that Venice's control of trade routes in the Mediterranean must rest on a firmer basis than mere usage. In gratitude for Venetian aid against the Normans, the Byzantine emperor Alexius I Comnenus granted Venice unrestricted trade throughout the Byzantine Empire, with no customs dues, a privilege that marked the beginning of Venetian activity in the East (1082). The Adriatic, however, was not yet under control, as the Dalmatian ports were threatened by the Hungarians and Slavs, with whom it was difficult to come to agreement. Zadar (Zara), the most important of these ports because it controlled the northern Adriatic, changed hands frequently until it finally became Venetian in 1409.

Norman expansion under Robert Guiscard

Toward the end of the 11th century the Crusades centred the newly awakened trading interests of the West on the Mediterranean. At first, however, Venice was concerned chiefly to gain control only of the European trading ports of the Byzantine Empire, leaving to private interests the trading opportunities with Syria and Asia Minor, although being prepared to intervene later if it should prove to be profitable to do so. As the 12th century progressed, the two other Italian merchant republics, Genoa and Pisa, came into conflict in the new trading area that had been opened up. The Venetians also, who were the first to win trade concessions from the Byzantine Empire, aroused the hatred of the Byzantines by their arrogance and highhandedness and by the conflict of interests continually rekindled by day-to-day contacts in the same market. Thus when the Byzantines and the Venetians should have been cooperating at sea against a revival of Norman expansion in Corfu (1143–44), they let slip the fruits of victory and vented on each other the fury of mutual hatred. In 1169 the Byzantine emperor Manuel I Comnenus made a trade agreement with the Genoese and in 1170 with the Pisans. In 1171 he tried to free himself from Venetian competition by arresting every Venetian in the empire and confiscating his goods.

The commune. All this time the expansion of Venice abroad and along the borders of the lagoon in the communes of Padua, Treviso, and Ferrara, as well as in the patriarchate of Aquileia, not only enriched its patrimony but also created an awareness of its own political power. Between 1140 and 1160 the most revolutionary change in Venice's political structure took place: the doge lost his monarchic character, becoming a mere official (though he still assumed resounding titles), and the commune took over the powers, functions, and prerogatives of the state. All political and administrative matters were placed in the hands of the *Maius consilium* (Great Council) of 45 members, which had evolved from a lesser body called the *Consilium sapientium.* A Minor Council of six members exercised executive powers alongside the doge, and magistrates were granted administrative and judicial functions. The financial system was elaborated and the market of the commune was controlled by executive organs.

Reorganization of the state

This systematic reorganization of the state was necessitated by the social development, the economic expansion corresponding to the scale of political needs, and the national communities formed on the Dalmatian coast. The reorganization was achieved in hard times under pressure of rivals in the Mediterranean; of an advancing policy by Manuel I Comnenus and by Frederick I Barbarossa, simultaneously converging on the Adriatic; of a spirit of revolt in the Dalmatian cities from Capodistria to Pola (now Pula, Yugoslavia) and Zadar; of disorders to be quelled in the mainland communes; of wary collaboration with the Lombard League, which governed the trade routes to the hinterland; of controlling the trading posts on the Italian coast from Ferrara to Ancona; and of guaranteeing free access to Apulian ports.

The hatred between the Byzantines and the Venetians reached its culminating point when the doge Enrico Dandolo diverted the Fourth Crusade to sack Constantinople in 1204. This reaction of the Western world in defense of its own interests brought great territorial gains on the continent to those who took part in it, but Venice preferred to assure for itself the dominion of the seas by acquiring the title of "lord of the fourth part and a half" (*Dominus quartae partis et dimidiae*) of the Byzantine Empire. The republic thus took possession of an extensive island patrimony from the Aegean to Crete and the lookout posts of Modon and Coron (modern Methoni and Koroni, Greece), and dominated the whole eastern Mediterranean, reaffirming an economic supremacy that quickly aroused the covetousness of the Genoese and the Pisans. The tremendous conflict between Venice and Genoa (the Pisans played a minor role) lasted for the better part of two centuries. Early attempts at cooperation were wrecked by the pressure of events, especially in Syria, where the Genoese had prior rights. Defeated at Acre (1258), the Genoese by the Treaty of Nymphaeum (1261) with the Byzantine Empire (then based at Nicaea) engineered the downfall of the Latin Empire of Constantinople set up by the Venetians in 1204. The Venetians, however, not only retained their island possessions but also recovered their trading privileges in Constantinople after a Genoese-Byzantine conflict had broken out. Syria had by then fallen to the Muslims, and the rivalry between Venice and Genoa was played out in Constantinople, dragging on in a long war of privateering on the seas.

The patriciate. Meanwhile at home the Venetian state was being built up. In 1242 the civil statutes of Jacopo Tiepolo regulated civil and economic relations, and maritime statutes had been established in 1239. From the Popular Assembly to the Great Council the scope of the commune was progressively enlarged. The elective members of the Great Council were raised from 45 to 60 and then to 100 by an increase in the ex officio members (the total of magisterial office holders). The Council of 40 (first mentioned in 1223) received powers of jurisdiction, and the *Consiglio dei Rogati* (60 members; founded mid-13th century), invested with the control of economic affairs, in time assumed all legislative functions and the honorific title of Senate. National, political, and economic interests abroad were protected wherever Venetian influence existed by obtaining trade concessions and reorganizing the national organs of jurisdiction in the colonial centres (from Constantinople to Tyre and Acre) or abroad (from Zadar to Ragusa, Crete, and Euboea).

In the 11th–12th century the Michiel and Falier families had tried to perpetuate ducal power, and restrictive electoral systems had been introduced to prevent the formation of family factions. In the 13th century similar attempts by the Ziani and Tiepolo families also failed, and the governing class strengthened its organization by translating into law the order already hallowed by custom. In 1268 an interlocking process of choice by lot and voting alternately among the members of the Great Council was introduced in order to select 41 persons who then, by a majority of not less than 25 votes, agreed on the next doge to take office.

Between 1290 and 1300 new laws restricted the right to take part in the government to families traditionally performing magistrate's duties. The patrician class was not created by the "closing of the Great Council" (*serrata del Maggior Consiglio*) achieved by these laws, but it received its legal status from them. Henceforward anyone claiming personal power had to act outside the patrician order and rely on the people; and the people were linked so closely to the patricians by economic needs that he would never find sufficient support. Thus the conspiracy of Marin Bocconio failed (1299), as did those of Bajamonte Tiepolo and the Querini brothers (1310), and later of Marin Falier (1354).

The special conditions of Venetian society created a governing class very different from that of the other Italian communes or of the continental states. To counter any attempts at sole personal rule, the Council of Ten was established (1310) to police the patrician order and defend the existing regime.

Struggle for naval supremacy. This consolidated internal regime aroused the jealousy of those who felt offended by the arrogance of Venice. The republic now had to fight at sea, to defend its maritime interests, and on the mainland as well. It had taken a cautious part in the negotiations leading to the Lombard League's truce of 1177 and had intervened more openly in the crusade against Ezzelino da Romano. By the beginning of the 14th century it was swept into struggles on the mainland of Italy and became involved in Adriatic and Mediterranean problems. Thus Venice took an active part in a war with Ferrara to safeguard the vital trade route of the Po, disputed by the Holy See, and when the Scaligers came to power in Verona the republic made alliance with the Carraresi of Padua, with the Florentines, and with the Visconti of Milan who feared the rise of a strong territorial lordship in the heart of northern Italy. Deviating from its strictly maritime policy, Venice established sovereignty over Treviso, thereby assuring itself of its own food supply but also providing itself with a land frontier to be defended.

Rivalry with Genoa

Meanwhile, the antagonism and rivalry with Genoa were rekindled. On Genoa's side was the King of Hungary, the Patriarch of Aquileia, and the Visconti of Milan, while Venice had the support of the Carraresi and the distant king of Aragon. The conflict, chiefly carried on in Dalmatia, was made more difficult for all by the spread of the Black Death (1348), by the economic and financial crisis caused by the prolongation of the war itself, and by the inanity of operations, which quickly disillusioned and dismantled the ill-assorted coalitions. The alternation of victories and defeats brought no conclusion to either side, but both exhausted their energies and resources. At last a second anti-Venetian coalition brought the war almost into Venice itself: at Pola and at Chioggia, Venice was first defeated and then won the victory (1380–81), triumphing under Vettor Pisani. The Peace of Turin (1381) eliminated Genoese political influence from the Mediterranean and the East, leaving the Venetian government as arbiter of the sea routes.

Conquest of the mainland. The Venetian victory over Genoa took place under the threat of Turkish advance in the East. The Venetians had to negotiate a state of neutrality with the Turks and find another base to compensate for the smaller yield now to be expected from the East. So they turned to the Italian mainland to rid themselves of the nuisance of neighbouring lordships and later to defend and exploit the rich land they had acquired, whose yield was no less important than overseas trade for the national economy. The restlessness of the Carraresi, Scaligers, Aquileiesi, and Visconti encouraged intervention, at first through subsidies to support one party against another, and later by the dissolution of the realm of Gian Galeazzo Visconti (1402). For a time Venetian territorial rule went no further than the rivers Mincio and Livenza, thus including the Treviso area and the Carraresi and Scaliger lands of Padua and Verona together with a lien on the Polesine (Rovigo) up to the Po, which belonged to the Este family. But beyond the Livenza lay the politically and economically important principality of the patriarch of Aquileia through which passed the main routes to Germany and to Istria. As the patriarch could not guarantee peace and order, Venice incorporated the principality in the Venetian domains (1420).

Venetian territory now covered roughly the areas of the modern regions of Veneto and Friuli-Venezia Giulia together with the Istrian Peninsula, and the doge Tommaso Mocenigo claimed that his city had reached its political and economic zenith. It had a solid base in Italy that could largely compensate for its losses in the East, and it should not expect indefinite progress—in fact, efforts to enlarge its conquests might be dangerous, and it was better to preserve, not to risk, its accumulated wealth. This warning was not heeded, however, by Mocenigo's successors, who feared that standing fast on the positions so far conquered would mean the beginning of a decline.

The doge Francesco Foscari risked a further policy of conquest in mainland Italy, dismembering the realm of Filippo Maria Visconti while the Turks encroached upon the Byzantine Empire in the East. Foscari carried out his first Lombard conquests in the territory of Brescia in 1426 just as Thessalonica fell to the Turks, and held the Adda River (1432–54) while the Turks took Constantinople (1453). Greed for territorial conquest made Venice forget the greater profit to be won in the East and involved Venetian policy, not only in the tangled web of Italian balance of power but in the conflicts between the great powers of Europe on a scale out of proportion to Venetian forces and aims. The Peace of Lodi (1454) did not ease the situation in Italy, which was threatened by foreign intervention. It was followed by the formation of the Italian League for the restoration of political balance among the Italian states, which produced a merely ephemeral accord.

The Peace of Lodi (1454)

Turkish expansion in the eastern Mediterranean after the fall of Constantinople involved Venice in war: Euboea (Negroponte) fell in 1470 to the Turks, with whom the Venetians finally made peace in 1479. But Venice was soon involved in another war, this time with Ferrara (1481), in which it conquered the Polesine (1484). This merely increased the opposition of the other Italian states to Venetian territorial expansion.

Europe against Venice. This internal discord made Italy a prey to invading foreigners, Spanish, French, and German. By 1508 these powers, together with the pope, the Hungarians, the Savoyards, and the Ferrarese, united to form the League of Cambrai against the Venetians, who were defeated at the Battle of Agnadello. Venice was saved from the worst results of this event by internal discords within the League of Cambrai, but Venetian territories on the mainland were diminished, and at the same time the consequences of the economic crisis could no longer be avoided. Not only was the Eastern market lost, but the discovery of new lands to the West and new trade routes to the East released Europe from dependence on the Venetian market. Venice ceased to be a Mediterranean power; it became a European power but without the advantages of the Atlantic states, now open to the New World.

Venetian policy in the 16th century was dictated by the need to keep intact its political, economic, and territorial heritage against the advance of the Turks on the one side and the pressure of the great western European powers on the other. This need supplied the reason for Venice's intervention in the Italian crisis of Charles V and in its struggle against the Turks, from the defeat of Preveza in 1538 to the victory of Lepanto and the loss of Cyprus in 1571; and for its tenacious resistance to ecclesiastical pressure from the pope. So Venice declined into economic stagnation, embittered by a constitutional conflict beween the *Consiglio dei Rogati* and the Council of Ten for control of the public finances. As the great Venetian statesman and historian Paolo Paruta (1540–98) pointed out, Venetian peace and neutrality meant defending the immediate interests of the nation but ceasing to take part in problems in which it was not directly concerned. Thus the spirit of political and religious conservatism grew increasingly tenacious in Venice.

Life and politics in the 17th century. Some scholars have detected symptoms of political and spiritual regeneration in the period between the end of the 16th and the beginning of the 18th centuries, but this is to overlook the slow and progressive withdrawal to prepared positions on which Venetian life was then based. The political crisis caused by the papal interdict of Venice in 1606 was not concerned with heresy or reform but with the temporal prerogatives of the papacy. Paolo Sarpi, the energetic defender of Doge Leonardo Donà's policy that had provoked the Roman Curia, never contested the legitimacy of papal power, but in the temporal sphere he denied that it carried any prerogatives superior to the sovereign rights of the state.

The only perspective open to Venetian policy was the defense of Crete, its surviving possession in the eastern Mediterranean, against the Turks and of its Adriatic possessions threatened not only by the Turks but also by the Spaniards and Austrians. After a long campaign (1645–69) Crete fell to the Turks, the Venetians being allowed to retain merely a few strongholds. This blow to Venetian morale was mitigated, however, by the preservation of Dalmatia, and the government, after allying itself with Austria, attempted to reestablish itself in the eastern Mediterranean by liberating the Morea (Peloponnese) from the Turks. There the brilliant campaign of Francesco Morosini in 1684–88 assured Venice of this new Greek territory, which was finally handed over by the Peace of Karlowitz (Karlovci; 1699). The conquest, however, proved a burden of great and unprofitable expense, and by the Peace of Passarowitz (Pozarevac; 1718) the Morea returned to the Turks. Thus ended Venetian activity in the eastern and southern Mediterranean, save for an unsuccessful attempt on Algerian and Tunisian pirates under Angelo Emo (1769).

The end of the Venetian republic. The last period of the Venetian republic was spent in neutrality, a policy inherited from its 16th-century theoreticians. It was neutral not merely politically but spiritually also, being estranged from the fervour of new ideas germinating in other nations. There was no lack of talent, no unwillingness to work or to understand, but Venetian life had crystallized into a system from which escape was not possible. Thus one cannot speak of Venetian society in the 18th century without referring back perpetually to that 16th-century crystallization that prevented any reform, whether agrarian, commercial, or political. The plans of Angelo Querini, Giorgio Pisani, and Carlo Contarini, the supposed reformers of the 18th century, did not go beyond the mentality of the noble class which for three centuries had controlled the government and which existed to uphold ancestral tradition or to satisfy personal ambition. In all the reformist literature of the century there is a chill dialectic.

Effects of the French Revolution on Venice

The end of the republic came after the outbreak of the French Revolution. Napoleon, determined to destroy the Venetian oligarchy, claimed as a pretext that Venice was hostile to him and a menace to his line of retreat during his Austrian campaign of 1797. The Peace of Leoben left Venice without an ally. The government offered no resistance, and Ludovico Manin, the last doge, was deposed on May 12, 1797. A provisional democratic municipality was set up in place of the republican government.

On the fall of the republic in 1797, during the French Revolutionary Wars, all Venetia east of the Adige River, together with the Polesine plain between the Adige and lower Po, passed to Austria under the Treaty of Campo Formio. Under the Treaty of Lunéville (1801) Austria lost the Polesine; and under the Treaty of Pressburg (1805) all western Venetia was added to Napoleon's kingdom of Italy, while western Istria was annexed to France. The rest of Istria, with the formerly Austrian Carniola, was likewise annexed to France by the Treaty of Vienna (1809). On Napoleon's downfall (1814) all Venetia was restored to Austria; and in 1815 the kingdom of Lombardy-Venetia was constituted for the Austrian emperor. Lake Garda, the Mincio, and a line drawn north–south from the Mincio west of Mantua (Mantova) to the Po formed the frontier between the Lombard and the Venetian halves of this kingdom: Venetia (German Venetien, as opposed to Venedig, Venice) was then bounded on the northwest by the county of Tirol including Trento and Bolzano; on the northeast by Carinthia; and on the east by the Austrian Küstenland (Gorizia, with Istria). In the revolution of 1848 a provisional republican government was set up by Daniele Manin, but it fell the following year.

The composite kingdom lasted until 1859, when Lombardy passed to Sardinia-Piedmont after the Peace of Zürich. After the Seven Weeks' War of 1866, in which Prussia defeated Austria, the Austrian kingdom of Venetia, but not the Küstenland, was ceded to Italy.

The Treaty of Saint-Germain (1919), after World War I, gave the Küstenland to the Italians, who then renamed it Venezia Giulia. The same treaty, however, also gave to Italy the southern Tirol. Though this area had never been Venetian, the Italians then named it Venezia Tridentina (Tridentum being the Latin for Trento) and spoke of *le tre Venezie* ("the three Venetias," namely Euganea, Giulia, and Tridentina). After World War II the Italian peace treaty of 1947 gave most of Venezia Giulia to the Yugoslavs, who received Plezzo (Bovec), Caporetto (Kobarid), and Tolmino (Tolmin) on the Isonzo (Soca) River in the north and also all Istria except the free zone of Trieste. The northern part of the free zone of Trieste was assigned to Italy in 1954, whereupon it was merged into the region of Friuli-Venezia Giulia; *i.e.,* the Italian rump of Venezia Giulia and the district of Friuli, transferred from the Veneto region.

In World War I the Austrians were defeated by the Italian forces at Vittorio Veneto in the northeast of the region. Considerable damage was sustained during World War II, notably at Verona and Treviso. The province of Udine was transferred from Veneto to Friuli-Venezia Giulia in 1947.

The region was subjected severely to storm and flood in November 1966; rivers overflowed the plain, and around Belluno in the southern Dolomites landslides caused by rains destroyed communications and engulfed houses and people. A similar landslide in 1963 had plunged into the Vaiont reservoir, drowning the village of Longarone and about 2,000 persons with the overspill.

(R.Ce./Ed.)

Tuscany

PHYSICAL AND HUMAN GEOGRAPHY

The central region of Tuscany (Italian Toscana) consists of the provinces of Arezzo, Grosseto, Florence (Firenze), Leghorn (Livorno), Lucca, Massa-Carrara, Pisa, Pistoia, and Siena, with a total area of 8,877 square miles (22,991 square kilometres). In the north and northeast the region is bounded by the Tuscan-Emilian Apennine (Appennino Tosco-Emiliano), these being separated by a series of long valleys (Mugello, Casentino) from the sub-Apennine uplands of Monte Albano, Pratomagno, and others. Quite separate from the Apennine and sub-Apennine systems are the so-called anti-Apennines of Tuscany, consisting of low mountains (*e.g.,* the Apuane Alps and the Colline Metallifera) and plateaus extending as far south as the volcanic uplands of the Lazio region. Farthest west are several isolated massifs along or near the coast (Monte Argentario, Monte Amiata). The lowlands of Tuscany are either interior valleys (*e.g.,* Val di Chiana and Valdarno) or littoral plains, the most important coastal plain being the Maremma. Watered chiefly by the Arno and the Ombrone, Tuscany has few rivers capable of supporting major hydroelectric projects, but the borax deposits at Larderello produce enough underground steam to power a major generating station. Among the mineral resources, easily worked iron ore from Elba is nearing exhaustion, but lead, zinc, antimony, mercury, copper, and iron pyrites are still produced in the region. Lignite is mined around San Giovanni Valdarno, and the marble of Carrara is world famous. Steel and iron are manufactured at Piombino, chemicals at Leghorn and near Pisa, ships at Leghorn, and textiles and ceramics in many cities. Besides larger manufacturing firms, Tuscany is famous for its artisan industries, especially in Florence (leather, lace, silver).

Agriculture

Agriculture in Tuscany is among the most prosperous in Italy, characterized by specific forms of land ownership and a variety of crops. The classic form of land ownership and utilization is share growing, the *mezzadria* system, with the landlord (who provides capital and current expenses) sharing the harvest with the tenant, who supplies the labour. Shares are fixed by law, and many tenancy contracts have been carried on for generations between landowner and tenants. Besides cereals, Tuscan agriculture is noted for wines (the wines of the Chianti district, near Siena, are the best known and most widely exported of all Italian wines), olives and olive oil (around Lucca), vegetables, and fruit. Cattle, horses, pigs, and poultry are raised in large numbers. The railway system is characterized by two trunk lines, the coast line, running from Rome through Grosseto, Leghorn, and Pisa to Genoa; and the interior line, from Rome through Arezzo to Florence, Bologna, and Milan. Construction began on a high-speed direct railway between Florence and Rome. There are numerous secondary railways and a good network of highways and of bus services. Florence is the largest city and Leghorn the leading port; Florence, Pisa, and Siena are important tourist resorts.

(G.Kh.)

HISTORY

The name Tuscany is derived from that of the Tusci, Tuscans or Etruscans. Their country, Etruria, which was finally annexed by the Romans in 351 BC, comprised not only Tuscany but also the northern part of what is now Lazio. In the 8th century AD, however, after Charlemagne occupied the Lombard kingdom in northern Italy, the name of Tuscia or Toscana became restricted to the area north of Viterbo and Bolsena. Tuscany then became a march, or frontier district, of the Frankish dominions, the principal authority in the march being, from 774, in the hands of the counts of Lucca. Boniface I, the first known count of Lucca, died in 823 and was succeeded by his son Boniface II, whose victories over the Arabs in the Mediterranean served to bring both Corsica and Sardinia into the Tuscan sphere of influence. With the decline of the Carolingian power in Italy the counts began to assume occasionally the style of duke or margrave of Tuscany. The 10th century, however, saw the rise of the House of the Attoni of Canossa, and a member of this house, Boniface, *c.* 1027 was invested with the margraviate of Tuscany by the emperor Conrad II. On the assassination of Boniface (1052), his widow, Beatrice, governed until 1076, when her daughter, the great Matilda of Tuscany, took her place.

The rise of communes

The quarrel about investiture between the empire and the papacy coincided with the rise of the communes in northern Italy, whereby a number of the more prosperous towns asserted their independence of their overlords. In Tuscany the first communes to emerge were Pisa, Lucca, and Pistoia, which, having obtained concessions from the emperor Henry IV, joined the Ghibelline faction. Subsequently Siena, Florence, and Arezzo also established communes. Florence, influenced to some extent by Matilda's benevolent attitude toward its commune, inclined, as she did, to the Guelph, or papal, side. As Matilda bequeathed all her extensive possessions to the church, her death (1115) was followed by a struggle over her inheritance between the popes and the emperors. This enabled the Tuscan cities gradually to confirm their independence until the old unity of the march was lost altogether. Widespread inundations in the Arno valley and in the Lucca area led to destitution that helped to bring about the politico-economic upheaval.

Pisan supremacy in the 12th and 13th centuries was contested by Florence and by Lucca (which eventually left the Ghibelline camp), and Pisa, though supported by Siena and Pistoia, had also Genoa for an enemy. After the defeat of the Pisan navy by the Genoese in the Battle of Meloria (1284), Florence grew to be the leading city of Tuscany. The Tuscan dialect as spoken in Florence and written by Dante came to be a standard form for Italian.

The Medici grand duchy. The later medieval history of Tuscany is chiefly that of the consolidation of Florentine supremacy and of the establishment in Florence of the dynasty of the Medici. Twice (1495–1512 and 1527–30) the Medici were expelled, but after the surrender of Florence to the emperor Charles V's forces in August 1530, they were restored, and Alessandro de' Medici became gonfaloniere for life, a dignity that was made hereditary in the family. As he held the title of duke of Città di Penna, he is generally called duke of Florence. Under his successor Cosimo de' Medici, Siena, the one remaining outpost of republicanism in Tuscany, was annexed (1559), and Cosimo was created grand duke of Tuscany (Cosimo I) by Pope Pius V in 1569. His son Francesco I was recognized as grand duke by the emperor Maximilian II in 1576. Under his descendants, Ferdinand I, Cosimo II, Ferdinand II, Cosimo III, and Gian (Giovanni) Gastone, Tuscany played but a small part in European history.

The House of Habsburg-Lorraine. Gian Gastone being childless, it was agreed in the preliminaries (1735) of the Treaty of Vienna during the War of the Polish Succession that Francis of Lorraine, future husband of the Austrian archduchess Maria Theresa of Habsburg, should succeed eventually to Tuscany in compensation for Lorraine, of which he was dispossessed. In 1737 Gian Gastone died, and Tuscany was governed for Francis, who resided in Austria, by a series of foreign regents.

Francis, who had been elected emperor in 1745, died in 1765 and was succeeded on the throne of the grand duchy by his younger son, Leopold I. Leopold resided in Tuscany and proved one of the most capable and remarkable of the reforming princes of the 18th century. He substituted Tuscans for foreigners in government offices, introduced a system of free trade in foodstuffs (at the suggestion of the Sienese Sallustio Bandini), and promoted agriculture. He reorganized taxation on a basis of equality for all citizens, reformed the administration of justice and local government, and suppressed torture and capital punishment. He also curbed the power of the clergy, suppressed some religious houses, reduced the mortmain, and rejected papal interference. With the aid of Scipione de' Ricci, bishop of Pistoia, he even attempted to reform church discipline, but Ricci's action was condemned by Rome and he was forced to resign. At the death of his brother Joseph II in 1790, Leopold became emperor and moved to Vienna. After a brief regency he appointed his second son, Ferdinand III, grand duke.

The French occupation. During the French Revolutionary Wars a French force entered Florence in 1799 and was welcomed by a small number of republicans. The Grand Duke was forced to flee, and a provisional government on French lines was established. But the mass of the people were horrified at the irreligious character of the new regime, and a counterrevolution broke out at Arezzo. Bands of armed peasants expelled the French from the countryside, with many atrocities. With Austrian help Florence was occupied and a government established in Ferdinand's name, but after Bonaparte's victory at Marengo the French returned in force, dispersed the bands, and reentered Florence (October 1800). They too committed atrocities, but they were more warmly welcomed than before by the people, after the experience of Austro-Aretine rule. Joachim Murat set up a provisional government.

Treaty of Lunéville (1801)

By the Treaty of Lunéville (1801), Tuscany was renounced by the Austrians. It was then erected into the kingdom of Etruria for the Bourbon prince Louis, son of Ferdinand, duke of Parma. When Louis died (1803), his widow, the Spanish infanta Maria Luisa, ruled as regent for her son Charles Louis until 1807. Then the emperor Napoleon obliged her father, Charles IV of Spain, to agree to the cession of Tuscany-Etruria to France, compensating Charles Louis with a Portuguese principality.

Annexed to the French Empire in 1808, Tuscany was divided into three *départements,* to be governed from 1809 by Napoleon's sister Elisa as titular grand duchess. French progressive ideas gained some adherents, but to most Tuscans the new institutions were distasteful as subversive of cherished traditions. The heavy taxes and conscription were especially resented.

The Habsburg restoration. After Napoleon's defeats in 1814, Ferdinand III returned to Tuscany, where he was received with some enthusiasm. The Congress of Vienna added some further territory to the grand duchy and guaranteed, moreover, that Lucca should in due course revert to it (as it did in 1847). The restoration in Tuscany was unaccompanied by the excesses that characterized it elsewhere, and much of the French legislation was retained. Ferdinand was succeeded in 1824 by his son Leopold II, who continued his father's policy of benevolent but enervating despotism. When the political excitement consequent on the election of Pius IX spread to Tuscany, Leopold, in February 1848, granted a constitution. For some months Gino Capponi was prime minister. A Tuscan contingent took part in the Sardinian (Piedmontese) campaign against Austria, but the increase of revolutionary agitation in Tuscany, culminating in the proclamation of the republic (February 8, 1849), led to Leopold's departure for Gaeta. Under the republican triumvirate of F.D. Guerrazzi, Giuseppe Mazzini, and G. Montanelli disorders continued, and Leopold was invited to return and did so, but forfeited his popularity by according the protection of an Austrian army (July 1849). In 1852 he formally abrogated the constitution, and three years later the Austrians departed. When in 1859 a second war between Sardinia and Austria became imminent, revolutionary agitation broke out once more. There was a division of opinion between the moderates, who favoured a constitutional Tuscany under Leopold as part of an Italian federation, and the popular party, led by Ferdinando Bartolommei, who aimed at the unity of Italy under Victor Emmanuel. At last a compromise was reached, and the Grand Duke was requested to grant a constitution and to take part in the war against Austria. Leopold having rejected these demands, the Florentines rose and expelled him (April 27, 1859).

Union with the Italian kingdom. A provisional government, led by Ubaldino Peruzzi and afterward by Bettino Ricasoli, was established. It declared war against Austria and then handed over its authority to a Sardinian royal commissioner (May 9). A few weeks later a French force under Prince Napoleon ("Plon-Plon") landed in Tuscany to threaten Austria's flank, but meanwhile the emperor Napoleon III made peace with Austria and agreed to the restoration of Leopold and other Italian princes. Victor Emmanuel was obliged to recall the royal commissioners, but together with Cavour he secretly encouraged the provisional governments, and the constituent assembly of Tuscany voted for annexation to Sardinia. The King accepted the annexation (March 22, 1860). On February 18, 1861, the kingdom of Italy, comprising Tuscany, was proclaimed.

Storms and floods, chiefly of the Arno and Ombrone rivers, dealt a severe blow to Tuscan agriculture and inundated Florence and Grosseto in 1966. (L.Vi./Ed.)

Emilia-Romagna

PHYSICAL AND HUMAN GEOGRAPHY

The north-central region of Emilia-Romagna is bounded by Venetia and Lombardy on the north, Liguria on the west, Tuscany on the south, the Marches on the southeast, and the Adriatic Sea on the east. With an area of 8,542 square miles (22,124 square kilometres), the region embraces the provinces of Bologna, Ferrara, Forlì, Modena, Parma, Piacenza, Ravenna, and Reggio nell'Emilia.

The northern plain

The northern portion is a great plain from the Via Aemilia to the Po. Its highest point is not more than 200 feet (60 metres) above sea level, while along the east coast are lagoons near the mouths of the Po and those called the Valli di Comacchio to the south of them. Farther south is the plain around Ravenna (10 feet), which continues as far as Rimini where the mountains come down to the coast. The region was badly hit by the great floods of 1966.

Immediately to the southeast of the Via Aemilia the mountains begin to rise, culminating in the central chain of the Ligurian and Tuscan Apennines. The boundary follows the summits of the chain in the provinces of Parma, Reggio, and Modena, passing over the Monte Bue (5,840 feet) and the Monte Cimone (7,103 feet), while in the provinces of Bologna and Forlì it stays along the northeastern slopes of the chain.

With the exception of the Po the main rivers of Emilia-Romagna descend from this portion of the Apennines, the majority of them being tributaries of the Po. The Trebbia (which rises in the province of Genoa), Taro, Secchia, and Panaro are the most important. Even the Reno, Ronco, and Montone, which now flow directly into the Adriatic, were, in Roman times, tributaries of the Po, and the Savio and Rubicon (Uso) seem to be the only streams of any importance from these slopes of the Tuscan Apennines that ran directly into the sea in Roman times. A considerable amount of electric power is derived from these rivers, and the stations are connected with the Alpine plants so that interchange at different seasons is possible. Emilia-Romagna is one of the leading regions of Italian agriculture. A large part of it is lowland, with water available both in the form of precipitation and from irrigation systems. Wheat and corn are the chief cereals, and sugar beets, tobacco, and potatoes the industrial crops. Vegetables are grown in the lowlands, while the best wines come from the Apennine slopes. There are large numbers of beef and dairy cattle. Cheese, meat, and smoked meats of various kinds support a large food-packing industry, and there are flour and sugar milling and the preserving of vegetables and fruit. Among the industries the manufacture of cars and trucks, farm machinery, chemicals and pharmaceuticals, and clothing are particularly important. The extraction of large deposits of natural gas (at Cortemaggiore north of Piacenza and near Ravenna) and of oil (at Cortemaggiore) gives the region a vital role in the energy economy of Italy. The main rail artery is the Milan–Bologna–Rimini–Bari line. Bologna also has trunk railway lines connecting it with Verona and Venice to the north and Florence and Rome to the south. The region is well served by secondary railway lines and highways. Express highways connect Bologna with Milan and with Florence.

HISTORY

The name Emilia comes from the Via Aemilia, the Roman road from Ariminum (Rimini) to Placentia (Piacenza) that traversed the entire district from southeast to northwest, its line being closely followed by the modern railway. In popular usage the name was transferred to the district (which formed the eighth Augustan region of Italy) as early as the time of Martial, and in the 2nd and 3rd centuries AD Aemilia was frequently named as a district under imperial judges, generally in combination with Liguria or Tuscia. From the 3rd to the 5th century the district of Ravenna was, as a rule, not treated as part of Aemilia, the chief town of the latter being Placentia. In the 4th century Aemilia and Liguria were joined to form a consular province; after that Aemilia stood alone, Ravenna being sometimes temporarily added to it. The boundaries of the ancient district correspond approximately with those of the modern.

The chief cities of Emilia

In the 6th century AD Ravenna became the seat of a Byzantine exarch. After the Lombards had for two centuries attempted to subdue the Pentapolis (Rimini, Ancona, Fano, Pesaro, and Senigallia), Pepin took these cities from Aistulf and in 754 gave them, with the March of Ancona, to the papacy to which, under the name of Romagna, they continued to belong. The other chief cities of Emilia—Ferrara, Modena, Reggio nell'Emilia, Parma, Piacenza—were independent. Whether belonging to the Romagna or not, each had a history of its own and, notwithstanding the feuds of Guelphs and Ghibellines, they prospered considerably. Pope Nicholas III obtained control of the Romagna in 1278, but the papal dominion during the Avignon period was only maintained by the efforts of Cardinal Albornoz, a Spaniard sent to Italy by Innocent VI in 1353. Even so, the papal supremacy was little more than nominal, and this state of things only ceased when Cesare Borgia, the natural son of Alexander VI, crushed most of the petty princes of Romagna, intending to found there a dynasty of his own. On the death of Alexander VI, however, it was his successors in the papacy who profited by what Cesare Borgia had begun. The majority of the towns were thenceforth subject to the church and administered by cardinal legates. In 1796–1814 Emilia was first incorporated in the Italian republic and then in the Napoleonic Italian kingdom. After 1815 Romagna returned to the papacy and its ecclesiastical government. The duchy of Parma was given to Marie-Louise, wife of the deposed Napoleon, and Modena to the archduke Francis of Austria, the heir to the last Este. In Romagna and Modena the government was oppressive, arbitrary, corrupt, and unprogressive, but in Parma conditions were better. In 1821 and 1831 there were unsuccessful attempts at revolt in Emilia, which were sternly and cruelly repressed. Chronic discontent continued, and the people joined again in the movement of 1848–49, which was crushed by Austrian troops. In 1860 the struggle for independence was finally successful, Emilia passing to the Italian kingdom almost without resistance. The region's name was changed to Emilia-Romagna in 1948. (G.Kh.)

Papal States

PHYSICAL AND HUMAN GEOGRAPHY

The Papal States (States of the Church) were the lands over which the popes, as heads of the Roman Catholic Church, had sovereignty from 756 to 1870; in particular they were those in Italy, with which this section is concerned. Included were the modern regions of Lazio, Umbria, the Marches, and part of Emilia-Romagna, with boundaries that shifted over the years.

The land. The west-central region of Lazio (ancient Latium) is composed of the provinces of Roma (Rome), Frosinone, Latina, Rieti, and Viterbo. Its area is 6,642 square miles (17,203 square kilometres).

In the east Lazio is dominated by the central Apennines. In its northeasternmost section, near Rieti, the region includes part of the Abruzzi. Monte Terminillo (7,260 feet [2,213 metres]), the highest peak, is in the Reatini range; other Apennine ranges within Lazio are the Sabini, Simbruini, and Ernici. Limestone is the main component of the Apennines, and karst phenomena are found in a number of places. The lower foothills of the pre-Apennines are fertile; some of the valleys, such as those near Rieti and Subiaco, are among the best farming areas of the entire Apennine chain.

Regions

The western part of Lazio centres around the Campagna di Roma. To the northwest and southeast of it are four groups of ancient volcanoes, the Cimini, Volsini, Sabatini, and Alban mountains, each containing one or more crater lakes, those of Bracciano, Albano, Nemi, Bolsena, and Vico. The Campagna di Roma continues northwestward into the Maremma and southeastward with the Pontine marshes, while beyond Terracina lie the plains of Fondi and Formia. Southeast of the volcanic Alban hills, the Lepini, Ausoni, and Aurunci, stark, denuded mountains, extend to the Garigliano River, the southern limit of Lazio.

The central region of Umbria comprises the provinces of Perugia and Terni. Its area is 3,265 square miles. Umbria's core is the upper and middle valley of the Tiber, flanked on the west and east by low hills that, in the east, gradually rise to the Umbrian-Marchigian Apennines (Roman Apennines). The characteristic feature of the region's physiography is the prevalence of wide basins, some of lacustrine origin (Lago Trasimeno); others form sections of river valleys, such as the Umbrian Valley between Perugia and Spoleto and the Tiber Valley from Sansepolcro to Umbertide, or small depressions, such as the plains of Gubbio and Terni.

The central region of the Marches is composed of the provinces of Ancona, Ascoli Piceno, Macerata, and Pesaro-Urbino. Its area is 3,742 square miles. The Marches is a region of mountains and hills, the only pieces of level land being scattered along river valleys and on the Adriatic shore. Its mountain backbone is the Umbrian-Marchigian section of the Apennines; the administrative boundary between the Marches and neighbouring Umbria is the watershed between the Tyrrhenian and Adriatic slopes. The highest peak of the region is Monte Vettore (8,130 feet). Except for the northernmost part, the hills of Montefeltro, the Marches is characterized by rivers running at right angles to the main Apennine crests out to the Adriatic, separated by low parallel ridges. The most important of these rivers are the Metauro, the Foglia, the Esino, the Potenza, the Chienti, and the Tronto, the last named forming the boundary between the Marches and the Abruzzi. The upper valleys of these streams are narrow, often passing through deep gorges, while the lower sections are wider. The valley bottoms are thoroughly cultivated, and most of the lower slopes are either in meadows or in well-tended fields.

The economy. *Lazio.* Until the latter part of the 19th century, much of the lowland area of Lazio was marshy and malarial, and its agriculture was characterized by migratory grazing and growing of wheat in the lowlands and grapes on the hills. Major reclamation works in the Maremma, the Campagna di Roma, and the Pontine marshes during the first half of the 20th century transformed farming in Lazio completely. Migratory grazing was greatly reduced in importance. Wheat and corn, vegetables and fruit, meat and dairy products are dominant in the lowlands, while olive groves, orchards, and vineyards cover the slopes. Commercial vineyards are especially important around Montefiascone, on the Alban Hills, and around Terracina and Formia. Civitavecchia and Gaeta are the main fishing ports.

Industry plays a subordinate role in the economy of Lazio. Rome, the capital of the region and of Italy, is the major urban and industrial centre and the region's banking and commerical core. However, Rome's industries, with a few exceptions, are either of the artisan type (fashion) or highly specialized (motion pictures). Rome, including the Vatican city-state, is Italy's largest tourist centre, and tourism is one of the city's largest employers. Large numbers of persons are also employed by the government. In the remainder of the region chemical and pharmaceutical plants, food industries, and a few small machine industries are the only ones of importance.

The transportation pattern of Lazio is dominated by Rome. The city is connected by trunk railroads with Genoa, Florence, Ancona, Pescara, and Naples, and it is the centre of the highway system of central Italy. Construction began on a high-speed direct Florence–Rome railway. After World War II Rome also became one of Europe's busiest airports, through which many trunk lines pass to the Far and Middle East and to Africa. Civitavecchia, the only port of importance, is noted chiefly for its trade with Sardinia.

Umbria. Farming in the hills and valleys is prosperous, characterized by intensive land use, especially intercropping. The principal cereals are wheat and corn, the latter supporting a sizable number of pigs. Potatoes, sugar beets, grapes, and olives are grown, and the wine of Orvieto is well known throughout Italy. The major power centre of Umbria is the hydroelectric complex of Terni, which supports the steel, chemical, and electrochemical industries at Terni, Narni, and Foligno. Textiles and food industries at Perugia are important. The main Rome–Florence railway passes through the southernmost part of the region, with a branch through Terni to Foligno and Perugia, joining the main line south of Cortona. The other main line (Rome–Ancona) passes through Terni and Foligno to Fabriano and Ancona. There is an excellent system of highways and bus communications.

The Marches. The economy of the Marches is predominantly agricultural. Wheat and corn are the main cereals, and there are vineyards on all the sunny slopes, a white wine produced in the area being especially popular. The local type of cattle, beasts of burden rather than meat or milk producers, are known throughout Italy, and there are large numbers of horses, pigs, sheep, and poultry. Fishing is important in several of the Adriatic ports, especially in Ancona and San Benedetto del Tronto. Industries are mostly small or medium sized and include shipbuilding in Ancona, paper in Fabriano, textiles in Iesi, musical instruments in Castelfidardo, and pottery in Pesaro and Recanati. Ancona is the largest city and the only one with a natural harbour; Pesaro is the only other urban centre of any size on the seacoast. The other cities are hilltop settlements, like Urbino, Macerata, and Ascoli Piceno, or valley centres, like Iesi and Fabriano. The main artery of northwest–southeast travel is the coast railroad, from Bologna through Rimini to Pesaro and Ancona and on to Foggia and Bari. Ancona also has a direct rail line to Rome.

For Romagna, see the section *Emilia-Romagna* in this article.

HISTORY

From the 4th century AD onward the Roman Catholic Church was the recognized proprietor of extensive estates throughout and even beyond Italy, but it held these *patrimonia* in the manner of a landowning corporation, under the Roman Empire, not in that of a sovereign ruler. By the middle of the 8th century, however, the Lombards had overrun most of Italy. The duchy of Rome was then still theoretically dependent on the Byzantine, or East Roman, Empire; but the Byzantines could not protect the duchy, within which the bishops of Rome, supported by their clergy, exercised an authority counterbalancing that of the local barons and their army. When Pope Stephen II (III) appealed for help against the Lombards to the Frankish ruler Pepin III the Short, Pepin in 754 made the famous and controversial Donation, whereby he undertook to "restore" to the Roman Church and the "Republic of the Romans" numerous lands of which they had been despoiled, regardless of the fact that these lands ought juridically to have been restored to the Exarchate of Ravenna.

Pepin's "restoration" of lands

The Lombard king Aistulf, by the Treaty of Pavia in 756, ceded (1) Comacchio and Ravenna; (2) the country from Forlì to Senigallia and Jesi between the Apennines and the Adriatic; and (3) Gubbio in the Apennines, linking the northeastern territories with the duchy of Rome by way of Perugia. This cession, which comprised less than what the Donation of Pepin had adumbrated, was the beginning of the temporal power of the papacy. Bologna, Ferrara, Imola, Faenza, and Ancona were added to the dominion soon afterward; but Charlemagne, whose Donation of 774 had promised far more, *c.* 781 limited the further acquisitions of the Holy See to the Sabina (Rieti and its vicinity, on the frontier of the duchy of Spoleto), to some Campanian cities (soon lost), and to some parts of the formerly Lombard Tuscia, including Orvieto and Viterbo.

With the breakup of the Carolingian Empire an era of vicissitudes began. The history of the Papal States is thereafter for centuries practically inextricable from that of Italy and of the papacy, or from that of Rome and of the other cities of the dominion. The extent over which papal authority was effective shrank or grew with the Holy See's prestige. The temporal power suffered from the growth of feudalism, from disorders in Rome itself, and from the domination of the Saxon dynasty after the revival of the Holy Roman Empire for Otto I in the 10th century; but it reasserted itself in the second half of the 12th century thanks chiefly to the alliance with the Norman conquerors of southern Italy and to the genius of Pope Gregory VII. The duchy of Benevento was recognized as papal in 1052 and definitely acquired in 1077; and the countess Matilda of Tuscany bequeathed her great inheritance to the Holy See. Pope Innocent III took great advantage of the dispute between the Hohenstaufen and their rival Otto IV for the imperial crown to promote his claims, notably in the March of Ancona; and Otto in 1201 acknowledged the church's right to the duchy of Spoleto.

The rise of the communes and the subsequent emergence of the *signorie* weakened papal authority, especially in the Romagna. The translation of the papacy to Avignon (1309) left the dominion in Italy to chaos; and the brilliant work of reconquest and rehabilitation carried out by Cardinal Albornoz in the 1350s and '60s was undone in the '70s by the War of the Eight Saints and by the beginning of the Great Schism.

At the end of the Great Schism in 1449 the Romagna, the Marches, and Umbria were still mostly in the lands of signorial houses exercising "vicariates" granted to them by the Holy See but in fact ruling as they saw fit. It was to subdue these places that Pope Alexander VI launched his son Cesare Borgia on his expeditions. Much of Borgia's conquests, however, fell away on Alexander's death (1503), and the restoration of the Papal State had to be undertaken again by popes Julius II and Leo X in the period 1510–21: they also won Modena and Parma and Piacenza for the church. Modena, however, was recovered in 1527 by the House of Este; and Parma and Piacenza were granted in 1545 to the House of Farnese, to which moreover the ancient papal territory of Castro had been given as a duchy in 1537. Efforts to recover Ferrara from the Estensi were finally successful in 1598, and Urbino returned to direct papal rule in 1626. The attempt of the Barberini pope, Urban VIII, to take Castro back by force (1641–44) was frustrated, but the duchy was reannexed to the papal state in 1649.

In the 18th century, though they were traversed by foreign armies in the course of dynastic wars, the Papal States enjoyed a period of prosperity under paternalistic government. The French Revolutionary Wars and the Napoleonic Wars changed everything: Bologna, Ferrara, and the Romagna were ceded by the Treaty of Tolentino (1797), under which the French also occupied the Marches and Umbria; Rome was a republic from February to November 1799; and after an interval in which the lands south of the Romagna returned to papal rule, the Marches were annexed to the Napoleonic Kingdom of Italy in 1808 and the remnant of the Papal State to the French Empire in 1809.

Restoration of the Papal States

The Congress of Vienna in 1815 restored the Papal States; but the liberalizing influence of Cardinal Consalvi, which had culminated in the Statute of 1816, was largely counteracted in Pope Leo XII's pontificate. Administratively, the state was divided between (1) Rome and its Comarca, or district, under a special regime; (2) the Legations, or *Legazioni,* under a cardinal legate or a vice-legate; and (3) the Delegations, under prelates.

The Italian Risorgimento gradually destroyed the temporal power. Austrian intervention against a revolt in the northern Legations (1831–32) was followed by the French occupation of Ancona until 1838. The conduct of Pope Pius IX in 1848 led to the proclamation of the short-lived Roman Republic in 1849. Thenceforward the temporal power depended on Austrian or French protection. Through the defeat of Austria in 1859 and the Battle of Castelfidardo in 1860 the Romagna and the Marches, with Perugia, Spoleto, Orvieto, and Rieti, were annexed to the Kingdom of Sardinia-Piedmont, which in 1861 became the Kingdom of Italy. Garibaldi's attack on the remnant of the Papal States in 1867 was defeated at Mentana; but in 1870 the final annexation to Italy was achieved. The Lateran Treaty of 1929 recreated a temporal power in the Vatican City State.

(Ed.)

Mezzogiorno

The Mezzogiorno, or southern, region of Italy consists roughly of the area south of the Garigliano River, plus the islands of Sicily and Sardinia, and corresponds approximately to the former Kingdom of Naples. Because of its economic backwardness, particularly in the 19th and 20th centuries, the Mezzogiorno grew into a major social and political problem for Italy. Thus the Cassa per il Mezzogiorno, or the Italy Development Fund, was initiated in order to alleviate some of the strains experienced by the south.

PHYSICAL AND HUMAN GEOGRAPHY

The Mezzogiorno includes the regions of Campania, Puglia, Basilicata, Lucania, and Calabria, as well as Sicily and Sardinia.

The land. *Campania.* The region of Campania comprises the provinces of Avellino, Benevento, Caserta, Naples, and Salerno. Campania extends from the Garigliano (the lower Liri River, ancient Liris) in the north to the Gulf of Policastro in the south, facing the Tyrrhenian Sea; on the inland side it is bordered by Latium (Lazio), Abruzzi, Molise, Puglia, and Basilicata. Its physiography is dominated by volcanic and seismic phenomena that characterize the area surrounding the Gulf of Naples (Campi Flegrei, or Phlegraean Fields; Vesuvius). Around these areas of volcanic activity extend the principal lowlands of Campania; the Volturno lowland; the plains called Terra di Lavoro from the middle Volturno (ancient Volturnus) River to the eastern flanks of Mt. Vesuvius; while the only other lowland of any size, the plain of the lower Sele River (ancient Silarus), is separated from the others by the Lattari Mountains between Pompeii and Salerno. To the east are the complex uplands of the region, part of the Apennine system, the Matese, the Picentini, and the Cilento mountains, and beyond these, the Neapolitan Apennines (Appennino Napoletano). The only rivers of any size are the Volturno in the northern and the Sele in the southern part of the region. Among the intermontane basins that of Benevento is the most important. While communications between the coastal areas of Campania are relatively easy, the highly dissected character of the interior makes rail and road travel "across the grain," in the west–east direction, very difficult.

Puglia. The region of Puglia (Apulia) extends from the Fortore River in the northwest to Cape Santa Maria di Leuca, the "heel" of the peninsula, in the southeast. It is composed of the provinces of Bari, Brindisi, Foggia, Lecce, and Taranto. The northern third of the region is centred on the Foggia Plain, or Tavoliere, flanked by the Gargano massif in the north and the Neapolitan Apennines in the west. The central third is occupied by the low plateau of the Murge, limited in the west by a depression, the "fossa premurgiana," while in the east it slopes gradually to the narrow coastal plains of the Adriatic. The southern third, southeast of the Taranto-Ostuni line, is the Salentine Peninsula, consisting of the lowland of Lecce and the low plateaus east of Taranto and south of Lecce. The predominant rock material of Puglia is limestone, and karst phenomena of underground drainage and large cave formations are present in many areas.

The coastline for the most part is low and sandy, except in the Gargano Peninsula and in the southeasternmost tip of Puglia. The only rivers of significance are the Fortore and the Ofanto, but there are numerous springs, some under the sea near the coastline. Absence of surface water over large areas led to construction of the Apulian Aqueduct, largest of its kind in Italy, which traverses the region as far as Cape Leuca.

Basilicata. The region of Basilicata roughly corresponds to the ancient region of Lucania (see below) and comprises the provinces of Potenza and Matera. It falls into a western mountainous section, dominated by the Lucanian Apennines, and an eastern section of low hills and wide valleys, while along the Ionian sea the sand and clay hills overlook narrow coastal plains. The extinct volcano of Mt. Vulture stands isolated from the Apennines in the north.

Lucania. Lucania is an ancient territorial division of southern Italy, corresponding to most of the modern region of Basilicata with much of the province of Salerno and part of that of Cosenza. Its boundaries were, approximately, the Silarus (Sele) River on the northwest, the Bradanus (Bradano) on the northeast, and the Crathis (Crati) and Laus (Lao) rivers on the south; Eburum (Eboli) and Volceii (Buccino) beyond the Silarus and Bantia (Banzi) beyond the Bradanus were, however, also included in Lucania. Apart from the east coast and the Silarus Valley, the whole of Lucania was occupied by the Lucanian Apennines.

Calabria. The region of Calabria comprises the provinces of Catanzaro, Cosenza, and Reggio di Calabria. Sometimes referred to as "the toe of the Italian boot," Calabria is a peninsula of irregular shape, jutting out in a northeast–southwest direction from the main body of Italy and separating the Tyrrhenian and Ionian seas. Most of the region is mountainous or hilly, the only extensive lowlands being those of the lower Crati Valley near Sibari (which derived its name from ancient Sybaris), of the Marchesato near Crotone, of Sant'Eufemia, and of Gioia Tauro. In the north Calabria is linked to the Lucanian Apennines (Appennino Lucano) by the massif of Monte Pollino; the Pollino is continued southward along the west coast by the coast range, which in turn is separated by the Crati River from the extensive La Sila massif. A narrow isthmus between the Gulf of Sant'Eufemia in the west and the Gulf of Squillace in the east separates the northern from the southern part of the region; the uplands continue as the Calabrian Apennines (Appennino Calabrese) and culminate in the southernmost part as the massif of the Aspromonte.

Agriculture

The economy. The economy of the Mezzogiorno is based largely on agriculture, although many areas produce low yields due to poor soil and lack of labour-saving machinery. Principle crops include cereals, fruit, olives, grapes, and tobacco; wine production and fishing are also important.

Campania. The most important farming areas of Campania are the lowlands of the Terra di Lavoro and of the circum-vesuvian plain. The land is fertile, and utilization is extremely intensive, characterized by interculture, with plots of land producing cereals on the ground, fruit on trees along the edges of the plots, and grapes from vines trailing between trees. Farms are usually small, and human labour is used for most farming operations. The most important crops are fruit (apricots, apples, peaches, nuts, citrus, and grapes), early vegetables and flowers, and such industrial crops as tobacco and hemp. Wines of Campania, especially those from Vesuvius (Lacrima Christi), from Ischia (Epomeo), and from the Sorrento peninsula, are famous throughout Italy. Fishing is important in the Bay of Naples, Procida and Torre del Greco being the leading ports. Metallurgy, chemicals, machinery and tools, textiles, agricultural industries (canning, flour milling, tobacco), and shipbuilding are the most important industries. In Naples and its suburbs there is a flourishing artisan industry working coral, pearls, tortoise shell, leather, and lace. The tourist trade in Naples, on the Sorrento peninsula, and on the islands of Capri and Ischia, is an important source of income. Naples is a leading Italian port. The transportation system is also centred on Naples. Main lines connect the city with Rome (via Formia and via Cassino), with Benevento, Foggia, and the Adriatic coast, with Potenza and Taranto, and with Reggio di Calabria and Sicily.

Puglia. Wheat and oats are the principal cereals, raised in the Foggia Plain and in the more fertile parts of the plateaus; olives, grapes, almonds, and figs are grown intensively in the coastal and some inland areas; tobacco is a specialty of the Lecce Plain. The wines of Puglia are among the strongest of Italy and are used to fortify other, lighter varieties. Fishing is carried out in many ports; those of the Gargano, of Barletta, of Monopoli, and of Taranto are the most important. Salt is produced from seawater at Margherita di Savoia, near Foggia. Bari is the largest city and the leading port, as well as the biggest industrial centre (especially chemicals and petrochemicals). The largest railroad centre is Foggia, with lines connecting it to Naples–Rome, Bologna–Milan, and Bari–Taranto–Brindisi–Lecce. The so-called "Ionian" railroad follows the Ionian coast from Taranto to Reggio di Calabria. After World War II Puglia became one of the principal areas of Italian land reform.

Basilicata. The mainstay of the economy is agriculture; most of it, however, is characterized by low yields. Wheat and rye are the principal cereals; sheep, goats, and pigs the farm animals. New crops introduced in the eastern and coastal areas include tobacco, vegetables, sugar beets, and flowers. Dairy and beef cattle are also raised. Except for olive presses and flour mills, industry in Basilicata was very late in developing. Potenza, the administrative centre, and Matera are the principal cities. The main railroad line of the Basilicata links Naples and Battipaglia in the northwest through Potenza with Taranto to the east, connecting also with the Ionian railroad (Taranto–Reggio di Calabria). Local lines radiate from Potenza and from Matera in several directions, and bus services reach all of the villages of the region.

Calabria. The mainstay of Calabria's economy is farming, once characterized by large landed estates and tiny peasant holdings. Under the Italian land reform the majority of the former latifundia were broken up and new, small peasant holdings created, with rural service centres, new houses, and new roads. Formerly the agriculture of Calabria concentrated almost entirely on cereals and the raising of sheep and goats, with occasional work in the forests of the Sila uplands. New commercial crops, citrus fruit (mostly on the west coast), figs, and chestnuts, were introduced. Hydroelectric power was developed during the 1920s and 1930s in La Sila and is an important feature of the Calabrian economy, supplying electric railways and the industrial centre of Crotone on the Ionian coast, which has chemical industries. The highways are well developed, with extensive bus services. A railway and car ferry links the ports of Reggio di Calabria and Villa San Giovanni with Messina in Sicily.

Calabria is one of the few areas of southern Italy with a non-Italian minority: the Albanians who settled in the region during the 15th and 16th centuries under Turkish pressure, and who have retained their speech, the Greek Catholic rite in their churches, and, on festival occasions, their colourful national costumes.

HISTORY

Ancient history. *Campania.* Ancient Campania was smaller than the modern region and roughly extended over the area now comprising the provinces of Naples, western Caserta, and northern Salerno. It was bounded on the northwest by the territory of the Aurunci, on the northeast by Samnium, on the south by the Sorrento peninsula, and on the west by the Tyrrhenian Sea. By the 1st century AD the northwestern boundary extended as far as Sinuessa (near Mondragone) and the Volturnus (Volturno) River, and the northern boundary came between Venafrum (modern Venafro) and Casinum (Cassino); the Volturnus valley and foothills of the Apennines as far as Abellinum (Avellino) formed the boundary on the northeast. The southern boundary remained unchanged.

Campani was the Roman name for inhabitants first of the town of Capua (modern Santa Maria Capua Vetere) and its district and then for inhabitants of the Campanian plain. The name is pre-Roman and appears with Oscan terminations on coins of the early 4th or late 5th century BC struck for or by the Samnites, the conquerors of the Etruscans in Campania at the end of the 5th century. Cumae was taken in 428 or 421, Nola about the same time, and the local dialect, henceforth known as Oscan, spread over all Campania except for the Greek cities, although Etruscans remained for at least another century. Latin became general soon after the Social War (90–89 BC) except in Neapolis (Naples), where Greek was the official language during the Roman Empire. The Samnites took over many Etruscan customs; the haughtiness and luxury of the men of Capua were proverbial at Rome. This town became the ally of Rome in 338 BC. By the end of the 4th century Campania was completely Romanized and was granted a limited form of Roman citizenship (*civitas sine suffragio*). Certain towns with their territories (Neapolis, Nola, Abella [Avella], Nuceria Alfaterna [Nocera Inferiore]) were nominally independent in alliance with Rome. These towns were faithful to Rome throughout the war against Hannibal (218–202 BC), but Capua and its satellite towns revolted. After its capture by the Romans in 211 the people of Capua were severely punished and their land confiscated. During the Roman Empire, however, it flourished as a *colonia*. In the division of Italy into *regiones* by the emperor Augustus, Campania with Latium formed the first region. From *c.* AD 285 the name Campania was extended northward to include the whole of Latium. This district was governed by a *corrector* who received the title of *consularis c.* AD 333.

Spread of the Latin language

Puteoli (modern Pozzuoli), the chief ancient harbour, was most important in the 2nd–1st century BC. The road system of Campania was extremely well developed, the most important road centre being Capua. The Appian Way met the via Latina at Casilinum, three miles to the northwest.

After the fall of Rome the region was occupied successively by the Goths and Byzantines. The Normans conquered it in the 11th century, and it was incorporated in the kingdom of Sicily in the 12th century.

Puglia. The southeastern extremity of Puglia was Roman (not to be confused with modern) Calabria, referring from the 3rd century BC to the district in the southeastern extremity of the peninsula between the Adriatic and the Gulf of Tarentum ending in the Iapygian promontory (Salentina). Calabria occupied the southern part of modern Puglia (Apulia), consisting of the provinces of Lecce, Brindisi, and Taranto, though the latter extends farther westward than the ancient district (modern Calabria comprises the ancient territory of the Bruttii, the southwestern extremity of Italy).

Between 272 and 266 BC six triumphs were recorded in the Roman *fasti* (calendar) over the Tarentini, Sallentini, and Mesapii, while the name Calabria does not occur; but after the foundation of a colony at Brundisium about 246 and the final subjection of Tarentum in 209, Calabria became the general name for the peninsula. According to Strabo (1st century AD) earlier Calabria had been extremely prosperous and had had 13 cities, but all except Tarentum and Brundisium had dwindled to villages. The Appian Way, extended to Brundisium probably by 244 BC, passed through Tarentum; the shorter route by Canusium (Canosa di Puglia), Barium (Bari), and Egnatia (near Fasano) was only made a main route by the emperor Trajan. When the emperor Augustus divided Italy into *regiones* he joined Calabria to Apulia and the territory of the Hirpini to form the second region. From the end of the 2nd century Calabria was associated for juridical purposes either with Apulia or with Lucania and the district of the Bruttii, while the emperor Diocletian placed it under one *corrector* (governor) with Apulia. When the Lombards seized Calabria about AD 668 its name became transferred to the southwestern peninsula of Italy. After the tumultuous times following the disintegration of the Roman Empire, Puglia was ruled by the Byzantines for more than two centuries and came to know its greatest glory under Hohenstaufen emperors.

Basilicata. During medieval times Basilicata was first under Lombard rule, controlled by the duke of Benevento and, later, of Salerno; after an interval of Byzantine control, the Normans took over and made Melfi the capital of one of their important dominions. Until the fall of the Hohenstaufen, Basilicata played a significant part in the affairs of southern Italy.

Lucania. Lucania was so called from the Lucanians (Lucani) who conquered it about the middle of the 5th century BC. Before that it was included under the general name of Oenotria, applied by the Greeks to southernmost Italy. The mountainous interior was occupied by Oenotrians and Chones, while on the coasts on both sides were powerful Greek colonies that doubtless exercised a protectorate over the interior. The Lucanians were a southern branch of the Samnite group, who spoke Oscan. After much intertribal conflict they began to attack the Greek cities, especially Tarentum, which appealed first to Archidamus III of Sparta (killed 338), then to Alexander, king of Epirus (killed 330). In 298 BC the Lucanians made alliance with Rome, and Roman influence was extended by the colonies at Venusia (Venosa; 291 BC), Paestum (273), and above all Tarentum (272). On the landing of Pyrrhus in Italy (281 BC) they were among the first to declare in his favour and found themselves exposed to the resentment of Rome when the departure of Pyrrhus left his allies at the mercy of the Romans. After several campaigns they were reduced to subjection (272 BC). They sided with Hannibal during the Second Punic War (216 BC), and their territory during several campaigns was ravaged by both armies. The country never recovered from these disasters and under the Roman government fell into decay, to which the Social War, in which the Lucanians took part with the Samnites against Rome (90 BC onward), gave the finishing stroke. For administrative purposes under the Roman Empire, Lucania was always united with the district of the Bruttii to the south. The two together constituted the third region of Augustus' reorganized Italy.

Calabria. In classical times the region was a centre of Greek colonization; Crotona, Sybaris, and Rhegion (now Reggio di Calabria) were Greek cities of wide fame. After the Roman conquest the splendour of the Greek cities slowly gave way to a remote provincial existence, and eventually *Ager Bruttius* as it was then called passed to the Byzantines, who applied the name Calabria (which was also the Roman name for the southeast extremity of the Italian peninsula). The Lombards controlled the region from Benevento and, later, from Salerno. After another period of Byzantine rule Calabria shared with the rest of southern Italy its Hohenstaufen, Angevin, and Aragonian rulers.

Kingdom of Naples. The term Kingdom of Naples has been conventionally given, since the end of the 14th century, to the kingdom of Sicily *citra Pharum* (*i.e.*, "on the hither side of the Strait of Messina"); in other words, to the mainland part of the greater medieval kingdom of Sicily, from which the island had been detached to form the separate kingdom of Sicily *ultra Pharum* ("beyond the Strait"). The mainland territory was bounded in the north by the Papal States, with a roughly S-shaped frontier running from the mouth of the Tronto River on the Adriatic coast to a point just east of Terracina on the Tyrrhenian coast, so that Abruzzi, Molise, and the Terra di Lavoro were its northernmost provinces.

The Normans and the Hohenstaufens. The Byzantines, Lombards, Arabs, and indigenous lords who had long been rivals for control over southern Italy and Sicily were displaced in the 11th century by bands of Normans from France. Among these, two brothers were preeminent: Robert Guiscard and Roger I. The papacy, which in 1059 had accepted Robert Guiscard as its vassal for his past and future conquests, looked to the Norman power in the south to prevent the German kings or Holy Roman emperors from extending their dominion over the whole

Roger II, king of Sicily

of Italy. In 1130 Roger I's son Roger II, having united all the Norman acquisitions, assumed the title king of Sicily. An energetic ruler with a splendid capital at Palermo, he welded his heterogeneous subjects—Catholic Italians and Normans, Orthodox Greeks, and Muslim Arabs—into a strong state.

Roger II died in 1154, his son and successor William I in 1166, and the latter's son William II in 1189. The heiress of the Sicilian kingdom was then Roger's daughter Constance, who had in 1186 been married to the German king Henry VI, son of Frederick I Barbarossa, of the Hohenstaufen dynasty. The consequent accession of Henry to Sicily was resisted by a national party, which set up first Tancred of Lecce as king, then Tancred's son Roger as joint king with him; but they were defeated by the Germans in 1194. Installed as king at Palermo, Henry now linked the Sicilian crown to the imperial (his since 1191).

The union of the crowns, dissolved by Henry's death (1197), was reconstituted by his son. Acknowledged in Sicily as king in 1198 (when Constance died), Frederick received the Sicilian crown from Pope Innocent III in 1209 and became German king in 1212 and emperor in 1220. His subsequent disregard of Sicily's feudal dependence on the Holy See and his attempt to override the autonomy of the Italian communes and of the feudatories provoked a formidable conflict. For the Kingdom of Sicily itself, his Constitutions of Melfi (1231) and his changes of the Norman dispositions there reflected his absolutist attitude. Though he patronized the arts and developed the economy, the kingdom had to contribute disproportionately to the expenses of his wars. When he died (1250), the Hohenstaufen succession had already been called in question by his second excommunication (1245).

Frederick's heir, the German king Conrad IV, died in 1254. Then Manfred, a bastard son of Frederick II, assumed the regency for Conrad's son Conradin. In 1258, on a false report of Conradin's death, Manfred had himself crowned king of Sicily, which he then tried to make independent of pope and emperor alike. The papacy, meanwhile, had been offering the Sicilian crown to others: Henry III of England accepted it in 1254 for his son Edmund, but this scheme came to nothing. At last, in 1265, Charles of Anjou, count of Provence and brother of Louis IX of France, accepted the crown as a vassal of the papacy. At the Battle of Benevento (1266) Manfred was defeated and killed.

The Angevins. So as to be nearer to his Guelph allies of northern and central Italy and to Rome, Charles made Naples his capital instead of Palermo. A revolt linked with Conradin's attempt to assert his claims (1268) was savagely put down. Charles, however, had further ambitions—in Italy, in the Balkans, and in the Arab Mediterranean—and subjected his kingdom to oppressive taxation. Discontent was exploited by exiles abroad, in particular the physician Giovanni di Procida and Ruggiero de Lauria, at the court of Peter III of Aragon, who had married Manfred's daughter Constance. When the revolt known as the Sicilian Vespers broke out at Palermo in 1282, Charles was taken by surprise. Rebellion spread over all Sicily and into Calabria.

Peter III's landing in Sicily began a war in which the Ghibelline states of Italy took the Aragonese side while the papacy and France took the Angevin. The Aragonese navy had several successes; *e.g.,* in the battle of Naples in 1284, in which Charles I's son, the future Charles II, was taken prisoner. Charles I died in 1285, and Charles II had to accept humiliating terms before being restored to Naples. Yet he finally expelled the Aragonese from the mainland; and at the Peace of Caltabellotta (1302) it was agreed that the island should revert to Angevin rule on the death of Peter III's son, who took the misleading style of Frederick III. Hostilities, however, were resumed, and warfare punctuated by truces went on until Joan I of Naples in the 1370s waived her rights to Sicily and accepted Frederick's grandson, the rightly numbered Frederick III, as her vassal.

Peace of Caltabellotta (1302)

When the Sicilian crisis arose, Charles I issued the *capitoli,* or Articles of San Martino (1283), which extended the jurisdiction and immunities of barons and clergy and strengthened feudalism. The capital city, moreover, with its administration monopolized by an urban aristocracy and with its swollen population largely destitute and often workless, became another privileged entity whose particular interests were allowed to overshadow those of the nation as a whole. While the parliament remained under the control of the feudatories, municipal life stagnated, since there was no way for an active middle class to make its influence felt. Poor in natural resources, the country based its economy on agriculture and livestock. Foreigners—mostly Florentines, Venetians, Genoese, Catalans, and Marseillais—competed between themselves for what trade there was and dominated the markets.

Conditions improved under Charles II's son Robert, who succeeded to the crown in 1309. Against the emperor Henry VII he took the Guelph side. Successfully defended by his troops, the Guelph communes in Tuscany finally put themselves under his protection; and the Avignon pope, John XXII, later made Robert his vicar for the Papal States. A capable ruler, Robert enhanced the prestige of his capital at home and abroad.

Robert died in 1343, leaving the crown to his granddaughter Joan I. Two years later dynastic strife broke loose. The Angevin princes of the branch of Taranto (sons of Robert's brother Philip) procured the murder of Joan's husband Andrew, an Angevin of the Hungarian branch (grandson of Robert's brother Charles Martel). Suspected of complicity in the murder, Joan fled to Avignon. Louis I of Hungary, Andrew's brother, twice tried to win the kingdom for himself (1348 and 1350–52), but the brutality of the Hungarians provoked a reaction in Joan's favour. She and her second husband, Louis of Taranto, were recognized as sovereigns of the kingdom in 1352.

Avid for power, Louis of Taranto quarrelled with his queen. The tranquillity that the grand seneschal Niccolò Acciaiuoli had worked to promote was shattered. When the schism of the papacy began (1378), Joan supported the antipope Clement VII. Pope Urban VI therefore declared the crown forfeit and offered it to another Neapolitan Angevin, Charles of Durazzo. Joan turned for help to Louis of Anjou-Provence (a brother of the French king Charles V), but Charles of Durazzo took Naples, imprisoned Joan, and had her put to death (1382).

Anarchy followed Charles's death (1386), the kingdom being alternately ravaged by the mercenaries of rival kings and embroiled in the conflicts of other Italian states. There was an interval of resurgence while Charles III's son Ladislas was king, from 1400 to 1414. An able soldier with no political scruples, Ladislas subdued the great feudatories, expelled his rival Louis II, and, by exploiting the Italian situation, made himself master of the Papal States, from which he invaded Tuscany. Premature death cut short his plans. His sister and successor Joan II, who reigned until 1435, was dominated by condottieri, favourites, and barons one after another. She first adopted Alfonso V of Aragon as her heir (Alfonso I of Naples), then put up Louis III of Anjou-Provence against him. The court could get no obedience from the provinces, over which the Angevin and Aragonese factions fought ruinously. Finally the Aragonese took Naples in 1442, driving René I, Louis III's brother and titular successor, into exile.

The Aragonese. Alfonso, whose predecessors had from 1409 united the crowns of Aragon and Sicily, fulfilled an ancient Aragonese aspiration when he conquered Naples; but in 1443 he decided that, whereas Aragon and Sicily would pass to his brother John, the Neapolitan succession should go to his bastard son Don Ferrante, or Ferdinand I. The Neapolitan barons helped him to win the approbation of Pope Eugenius IV, who had hitherto been against him. Alfonso's policy of war and expenditure in Italy was designed to strengthen his son's position and to open up new vistas for Aragonese diplomacy and commerce. The result, however, was to aggravate the disequilibrium in the kingdom.

Succeeding Alfonso in 1458, Ferdinand I was soon faced with a baronial revolt in favour of René or the latter's son John of Calabria. Centralization and modernization antagonized the barons; and they revolted again in 1485–87, with the intention, this time, of putting his younger son Frederick (born in 1451) on the throne. Ferdinand prevailed by trickery and ruthlessness.

Invasion of Italy by Charles VIII

Meanwhile Charles VIII of France had inherited the Angevin claims to Naples. In 1494 he invaded Italy to conquer the kingdom. Alfonso II of Naples, Ferdinand I's elder son and successor, abdicated in favour of his son Ferdinand II, or Ferrandino, in 1495; but the kingdom, torn by faction, could not withstand the French, and Charles was crowned king. The formation of the League of Venice obliged Charles to withdraw from Naples, and Ferdinand II was restored to his throne a few months before his death (1496). The crown then passed to Frederick (Alfonso II's brother), under whom the kingdom enjoyed a few years of security. Finally, however, Louis XII of France and Ferdinand II of Aragon, by the Treaty of Granada (1500), agreed to divide the kingdom between themselves, on the pretext of forestalling a Turkish occupation. Frederick gave himself up to Louis XII in 1501 and was pensioned off in France, where he died in 1504. The French and the Spaniards soon came to blows over the partition. Gonzalo de Córdoba's victories at Cerignola and on the Garigliano River (1503) decided the issue in Spain's favour.

The Spanish viceroyalty. The Spanish kings ruled Naples through viceroys. After the failure of the last French effort to recover the Angevin inheritance (1528, when the siege of Naples had to be abandoned), the major defensive concern of the viceroys was against the Turks and the Barbary pirates. Internally, they applied themselves to breaking the political power of disloyal feudatories. Pedro de Toledo, viceroy from 1532 to 1553 for the emperor Charles V, was especially successful, but his attempt to reinforce absolutism by introducing the Spanish Inquisition was foiled by a vigorous opposition.

Pedro de Toledo and his successors enhanced the importance of the capital, whose population grew to excess. While the aristocracy was still predominant in the *seggi*, or municipal administration of Naples, a popular element was brought in, so that the system became rather more democratic. The parliament of the kingdom, whose sole function was to vote supplies of money, was last convened in 1642; thereafter the viceroys treated the *seggi,* on whose loyalty they relied, as representatives of the kingdom.

The rising of Masaniello in July 1647 was provoked by taxation and by the high cost of living but was really organized and directed by Giulio Genoino, a lawyer who wanted the people's voice in the administration to be made equal to the nobility's. It failed, but insurrection broke out again: while the provinces rose against baronial oppression, the city of Naples, turning at last against Spanish rule, proclaimed itself a republic. As neither the papacy, nor Savoy, nor France could undertake its effective protection, the republic had to offer sovereignty to the Duc de Guise (Henry de Lorraine), but the Spaniards took prompt countermeasures, supported by those who wanted to see order restored. By spring 1648 the rebellion was crushed.

The second half of the 17th century was a period of stagnation for the kingdom. The last lights of Renaissance culture in southern Italy had gone out with Tommaso Campanella and with the economist Antonio Serra (*Breve trattato. . . ,* 1613). Society was paralyzed by the rivalries of pressure groups and by corruption; and the Spanish monarchy, which drained the country of money and soldiers to prop its power in Europe, made matters worse by its administrative experiments. There was, however, a slow reawakening of intellectual life as the currents of European thought touched Naples, Cartesianism being the most vital factor.

The Spanish succession and the Austrian interlude. When the War of the Spanish Succession was breaking out, a group of nobles, led by Gaetano Gambacorta, principe di Macchia, in reaction against the absolutism of Spain and its middle-class supporters, plotted to offer the crown to an Austrian Habsburg (1701). This conspiracy came to nothing; but the Austrians overran the kingdom in 1707, and by the peace treaties of Utrecht and Rastatt (1713–14) it was ceded to the emperor Charles VI. Sicily was at the same time ceded to Victor Amadeus II of Savoy; but a few years later, in accordance with the Quadruple Alliance of 1718, it was transferred to Austria likewise, in exchange for Sardinia. The Austrians undertook some measures of rehabilitation and were to some extent open to the influence of the Neapolitan intelligentsia, whose great spokesmen now were Giambattista Vico and Pietro Giannone.

Don Carlos de Borbón

The first Bourbon period. On the outbreak of the War of the Polish Succession in 1733, France, Spain, and Sardinia-Savoy agreed that the Spanish infante Don Carlos de Borbón should have Naples and Sicily if they could be conquered. Don Carlos defeated the Austrians in the Battle of Bitonto (1734) and was finally recognized in his possession of the two kingdoms by the Peace of Vienna (1738). He had also the Presidi, that is, the former Habsburg possessions on the Tuscan coast. Welcome to the Neapolitans because it brought independence to the kingdom, he made his reign glorious by careful reform of the political and administrative system and by splendid building. His most able minister was the great Bernardo Tanucci.

In 1759, on succeeding unexpectedly to the Spanish throne, Charles abdicated his Italian possessions to his third son, who became king as Ferdinand IV of Naples (III of Sicily). There was a council of regency, on which Tanucci sat, until the new king came of age in 1767. The next year Ferdinand married the Austrian archduchess Maria Carolina, an ambitious woman who easily took control of her pleasure-loving husband's government. By ousting Tanucci and putting the Englishman John Acton at the head of affairs, she opened the way for British and Austrian influence to replace Spanish.

The backwardness of the kingdom, where the aristocracy and the clergy were still highly privileged and the lower class lived in degrading poverty, was meanwhile being denounced by progressive thinkers such as Antonio Genovesi; and the Neapolitan government, in the paternalistic spirit of "enlightened despotism," sponsored a vigorous program of legislation to rectify these injustices and to modernize the state. The expulsion of the Jesuits (1767) was another symptom of the current tendency.

The monarchy was halted in its course of reform by the example of the French Revolution, which released a flood of republican and democratic ideas. These ideas appealed strongly to those—middle-class intellectuals, nobles, and churchmen alike—who had seen the Bourbon reforms as designed rather to increase the king's power than to benefit the nation. "Patriots" began to conspire and were countered by persecution. Having entered the War of the First Coalition against France in 1793, Ferdinand had to withdraw from it in 1796, after the Franco-Sardinian Treaty of Paris; but the French intervention in Rome alarmed him, and the British admiral Nelson's victory in the Battle of the Nile encouraged him to take part in the War of the Second Coalition in 1798. His army, under the Austrian Karl von Mack, advanced against the Roman republic but was defeated at Civita Castellana by the French under J.E. Championnet, who then pursued it toward Naples. In Naples, while the local Jacobins were preparing to revolt on the arrival of the French, the populace rose to defend its king; but Ferdinand took panic and fled with his entourage and his treasure to Palermo, having ordered the burning of the fleet built up in the past decade.

The Republic and the first Bourbon restoration. The French entered Naples, and on January 24, 1799, the Parthenopean republic was proclaimed. The republicans, however, who had ingenuously welcomed the French as liberators, were idealists out of touch with the actual life of the people. With no army of their own, they were caught between the royalist-popular counterrevolution and the exorbitant demands of the French. Appointed vicar of the kingdom by Ferdinand, the cardinal Fabrizio Ruffo landed in Calabria and mustered an army of all sorts, including brigands and convicts. The city of Naples, abandoned by the French, fell to Ruffo on June 13, 1799, after desperate resistance by the "patriots." The castles that still held out were won over by a capitulation whereby their defenders were promised freedom to remain at home unmolested or to embark for Toulon. On June 24, however, Nelson's fleet arrived, and Nelson, in agreement with Palermo, repudiated the capitulation, despite Ruffo's protests. The patriots' ships were attacked, many prisoners were taken, and Francesco Caracciolo was hanged. After Ferdinand's return to Naples (July 1799), special tribunals sentenced the leading republicans to death. The jurist Mario Pagano, the physician Domenico Cirillo, the priest Francesco Conforti, the writer Luigia Sanfelice, and the noble Ettore Carafa, conte di Ruvo, were among the victims of a reaction that estranged open-minded intellectuals from the Bourbon dynasty.

Napoleon's terms in the Treaty of Florence (March 1801)

After the Franco-Austrian Peace of Lunéville, Ferdinand had to accept Napoleon's harsh terms in the Treaty of Florence (March 1801), whereby he surrendered the Presidi to France, agreed to grant a general amnesty, and consented to a French occupation of parts of his territory. When the War of the Third Coalition broke out (1805) he entered into secret dealings with Austria and Great Britain. The exasperated Napoleon, having defeated the Austrians at Austerlitz, sent his brother Joseph to conquer Ferdinand's kingdom. Ferdinand fled to Sicily again, and the French overran the mainland. The capital was taken easily, as the bourgeoisie, remembering the troubles of 1799, formed its own militia to hold the populace down (January 1806).

The Decennio. Napoleon first annexed the kingdom to his empire, then declared it independent, with Joseph as king (March 30, 1806). When Joseph was transferred to Spain (1808), Napoleon gave Naples to his brother-in-law Joachim Murat.

The Decennio, as the decade of French rule is called by Neapolitan historians, brought great reforms. Feudalism was abolished; seignorial domains were broken up and ecclesiastical property was confiscated and sold (thus promoting the growth of the agrarian middle class); the laws were unified, with the introduction of the Code Napoléon; army and navy were brought into being on modern lines; brigandage was suppressed; and the administration was reorganized. Neapolitans brought up on the principles of the 18th century but conscious of traditional values collaborated with French officials in the public service.

Murat was deservedly popular as king; but discontent was fomented by clandestine propaganda spread from Sicily by Lord William Bentinck, the British envoy there, who had saddled Ferdinand with a constitution and so could represent him as standing for liberalism. Moreover, Napoleon's tutelage was oppressive and suspicious. To win genuine independence, Murat negotiated secretly with Great Britain and with Austria, concluding a treaty with the latter in January 1814, after Napoleon's defeat at Leipzig. During the Hundred Days, however, he rallied to Napoleon's cause, took up arms against Austria and, on March 30, 1815, from Rimini, appealed to the Italians to join him in a fight for the national independence of the peninsula. Defeated by the Austrians at Tolentino (May 2), he abandoned his kingdom and retired to France.

The second Bourbon restoration. By his Treaty of Casalanza with the Austrians (May 20, 1815), Ferdinand IV recovered the Kingdom of Naples. On the recommendation of the Congress of Vienna, he unified Naples and Sicily to form the Kingdom of the Two Sicilies, of which he was styled Ferdinand I (1816). Though he had assured conservative Austria that he would not allow any institutions incompatible with the absolutism about to be reimposed on the Austrian north of Italy, he took no measures to undo the reforms of the Decennio when he returned to Naples in June 1815. On the contrary, he sought to integrate and to develop its achievement and retained much of its personnel in his service.

Sicily proved obstinately hostile to the centralizing policy of unification with Naples: the Sicilian barons, who had resented the viceroy Domenico Caracciolo's antifeudal program in the 1780s, had in 1812, with Bentinck's connivance, transformed their own parliament and forced the King to accept a constitution on the English model, so that autonomists and liberals were now combined in opposition to the Bourbons. On the mainland, the pathetic failure of Murat's audacious landing in Calabria (October 1815) and his subsequent execution, far from discouraging his partisans, increased their numbers and created a Murat cult which looked backward to the Decennio and forward to the united Italy envisaged by him at Rimini. Even stronger in opposition to the Bourbons were the Carbonari, who were in existence in the kingdom in 1810 and whose ranks, reinforced by men returning from Germany and Switzerland, now included nobles, priests, bourgeois, and, particularly, army officers. The Bourbon government antagonized them by its measures against secret societies, by its favouritism toward its own trusty friends, and by its wide concessions to the church in the Concordat of 1818 with the papacy.

Revolution

Revolution began on July 1, 1820, with a demonstration by a cavalry regiment at Nola. It spread quickly, under the leadership of Gen. Guglielmo Pepe. The rebels demanded the democratic and unicameral Spanish constitution of 1812; Ferdinand granted it on July 7 and swore to it on July 13. The Sicilians, on the other hand, rose for their own constitution of 1812, which was aristocratic and bicameral; and the Neapolitan parliament, consisting entirely of mainland deputies, resorted to armed force against Sicilian separatism. At the Congress of Troppau Austria's call for intervention by the powers against revolutionary threats to the European order was agreed to despite British and French objections, but a decision on Naples was postponed for the Congress of Laibach in 1821.

Granted leave by his parliament to go abroad, Ferdinand attended the Congress of Laibach and promptly broke his oath to the constitution. With his consent and the mandate of the Holy Alliance, Austria undertook to restore his absolute monarchy. Pepe's forces were defeated at Rieti, and on March 23, 1821, the Austrians entered Naples. They were to stay there, as the watchdogs of the Bourbons, until 1827, when King Francis I obtained their recall and engaged a Swiss bodyguard instead.

Reprisals were taken against the liberals; and the kingdom was dominated by the police for the rest of Ferdinand's reign and throughout Francis I's (1825–30). Resentment bred an outcrop of malcontents, a reemergence, under various names, of the Carbonari. A new revolt, in the Cilento, was put down by force (1828).

Francis I's son and successor Ferdinand II began his reign with measures expressly intended to amend things. Trade and industry increased; agricultural production was encouraged; the lot of the workers was improved; better budgets allowed a reduction of taxes; laws were brought up to date; and public works were undertaken. Ferdinand, however, allowed no political reflection of this economic progress, though he issued an amnesty on his accession and was ready to employ, as administrative experts, men who had previously been exiled or imprisoned. In foreign affairs he was inclined to neutrality, as he saw that alliances might compromise his precious independence.

Ferdinand II's attitude was thus quite irreconcilable with the spirit of the Italian Risorgimento, which was affecting his kingdom through various channels. Periodicals and intellectual groups sprang up to satisfy the revived taste for literature, history, philosophy, and the arts; and liberalism and nationalism thrived in the new atmosphere. Vincenzo Gioberti, who had a Neapolitan precursor in Vincenzo Cuoco (1770–1823), captivated the majority of would-be reformers with his book *Del primato,* which appealed to tradition and called for an Italian federation of the existing states under papal presidency. Others held less moderate views: Giuseppe Mazzini, who wanted a unitary republic of Italy, had converts in the kingdom; and the backward areas were sprinkled with revolutionary cells plotting to subvert the economic and social order, as for instance Benedetto Musolino's Calabrian association Giovane Italia (not to be confused with Mazzini's Giovine Italia). Gradually the old political world of Naples was undermined by a more or less clandestine and heterogeneous opposition: liberal constitutionalists (mainly former Carbonari), led by Carlo Poerio; Giobertians, led by the historian Carlo Troya; and democrats of all sorts, including extremists. Successive revolts proved that repression was not enough to stifle the revolutionary ferment. A violent rising at Cosenza in 1844 was put down rigorously, but its example prompted Mazzini to encourage the ill-fated expedition of the brothers Attilio and Emilio Bandiera to Calabria in 1845.

Rejoicings at Pope Pius IX's early marks of sympathy for the Italian cause (1846–47) were paralleled by overt attacks on Bourbon tyranny: Luigi Settembrini published his *Protesta del popolo delle due Sicilie* in July 1847. There followed the great year of European revolutions. Sicily, still obstinately separatist, was in revolt from January 12, 1848; and alarm for the mainland induced Ferdinand II to grant a constitution on January 29. Modelled on the French constitution of 1830, this was promulgated on February 11. A constitutional guard was formed; the kingdom joined the Italian customs union; and finally, under pressure from the people, war was declared against Austria on April 7. Pepe was sent with an expedition to Lombardy while the fleet sailed for Venice.

Constitution of 1848

Elections took place in April 1848. The democrats were agitating in the press for a parliamentary revision of the constitution and for other controversial measures. There were riots in Naples, and the peasants' seizure of national, communal, and private lands exposed the activity of socialist and communist cells. Parliament was to have met on May 15; but on May 13 the radicals demanded that the king's oath should acknowledge the parliament's right to revise the constitution. The King refused, and barricades were put up (May 14). Ferdinand would not open parliament unless the barricades were dismantled, but the rebels stood firm until troops had mastered them by force (evening of May 15). When Troya resigned from the ministry, Ferdinand replaced him with the reactionary principe di Cariati (Gennaro Spinelli). Eventually the King dissolved parliament, announced fresh elections, recalled his troops from Lombardy, and broke off relations with Sardinia-Piedmont (July). The democrats were now openly opposing the liberals and working for a republic. The provinces were still in commotion. When the elections returned as many radicals as before, Ferdinand postponed the opening of parliament from July 1848 to November and then to February 1849.

After both Pius IX and Leopold II of Tuscany had been driven by revolution to take refuge at Gaeta, Ferdinand dissolved the parliament and shelved the constitution (March 1849); and the Austrian victory at Novara gave a free hand to reaction in Italy. Sicily, which in summer 1848 had declared the Bourbon dynasty deposed, was subdued by Carlo Filangieri in April–May 1849. Sentences of death (commuted by Ferdinand to penal servitude), imprisonment, banishments, and dismissals, were widespread: Carlo Poerio, Silvio Spaventa, Francesco de Sanctis, and Luigi Settembrini were victims of the reprisals. The liberals, having identified themselves with the cause of Italian national unity, were wholly discredited, and vigorous absolutism was brought back. Believing that he had thus won tranquillity for his subjects, Ferdinand devoted himself again to public works and to economic improvement. Foreign liberals, such as Gladstone in the British press (1851) and Cavour at the Paris conference (1856), inveighed against the kingdom as morally bankrupt and politically anachronistic. Carlo Pisacane, a former officer of the Bourbon army who had fought for the Roman republic and had absorbed Mazzinian and socialist ideas, landed at Sapri with a small force in 1857 but was defeated and killed by the local police and townspeople. Next year, yielding to French and British pressure, Ferdinand released into exile some of his political prisoners. In May 1859, on his deathbed, while the French and the Sardinians were beginning their campaign in Lombardy, Ferdinand adjured his heir, Francis II, to stay neutral and to eschew innovations.

On his accession to the throne, Francis II appointed as prime minister the moderate Carlo Filangieri, who advised him to reinstate the constitution and to accept the alliance offered to him by Victor Emmanuel II of Sardinia. Francis hesitated, while his kingdom was shaken with the reverberation of Sardinian successes in Lombardy and central Italy. The Swiss regiments, hitherto the mainstay of royal absolutism, mutinied in July 1859 and were disbanded; Filangieri resigned in March 1860; and Palermo rose in revolt on April 4.

Sicily, the hotbed of antidynastic feeling, had already been infiltrated by emissaries of the Comitato d'Azione, an organization under Mazzinian influence whose aim was to unify Italy by means of revolution, instead of waiting for Cavour to do so by internationally acceptable means. These emissaries, notably Francesco Crispi and Rossolini Pilo, prepared the ground for Giuseppe Garibaldi, who landed with his Thousand at Marsala on May 11, 1860, declared himself for "Italy and Victor Emmanuel" (to the chagrin of the Mazzinians), defeated the Bourbon troops at Calatafimi on May 15, and was in Palermo (which the Bourbons had recovered since its revolt) by the end of the month.

Garibaldi

Francis on June 25, 1860, signed a proclamation promising a general amnesty for political offenders and a liberal ministry; and on July 1 he restored the constitution; but the amnesty only exposed Naples to the return of exiles who were either converts to Cavour's plan for a unitary Italian monarchy under the House of Savoy or partisans of Lucien Murat, Joachim's heir. Garibaldi's victory at Milazzo (July 20) put an end to Bourbon resistance throughout Sicily, except in the citadel of Messina; and Great Britain, France, Russia, and Sardinia ignored Francis II's plea that the invasion should be halted at the straits. Cavour, who, unlike Victor Emmanuel II, had at first objected to the expedition both because its democratic sponsors had opposed his cession of Nice and Savoy to France and because its revolutionary aspect might have compromised his plans, now began to abet it. On August 18, Garibaldi crossed the straits and took Reggio di Calabria. As his forces advanced northward, Bourbon rule collapsed, province by province. Francis fled to Gaeta on September 6; and Garibaldi, having entered Naples in triumph on September 7, assumed the dictatorship on September 8. The Neapolitan fleet refused to follow the King.

The revolutionaries with Garibaldi's forces might now have proclaimed a republic; and if they had marched on Rome, France and Austria would have intervened. To forestall this and at the same time to seize a role for Sardinia in the liberation of southern Italy, Cavour exploited the pretext furnished by the disorderly conduct of the foreign volunteers who were rallying to defend the Papal States. The Sardinian army began to overrun the Papal States on September 11, 1860; and on October 12 it crossed the frontier into Neapolitan territory. A Bourbon counteroffensive against the Garibaldians was defeated in a desperate and hard-fought battle on the Volturno River (October 1–2).

In Naples, meanwhile, Cavour's followers, in favour of annexation by Sardinia, were disputing with anti-annexationists, whom Mazzini and Carlo Cattaneo arrived to reinforce. Plebiscites, held on October 21, 1860, one for the mainland and one for Sicily, came out in favour of annexation. On October 26, at Caianello near Teano, Garibaldi saluted Victor Emmanuel as "king of Italy," a title confirmed by the Italian parliament at Turin in February 1861.

The French naval squadron that had been protecting the heroic defenders of Gaeta was withdrawn, at British insistence, in January 1861; and on February 13 Francis and his family sailed for Rome, whereupon Gaeta surrendered. The citadel of Messina capitulated on March 12, that of Civitella del Tronto on March 20. The 700-year-old kingdom was no more.

The 20th century. The Mezzogiorno in the recent past has suffered economically and socially compared to the prosperous north. Not only was there a great chasm between large landholders and peasants, but the government structure, by inadequately providing for the welfare of the people, encouraged criminal organizations, such as the Mafia, to flourish. Feudal traditions remained very strong in the south; few resources, as well as preferential treatment by the government toward the north, kept the Mezzogiorno underdeveloped. In addition, southern politicians placated the wealthy with special advantages.

The spread of organized crime

In the late 19th century, awareness of the south's situation spawned a style of literature known as Letturatura Meridionalista. Writings portrayed the exploitation of the landless peasants and the detrimental effects of government policy there.

The central government undertook some public projects in the early 20th century, but these were insufficient and many residents decided to emigrate. Fascist policy in the 1920s and 1930s concentrated primarily on controlling the Mafia's power. After World War II some large estates were dismantled and redistributed. The Cassa per il Mezzogiorno (Fund for the South) represented the government's efforts to encourage economic development. Some of the programs have included providing capital for investment and for industrial projects.

For additional history, see *Sicily* and *Sardinia* in this article.

(E.Po./Ed.)

Sicily

PHYSICAL AND HUMAN GEOGRAPHY

Location and general character of the island

The largest and one of the most densely populated islands in the Mediterranean Sea, Sicily (with several small adjacent islands) is an autonomous region of Italy. Its northern shore, site of the capital, Palermo, is lapped by the Tyrrhenian Sea; to the east it looks across the narrow Strait of Messina to the toe of the Italian peninsula; and it lies about 100 miles (160 kilometres) northeast of Tunisia, in North Africa. Sicily occupies 9,830 square miles (25,460 square kilometres), and with its adjacent islands it covers 9,927 square miles.

Sicily's strategic location at the centre of the Mediterranean has made the island a turbulent crossroads of history, a pawn of conquest and empire, and a melting pot for a dozen or more ethnic groups whose warriors or merchants sought its shores virtually since the dawn of recorded history. The island and its people have lived much by the sea, flourishing when the Mediterranean traffic was intense and stagnating when it waned. Only a few days' sailing time from the ancient civilizations of Greece and the Aegean, Sicily was at times the major western outpost of the high culture of those regions. Its land also lies open to climatic influences from all directions. In spite of these many forces that so often have reshaped its political and social life and of the economic depression that continues to plague it, Sicily has been able to establish one of the most distinct identities among the Italian regions, a distinctiveness that is guarded jealously by its inhabitants.

The land. *Relief.* The landscape of Sicily is mainly one of mountains, sometimes reaching almost to the coasts, and of irregular coastal plains on which all the main cities are located. The present topography is the result of the fact that the rocks erode easily in a climate marked by intense exposure to the sun, persistent drought during the long summers, and the sea winds.

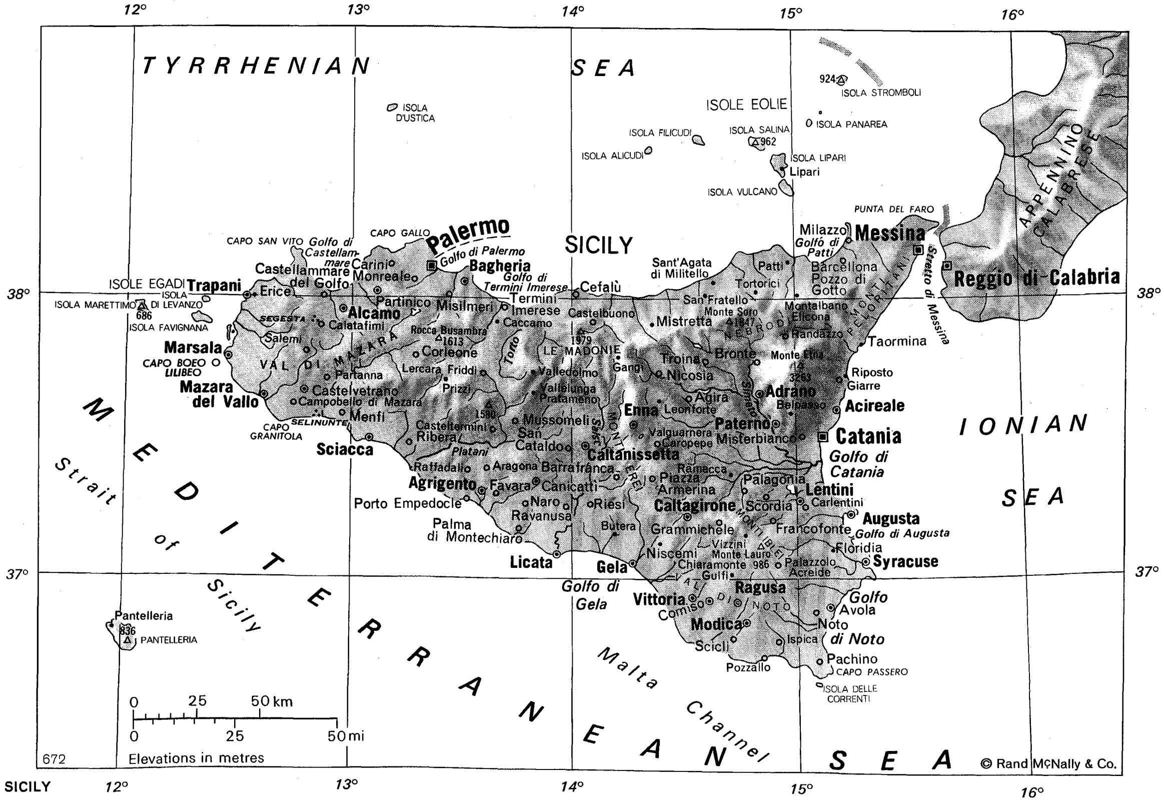
The mountain chains

The mountains are continuations of the Apennines of the Italian peninsula and the Atlas Mountains of northwestern Africa, which are parts of the discontinuous Alpine–Himalayan mountain belt circling much of the Earth from Asia to southern Europe and northern Africa. In northeastern Sicily the Peloritani chain rises suddenly from the Strait of Messina and continues the landscape of the Calabrian Apennines of the mainland, though without the typical high plateaus. The higher reaches have little or no vegetation and are deeply scarred by currents that, at lower altitudes, produce frequent flash floods and sweep away both rubble and fertile sediments. Forests of chestnut, beech, or Mediterranean scrub climb above the cultivated lower slopes. Maximum altitudes are about 4,300 feet (1,300 metres).

The Nebrodi and Le Madonie chains to the west are higher, reaching 5,900 and 6,600 feet, respectively. The forms are similar, though in Le Madonie they tend to be stark, standing out like great limestone massifs or peninsulas. Le Madonie is known for its underground drainage, which provides water to Palermo. Inland, the mountains gradually diminish in height toward the south, with great expanses of virtual tableland known for its sulfurous, desolate conditions; grasses and twisted olive trees dominate the landscape. The Iblei chain in the southeast is of volcanic origin and presents a refreshing scene of brightly coloured vegetation.

Mt. Etna

To the northeast of this chain, the great mass of Mt. Etna rises to 10,703 feet (3,263 metres), the highest point in Sicily, which is a region of strong seismic activity. One of the world's most active volcanoes and the biggest in all of Europe, Mt. Etna covers half the province of Catania and rises above the provincial capital of that name. It is thought that this massive mountain was gradually built by extruded volcanic debris from beneath the sea to its present height. Luxuriant wild and cultivated trees and shrubs cover its sides wherever a cover of decomposed debris occurs. Underground water and springs are plentiful for the many vineyards and citrus orchards that grow in the region.



SICILY

Climate. Sicily's most notable climatic factors are its southern latitude, its insularity, its elevations decreasing from north to south, and the moderate atmospheric pressure characteristic of the Mediterranean. Latitude and the pattern of winds make it a sun-drenched island, and desert winds from Africa often scourge the southern and central regions during the May-to-October summers.

Summer temperatures often climb over 85° F (29° C), but the climate is even more influenced by the rainfall, which is irregular in quantity from year to year but regular in its uneven annual distribution, virtually all of it falling in late autumn and winter. The south receives the least amount of rain and the northeast the most. Much of the fall is lost through deforestation (less than 4 percent of the island is woodland) and the soil's resultant inability to hold the torrents from the mountains; and, though the amounts are similar to those on the Italian mainland at equivalent latitudes, the irregularity of the fall and its concentration in short downpours contributes significantly to the island's agricultural depression and rural poverty. Neither irrigation facilities nor discoveries of underground water have begun to counteract this natural blight.

Plant and animal life. Only on the higher slopes does the chestnut replace the laurel, perhaps the most characteristic Sicilian tree. Pines grow on Mt. Etna, on whose upper reaches grow the low, isolated shrubs and grasses of the Alpine zone. Of the cultivated plants, only the flowering ash is native; the olive, grape, almond, pomegranate, hazel, and other fruit and nut trees, however, date from antiquity. Arabs introduced the date palm and citrus trees, though the common orange is of 16th-century Chinese origin, and other species came a century later from the Americas. Of the few animal species, the invertebrates, birds, and fish outnumber mammals, though remains of vanished large mammals occasionally are encountered in caves. Overall, the animal population represents a transitional zone between Italy and Africa.

The people. *Ethnic composition.* Sicily was already inhabited toward the end of the Pleistocene Epoch (more than 10,000 years ago), and throughout its history has received peoples of every coast on the Mediterranean. According to archaeologists, anthropologists, philologists, and historians, the peoples living there in the remotest times, the Sicani, were of Mediterranean origin. Some of the physical features of the present-day population may be traced to them: elongated heads, small stature, and dark pigmentation of the eyes and hair. Interbred with them were the Elymi, who left their mark mainly on the towns of Segesta, Erice, and Entella; they may have come in small numbers from neighbouring Africa.

Diversity of ethnic stocks and physical types

Then the Siculi, a short-headed Eurasian people who spoke an Italic language akin to Latin, came down the peninsula from the central and eastern Alps; they seem to have infiltrated slowly and in small waves among these ancient Mediterranean peoples. The Siculi and Sicani seem to have formed the most homogeneous group of the island's population. But as early as the transition from the Bronze to the Iron Age, they appear to have been molested by Greek pirates, coinciding and in a sense competing with Etruscan pirates in the lower Tyrrhenian. Some pirate groups appear to have established themselves on the island, forcing the inhabitants to withdraw inland and fortify their new settlements. Later, between the 8th and 6th centuries BC, migrations from Greece seem to have been more frequent and of larger size. The new peoples—among them Ionians from Euboea and Samos, Aeolians and Achaeans from Boeotia, and Dorians from Corinth and other locations—had apparently left their homelands to settle on the island.

At more or less the same time, the Carthaginians, Semites from the North African coast, set up some of their trading posts along the western coast of Sicily. Some of the tyrants then welcomed exiles from the neighbouring continent and recruited labourers to cultivate the land and mercenaries to fight their battles from the Italian peninsula, Greece, Asia, and Africa. These new immigrants remained and merged with their predecessors, an example that was later followed by Romans, Vandals, Goths, and Byzantines, who came for reasons of war or conquest.

From Arabia, Egypt, Syria, and Mesopotamia came the Muslims, who took only 80 years to conquer Sicily, settling mainly in the northern regions, while the Berbers were more predominant in the south. Many Sicilian Christians converted to Islām. Around the year 1000 the Normans arrived and held the island for more than a century; then came the Swabians, who held it for three-quarters of a century. The tall stature, blond hair, blue eyes, and pink colouring of some present-day Sicilians may be traceable to them. The Aragonese also governed Sicily for more than a century.

Demography. The Sicilian population is the most urbanized of the provinces in southern Italy; yet it displays many characteristics of a rural population, a fact attributable primarily to a generally low standard of living. The greatest proportion live in the provinces of Palermo and Catania, and only a small fraction live outside the narrow coastal strip around the island. Migrants long have been driven from the interior not only by the barrenness but also by malaria and by the land-tenure system, *latifundium,* characterized by large estates ruling over great wheat-growing acreages.

The economy. *Incomes.* The per capita income of Sicilians is substantially less than the national average. The tertiary sector of the economy—retailing and services—plus the civil service account for about one-half the island's income. Industry has gradually outstripped agriculture to contribute the greater proportion of the other one-half, especially in Palermo. Much of this industry, moreover, is based on the processing of the products of farm, forest, fishery, and mine.

Agriculture, industry, and energy resources

Components. Agriculture employs about one-quarter of the labour force, and a very high percentage of Sicily's land area is given over to it and to forestry. Much of this is sown in wheat, which does not require great fertility. In ancient Rome, Sicily was known as one of the breadbaskets of the republic and empire. Citrus fruits (including nine-tenths of Italy's entire lemon production) and vegetables, however, provide the largest part of agricultural exports to the peninsula and abroad. Dams, flood- and erosion-control measures, and irrigation canals are among the major projects undertaken in recent years to assist the farmer.

Among the more promising aspects of industry, which employs about one-third of all workers, have been the discoveries of petroleum and the building of refineries to complement the plants producing chemicals, pharmaceuticals, cement, fertilizers, and synthetics. These are in addition to the small-scale food- and sulfur-processing plants and apparel manufacturing. Mineral production also includes limestone, asphalt, potash, and salt.

Virtually all of Sicily's energy is generated in thermal plants; hydro potentiality is nil. The three main ports—Palermo, Catania, and Messina—have extensive shipping facilities and are interconnected by single-track railways. Main highways between cities are adequate, but secondary roads are poor. Many interior communities are reachable only by track.

Governmental involvement. Since World War II, the national government has given Sicily and other areas of southern Italy special funds to improve transportation facilities and other public works and to prepare them for industrial development. In its turn, the regional government took numerous steps to ease the location and capitalization of industry. Years of political instability and shortage of resources, however, have tended to impede even the legislated goals.

Administrative and social conditions. *Government.* An Italian constitutional statute of 1948 made Sicily an autonomous region, divided into nine provinces. Government is parliamentary, with the president elected from among the members of the regional assembly. With the assembly's executive committee, he promulgates laws and may attend meetings of the Italian council of ministers and speak on Sicilian matters. He represents the federal government and is charged with maintaining order with the help of the national police force. A Sicilian separatist movement won minority support after World War II but faded in the 1950s.

The social milieu. Life in much of the interior has remained in the late 20th century at a primitive level, with inadequate water supplies and sanitation, and rampant disease. Unemployment has been chronic, except around the developing petroleum-producing and petroleum-processing centres.

The regional government exercises complete control of primary education, but the illiteracy rate remains high. The universities in Palermo, Messina, and Catania are required to maintain national standards. The well-stocked civil service provides a well-organized system of social services that are generally of low calibre. Good housing and medical services, even in the cities, tend to be in short supply as well.

A peculiar feature of the separateness of Sicilian life from that of mainland Italy is the persistence of the Mafia, an organization dating from the Middle Ages that gradually evolved into a paralegal criminal brotherhood. It gives certain parts of the island virtually a dual government, standard of conduct, and system of enforcement—one is the legitimate regime and the other a shadow but a pervasive social, economic, and political network maintaining its powers through violence. In recent years it has tended to move away from the rural areas, in which it was spawned, to urban Sicily, in which through its traditional code of silence and frequently violent settlement of disputes it directs its activities at legitimate businesses through extortion and terror.

Cultural life. Sicily has made its contribution to Italian literature and art. The poets of the Sicilian school at the court of the emperor Frederick II were important in the development of Italian lyrical poetry in the vernacular early in the 13th century. The Byzantine and Norman architecture of the island with its splendid mosaic murals is celebrated. In the 15th century the great painter Antonella da Messina introduced the Flemish technique of oil painting to Italy.

In more recent times, in the realm of the "fine" arts, Palermo has its opera house and Catania its Bellini Theatre, named in honour of the island's major musical celebrity, the opera composer Vincenzo Bellini. In literature the influences of two Sicilians, the novelist Giovanni Verga and the playwright Luigi Pirandello, have profoundly affected the shape of modern Italian literature and have had great influence far beyond the shores of the Mediterranean. In 1959 the Sicilian poet Salvatore Quasimodo was awarded the Nobel Prize for Literature.

Italian song

Sicily has been called the cradle of the Italian popular song. Various kinds of song exist locally: love songs are the most numerous, followed by the work songs of such occupations as fishing (the *mattanza,* or tuna fishing song) and by the sung accompaniments to the many children's games. The songs of the beggars and the musical cries of the street vendors are still heard, though they, like much traditional culture, are disappearing under the homogenizing influence of radio and television and of modern life in general.

Sicily has strong traditions of folk art. The elaborately painted carts of the peasants, which are features of the provinces of Catania and Palermo, depict on their sideboards scenes of combat between the Christian knights and the Saracens. Though no extant carts predate the 19th century, their art may commemorate the era of Sicilian prosperity under the Norman kings and the themes of chivalric quest popular in the epic poetry of that time. Similar themes run through many of the puppet dramas, which remain popular on the island. Silversmiths' and goldsmiths' work, to be seen especially in Palermo, and peasant filigree settings for gems are of the most intricate refinement, and Sicilian embroidery has an uninterrupted tradition many centuries old.

Popular religious festivals make up colourful occasions for the sun-bleached land. The processions of triumphal cars of the feasts celebrating St. Rosalia in Palermo, the Annunciation in Trapani, and St. Lucia in Syracuse, as well as the historical pageants featuring symbolic figures of Christians and Saracens and of Normans and Turks in combat to the accompaniment of fireworks and dancing, represent the gayer sides of Sicilian life.

(Ed.)

HISTORY

Greek Sicily. At the coming of the Greeks three peoples occupied the island of Sicily: in the east the Sicels, or Siculi, who gave their name to the island but were reputed to be latecomers from Italy; to the west of the Gelas River, the Sicani; and in the extreme west the Elymians, a people to whom a Trojan origin was assigned, with their chief centres at Segesta and at Eryx (Erice) with its temple of Aphrodite. The Sicels spoke an Indo-European language; there are no remains of the languages of the other peoples. There were also Phoenician settlements on the island. The Greeks settled Sicilian towns between the 8th and 6th centuries BC. The mountainous centre remained in the hands of Sicels and Sicani, who were increasingly Hellenized in ideas and material culture.

After the foundation of the colonies, little is known of them until the 5th century BC. They prospered materially, building temples whose remains, at Syracuse, Akragas, and Selinus, are among the finest monuments of archaic and classical architecture. Their sculpture and other arts were vigorous, though provincial and largely dependent on impulses from mainland Greece. There is some reason to believe that they had a colonial economy, producing food (corn, sheep and cattle, fish) and importing manufactured objects (of which clay vases are the chief survivors) from Greece—in the period before 550 BC mainly from Corinth, afterward from Athens. The broad acres of the colonies and the use of Sicel serfs at Syracuse and probably also elsewhere gave rise to a life of easy circumstances and, in the 5th century, contributed to the great wealth of the tyrants. The fine series of Sicilian coins began toward the middle of the 6th century. The poet Stesichorus of Himera, who retold many of the stories of epic with a new spirit reflected in archaic vase paintings, belonged to this period.

Sicily's tyrants

There were many tyrants—unconstitutional monarchs—in the 6th century, especially in those towns that were pressed by the growing hostility of the Carthaginians; the most famous was Phalaris of Akragas. In general, however, this was a period of aristocratic or oligarchical constitutions. The early 5th century saw tyrants in most cities. The tyrannies were short-lived, however, and were succeeded by uneasy democracies in most cities, Syracuse still taking the lead. The Sicels then for the first time formed a united power, under Ducetius, who founded a new city at Palici in the plain of Catania and became a serious threat to Syracuse and to Akragas. He was defeated and his political work largely undone, except for the foundation of Cale Acte (Caronia) on the north coast. The Sicels did not again attempt to combine as a force, but their Hellenization continued; many Sicel cities began to coin on the Greek model.

The Athenians looked for allies in Sicily at least as early as 458/457. The ultimate object of this expedition was the subjection of the whole of Sicily. The Athenians were received with suspicion, even by some of their allies, and Naxos and Catania were the only Greek cities to support them; they also had help from some of the Sicels. The utter defeat of the Athenian force under the walls of Syracuse in 413 was achieved in part by the arrival of the Spartan leader Gylippus and was followed by the dispatch of small forces from Syracuse and other Sicilian cities to help the Spartans against Athens.

Carthaginian wars. The failure of Athens left the field open for Carthage, with whom the Athenians had sought an alliance. Dionysius the Elder, leader of Syracuse, fought a series of inconclusive wars against the Carthaginians, finally winning a victory *c.* 396 BC. He had to fight other Carthaginian wars, however, in 392–391, 383–378, and 368–367, the year of his death. In the peace made in 378 the boundary between Greek and Carthaginian Sicily was fixed at the River Halycus (Platani).

Dionysius of Syracuse

Dionysius used his position as defender of Greek Sicily to build up a personal power that anticipated the Hellenistic monarchies in many ways. The transplantations of population and refoundations of cities that had been a feature of the rule of Gelon and Hieron continued.

The rule of Dionysius at Syracuse depended on foreign mercenaries and on secret police, and many of the typical features of Plato's and Aristotle's accounts of Greek tyrants are no doubt derived from him. But he was well served by efficient ministers, such as Philistus the historian (who, however, went into exile). He was himself a poet and, like earlier tyrants, kept a court, which was visited by Plato and other philosophers and poets. He made Syracuse the greatest power in the Greek world, and it became the largest and, probably, the most populous of Greek cities, with its extensive fortifications. His power, however, did not long survive him. Most of his conquests in Italy fell to the Lucanians and Bruttii, and at Syracuse his son Dionysius the Younger, after ruling from 367 to 356, was expelled by his uncle Dion. Dion was killed in 354, and there followed 10 years' confusion: fighting in Syracuse between Dionysius the Younger and the citizens; tyrannies in other cities; and renewed danger from Carthage, allied with Hicetas, tyrant of Leontini and rival of Dionysius. The Syracusans appealed to their mother city Corinth, and Timoleon came as a deliverer. He freed the cities from tyrants and defeated the Carthaginians at the Battle of the Crimisus (probably 341), but the boundary remained at the River Halycus. Timoleon's reputation was high because he restored the democracy at Syracuse and retired after his work of liberation was done, but his settlement did not last long after his death (*c.* 336). The Carthaginians, who had already played off one city against another, continued this policy, while the Greeks consumed their energies in struggles between would-be tyrants. They found another leader in Agathocles, one of the ablest soldiers of the time, who made himself tyrant of Syracuse in 317. Akragas, strengthened by Syracusan exiles, stood out again as the rival of Syracuse; and Hamilcar, son of Gisgo, won many Greek cities to the Carthaginian alliance and blockaded Agathocles in Syracuse. Agathocles broke through and carried the war into Africa, where he won many successes (310–307) but was finally completely defeated and had to flee back to Sicily. In spite of this defeat he maintained his position at Syracuse and made peace on the old terms with Carthage. He formed marriage alliances with Ptolemy I of Egypt and with Pyrrhus of Epirus, who married his daughter. He was the first of the Sicilian tyrants to take the title of king. He died in 289. In spite of his reputation for treachery and massacre, his rule was remembered as a period of prosperity.

The Mamertini

In the troubles which followed Agathocles' death, his disbanded Campanian mercenaries seized Messana and called themselves the Mamertini, children of Mamers, or Mars. The fortunes of Sicily were thus linked with Rome. The attack on the Mamertini by Hieron II, king of Syracuse, led in 264 to the intervention of Rome and the First Punic War (264–241). The war began as a three-cornered event between Rome, the Carthaginians, and Hieron; but in 263 Hieron turned from the Carthaginians to Rome and formed an alliance to which he remained loyal for the rest of his long life. The Romans were thus free to use Greek Sicily as a base for war with Carthage. The western part of the island, both Greek and Phoenician, suffered greatly in this long war. By the treaty which ended it Carthage ceded to Rome all its possessions in Sicily, which thus became the first Roman province (241). Hieron retained possession of eastern Sicily, south of Messana. His rule was able and enlightened, and his financial enactments, particularly his corn laws, were taken over when Rome incorporated his kingdom. This period of peace was the last golden age of free Sicily.

Roman Sicily. At the outbreak of the Second Punic War Hieron held firm to the Roman alliance, but after his death in 216 his grandson Hieronymus repudiated it. Hieronymus was overthrown by revolution at Syracuse (215), but the city had to stand a siege from the Romans. The great fortress of Euryelus perhaps took its final form then, under the inspiration of Archimedes, though there is no reason to doubt that its construction began in the time of Dionysius I. Syracuse was taken and sacked in 212, Akragas after a further campaign in 210. The whole of Sicily then became Roman.

Little is known of the early organization of the Roman province. It was governed by a praetor sent out yearly from Rome, who after the annexation of Syracuse had his capital there. Two quaestors were appointed, one with his office at Syracuse, the other at Lilybaeum. The province included a number of free cities: Messana, Tauromenium, and Netum (Noto) were allied cities (the two latter had probably taken the Roman side in the Second Punic War); a number of others, including allies from the First Punic War, were also free—Segesta, Halicyae, Panormos, Halaesa, and Centuripe. The rest paid tithes to the Roman people according to the law of Hieron, which was extended to the whole island. Sicily had long had a surplus of corn for export, and Livy records occasional dispatch to Rome as early as the 5th century BC; the island now became the granary of the Roman people. The rolling country of the central part of the island was suitable for pasture and cultivation on a large scale in latifundia (landed estates), and slave gangs were introduced on the estates both of rich Sicilians and of Roman citizens. Hence arose the two great slave revolts of the second half of the 2nd century BC, the first, led by Eunus, from 135 to 132, with Enna and Tauromenium as its centres, the second from 104 to 100, both periods of internal and external stress at Rome. The settlements after these two wars by Publius Rupilius (131) and Manius Aquilius (99) modified the organization of the province.

In spite of slave wars and the burden of Roman provincial governors and tax farmers, Sicily was not unprosperous under the Roman republic. It was free from external dangers; and even the unprivileged cities kept their own laws, magistrates, and assemblies, and provision was made for lawsuits between Romans and Sicilians and between Sicilians of different cities. There seems not to have been much commercial exploitation; tax collecting was normally in the hands of the Greeks themselves, not of Roman *publicani;* and smallholdings continued to be the rule in many cities. The wealth of the cities, both free and tributary, may be seen from the speeches of Cicero in prosecution of Gaius Veres, who in three years' governorship (73–71) had plundered widely, with especial attention to works of art. He also failed to defend the province against pirates.

Sicily was again a battlefield between 43 and 36 BC, when Sextus Pompeius held Messana and cut off the corn supply of Rome. In the division of provinces between Augustus and the Senate, Sicily fell to the latter. It had perhaps received Latin rights from Julius Caesar. Augustus planted colonies at Panormos, Syracuse, Tauromenium, Thermae, Tyndaris, and Catania. But the island remained Greek; not only the old Greek cities but also the old Sicel towns, which had long been completely assimilated to the Greek, used Greek as their everyday language, though Latin was the official language. Christianity was early introduced to Syracuse, where the catacombs and early churches (belonging mainly to the Byzantine period) are second only to those of Rome.

Byzantine rule. The Byzantine general Belisarius occupied Sicily in 535, when his emperor began hostilities against the Ostrogoths in Italy, and after a short time Sicily came under Byzantine rule. When the emperor Heraclius and his successors divided the empire into themes (provinces), Sicily became one of them, placed between the exarchate of Ravenna in the north and that of Carthage in the south; it was administered by a *patricius* responsible to the government at Constantinople. Probably after the Italian revolt of 726 and certainly after the fall of the exarchate in the middle of the 8th century, the Byzantine dominions in southern Italy were incorporated in the theme of Sicily, an arrangement that lasted until the time of the Arab conquest, when the mainland dominions were formed into the themes of Calabria and Longobardia.

Ecclesiastically, Sicily remained at first under the papacy, which in addition to rights of jurisdiction had considerable interests in the island, arising from its vast Sicilian estates. Gregory I's letters admirably illustrate the importance attached to them by the pope. The Iconoclastic Controversy and the ensuing revolt against Byzantine rule under papal leadership (726) led to the confiscation by the emperor of the papal estates in Sicily and southern Italy; soon after, the ecclesiastical jurisdiction of these regions passed to the patriarch of Constantinople.

These political and ecclesiastical changes corresponded to demographic and cultural developments of longer standing. In Sicily and southern Italy, the Greek element had been greatly strengthened since the end of the 6th century as a result of emigration from other Byzantine provinces after the Avar and Slav invasions of Greece and, later, of the Persian and Arab conquests. The Hellenization of Sicily, which appears to have been all but complete in the 8th century, is revealed in the history of the Sicilian Church. The Greek rite, already used at the time of Gregory I, spread during the 7th century from the eastern coast over the whole of the island. By the time of its separation from Rome, the Sicilian Church was virtually Greek. Sicily was thus well on the way to becoming a fully integrated part of the Byzantine Empire when the Arab conquest began in 827.

Spread of the Greek rite

Arab conquest. Ever since the 7th century, Arab expansion in North Africa had constituted an immediate threat to Sicily and to southern Italy, the occupation of which would moreover expose Greece, the exarchate of Ravenna, and Dalmatia to Saracen attack. Sicily consequently became a vital link in the imperial defense against Islām.

Rebellions within Sicily significantly weakened Byzantine rule, leading finally to the Arab conquest. The last Byzantine stronghold, Rameta, was not lost until 965. The Byzantines, however, did not abandon hope of reconquering Sicily. Basil II was planning a Sicilian expedition at the time of his death (1025); it was actually dispatched in 1038 under the great Byzantine general George Maniaces. The campaign was highly successful, and a large part of eastern Sicily, including Messina and Syracuse, was recaptured; but after Maniaces was suddenly recalled, the Byzantine position on the island collapsed.

The Arab conquest separated Sicily not only from Constantinople but also from the Italian mainland, where the Arabs did not succeed in establishing themselves permanently. The history of Arab Sicily, on the other hand, was marked by growing independence from Africa. Until 909 it was under the Aghlabids; after their fall, it passed under the Fāṭimids, who moved their capital to Cairo in 972. But from the middle of the 10th century the office of the governor (amir) became hereditary in the dynasty of the Kalbids, until the anarchy after Maniaces' conquests led to the fall of that dynasty and to the division of Sicily into a number of principalities, while Palermo acquired self-government.

During and after the conquest, large Arab immigration from Africa took place; this, together with the conversions of Christians, contributed to make Sicily not only politically but also culturally part of the Arab world. However, the Greek Christian element remained predominant in the Val Demone; and even after the Arabs gained a majority in the rest of Sicily, Christian groups remained scattered over the island.

As in other Arab countries, the Christians, although placed in a position of legal inferiority, enjoyed religious toleration and a measure of self-government in return for paying taxes, which may have been often less burdensome than those levied by the Byzantines. Relics of the Greek episcopate seem to have survived to the end of the Arab period; so did a number of Basilian monasteries; both provided the principal link with the Byzantine world outside (particularly with Calabria, where the Greek population had been strengthened by emigrants from Sicily). As far as the Sicilian clergy, secular and regular, was concerned, this emigration seems to have been less the result of persecutions than of the gradual spread of Islām over Sicily.

The Norman conquest: Roger I. When the Normans began to conquer Sicily in 1060, they were welcomed as liberators, and their progress was doubtless assisted by the Christian population. They were no newcomers to the island. Norman mercenaries from the mainland, among them two sons of Tancred of Hauteville, had fought in Maniaces' army and taken part in the capture of Messina. In 1059 Pope Nicholas II invested another son of Tancred, Robert Guiscard, with his past and future conquests not only in Apulia and Calabria but also in Sicily. In his oath of allegiance Robert styled himself "by the grace of God and St. Peter duke of Apulia and Calabria and, with their help, hereafter of Sicily."

The conditions under which the Norman expedition took place were to leave a profound mark on the history of Sicily. The papal enfeoffment of Robert with Sicily may have been legally contestable—it was naturally not recognized by Constantinople—but it forged a link between the papacy and Sicily that long outlasted Norman rule.

It was Robert's brother, Count Roger I, who as the former's vassal and with his assistance became the real conqueror of Sicily. Internal conflicts among the Arabs served him well. The amir of Syracuse and Catania, at war with his brother-in-law, the amir of Girgenti (Agrigento) and Castrogiovanni (as Enna came to be called), went so far as to offer Roger his help to conquer the island. The first landing (1060), near Messina, was followed by an equally inconclusive attack on that town in 1061; the third, made in greater strength and with the participation of Robert, succeeded (summer, 1061). The possession of Messina gave the Normans control of the straits and a military base for further advance; the capture of Palermo (1072) concluded the first phase of the conquest; and the capture of Noto (1091) completed it. Although Robert's assistance was at first all-important, Roger soon assumed the leading role. After the conquest Robert probably retained only Palermo—apart from the suzerainty over the whole island, which suzerainty became entirely nominal under Roger Borsa, Robert's weak successor as duke of Apulia (1085–1111), who in 1091 surrendered half of Palermo to his uncle.

The Norman capture of Messina

If Roger was the real conqueror of Sicily, he was also the founder of the Norman Sicilian state. By distributing fiefs sparingly, he gave feudalism a less important place in Sicily than it had acquired on the mainland, where Robert Guiscard had established his ducal power by making the Normans accept him as their ruler. At the same time he accepted much of the existing law and institutions. In his treatment of the Arab majority, his policy of religious toleration resembles that previously practiced by the Arabs. Their legal conditions varied considerably, ranging from the liberties that they enjoyed at Palermo, where they had their own quarter and mosques, to the serfdom of the mass of the country population. Roger made use of their military and administrative services; many leading Arabs, however, seem to have left the country after the conquest.

Roger showed much favour to the Greeks, as appears from his lavish patronage of Basilian monasticism. He founded or restored at least 14 Greek monasteries as against four Benedictine ones. The Norman conquest was followed by an increase of the Greek population, Basilian monks from Calabria forming only one element of the new immigration. The Sicilian Church, however, became Latin, according to the promise made by Robert Guiscard to Pope Nicholas II in 1059. But the papacy was left only little influence in it; the concession of the apostolic legation to Roger (1098) made the Sicilian Church practically independent of Rome by sanctioning an already existing state of monarchical control.

Roger II and the foundation of the kingdom. The survival of Roger I's work during the difficult period after his death (1101) was a measure of his success. Roger's son and heir Simon died in 1105 and was succeeded by his brother Roger II; but their mother, Adelaide, ruled as regent from 1101 until Roger attained his majority in 1112. Roger continued his father's efforts to take advantage of the difficulties of the Duke of Apulia (Roger Borsa had been succeeded by his son William in 1111), not only eliminating the last relic of ducal authority on the island by obtaining the second half of Palermo in 1122 but also extending his lands and influence in the duchy of Apulia and in Calabria. After Duke William's death in 1127, Roger II crossed over to the mainland to assert his claims as his heir.

Roger's expedition opened a period of struggles that lasted until 1139. His claims were opposed not only by many lords and towns but also, until 1128, by Pope Honorius II. By supporting the antipope Anacletus II against Innocent II, Roger obtained the royal title for Sicily, Apulia, and Calabria (1130). But this led to an alliance of Innocent II and the emperor Lothair against the ruler who by his support of the antipope was primarily responsible for the prolongation of the schism. Lothair's invasion of the kingdom in 1136 in alliance with Pisa and Venice and, perhaps, Byzantium brought Roger's fortunes to their lowest point. But after Lothair's withdrawal, Roger soon recovered lost ground; and after the death of Anacletus (1138), Innocent II confirmed him in the royal title. The Treaty of Mignano (1139) meant the final acknowledgement by the papacy of the Norman kingdom, which included the mainland provinces of Apulia, Calabria, and Capua as well as Sicily.

The political problems of the kingdom were, to no small extent, a legacy of the different territories of which it was formed. Thus the conflicts with the German and Byzantine emperors belong largely to the history of the southern Italian provinces, while Roger's and his successors' African policy is rooted in the earlier history of Sicily. The first expeditions against the Zirid prince of Mahdia in North Africa (1118, 1123) were failures; but in 1134–35 internal discords in the Zirid state provided Roger with fresh opportunities for intervention, which was facilitated by the growth of the Sicilian navy; the expedition led to the occupation of the island of Djerba. The capture of Tripoli in 1146 initiated a series of conquests that culminated in that of Mahdia in 1148; by that year, Roger's African empire extended from Tripoli to Tunis, from the desert of Barca to Kairouan. That it was short-lived was primarily due to the failure to check the growing power of the Almohads; and the death of the grand admiral George of Antioch, the conqueror of Africa, in 1151–52 and that of Roger II in 1154 jeopardized its survival. The crisis was to begin in 1156 with an Arab rising in Africa; by 1160, the African empire was lost. The Norman kings, however, did not abandon their African ambitions: William II was to send a fleet to lay siege to Alexandria in 1174, and further expeditions were to be undertaken against the North African coast in the following years.

Extension of Roger's African Empire

Expansion in Africa had sharpened the antagonism between Norman Sicily and Constantinople. At first, when the Byzantines had received support from the Western emperors, the alliance of the two empires constituted a formidable threat to the Sicilian kingdom; but the clash between Byzantine and Western imperial claims led, after Frederick I Barbarossa's Roman coronation (1155), to the end of the alliance, Frederick continuing an anti-Norman policy of his own. Roger took the initiative against the Byzantines in 1147 and 1149, seized Corfu, and invaded Greece. The Byzantines, on the other hand, reconquered part of Apulia in 1155, after Roger's death; but they lost the conquered territories again after their crushing defeat at Brindisi (1156). Peace was concluded in 1158 and was not broken until 1185, when King William II invaded the Byzantine Empire, took Durazzo (Durres) and Thessalonica, and advanced toward Constantinople, with the ultimate aim of seizing the Eastern imperial crown. But the counteroffensive under the new emperor Isaac II Angelus put an end to this expedition. It was the last great enterprise of the Norman kings.

Internal development of Norman Sicily. Roger II's internal policy was based on that of his father and was continued by his son William I (1154–66) and by the latter's son William II (1166–89); but his own contribution to the building of the Sicilian state was very great. The combination of Norman, Arabic, and Byzantine elements is perhaps the most striking but at the same time a very natural characteristic of the government and civilization of the kingdom. Roger's Assizes of Ariano (1140) derive largely from Justinian and later Byzantine law; very strong Byzantine influence can be found in the judicial and fiscal administration and in the central financial department, the *duana* (diwan), the personnel of which was originally Arabic. The Norman chancery issued documents in Latin, Greek, and Arabic. On the other hand, the curia was primarily modelled on that of the northern European states; and not only were feudal institutions accepted but feudal barons also played an important part in the provincial and local administration. At the same time, some of the most influential ministers of the Norman period, the grand admirals such as George of Antioch, were Greeks. Roger's capital was cosmopolitan Palermo.

But despite the role played by Greeks and Arabs in the bureaucracy, Sicily became progressively Latinized. While the Muslim population was decreased by conversions, the Latin element was strengthened by the settlement of colonies of "Lombards" (*i.e.,* mainlanders), Greeks and Arabs being gradually reduced to small minorities. This process is also reflected in the large number of Latin monasteries founded by the kings, after the initial favour shown to the Basilian monks. But Sicilian civilization retained the composite character that it had possessed from the beginning; Sicilian scholars translated Greek classical texts into Latin; Idrisi, on the orders of Roger II, composed one of the outstanding works of Arab geography; and Sicilian architecture was the product of Roman, Byzantine, Arabic, and Norman influences.

The Hohenstaufen accession. William II's death (1189) was followed by a struggle for the succession. William I and William II had long supported Pope Alexander III in his conflict with Frederick Barbarossa; but a truce had been concluded in 1177 and the subsequent rapprochement between the King and the Emperor had resulted in William II's sanctioning the betrothal of Frederick's son Henry to Constance, daughter of Roger II and heiress apparent to the kingdom (1184; marriage 1186). This created a situation of great potential danger for the papacy, as it strengthened imperial claims on the kingdom of Sicily by the addition to them of Constance's rights. On William's death, however, there was strong Norman opposition to the prospect of German rule, and Tancred, an illegitimate grandson of Roger II, was crowned king. Frederick's son, who had succeeded his father as Henry VI, failed to make good his claims in 1191, when he was forced to raise the siege of Naples; but the sudden death of Tancred early in 1194 proved a turning point. The reign of his young son, William III, under the regency of Queen Sibylla lasted only 10 months; Henry VI finally occupied the kingdom and was crowned king at Palermo in 1194. A new period in the history of Sicily had begun.

Struggle for succession

Henry aimed at including Sicily permanently in the empire, which he unsuccessfully tried to make hereditary in his family, the House of Hohenstaufen; and the distribution of high offices and lands among his followers was to strengthen his control of the kingdom. His death in 1197 temporarily put an end to imperial domination. Constance, as ruler with her young son, the future emperor Frederick II, returned to a strictly Norman policy. But she died in 1198, leaving Frederick under the guardianship of Pope Innocent III. The following years were marked by growing anarchy, due to the weakness of the monarchy, to the attempts of German lords to seize hold of the kingdom, to the Arabs' endeavours to improve their position, and to the commercial expansion of Pisa and Genoa. After Frederick had reached his majority (1208), Innocent handed the government over to him.

The emperor Frederick II (Frederick I of Sicily). In 1211 Pope Innocent saw himself forced to support Frederick's renewed election as king of the Romans or as German king. Frederick promised Innocent, just before the latter's death in 1216, to renounce the Sicilian kingdom in favour of his son Henry after his imperial coronation; but the papacy's hope to prevent in this way the reunion of Sicily and the empire was not fulfilled, for in 1220 Frederick succeeded in having Henry elected German king. The eventual union of the German and Sicilian crowns, foreshadowed by this election, made papal insistence on Frederick's renunciation obsolete; and when Frederick was crowned emperor in the same year, Honorius III tacitly recognized him also as Sicilian king.

Frederick devoted the following years to the restoration of royal power in the kingdom; and in this formidable task he revealed himself the true successor of Roger II. In Sicily Frederick acted as Norman rather than German ruler: symptomatically, his edict on the resignation of privileges (1220) put the deadline at the death of William II, not of Henry VI. His preoccupation with the affairs of the kingdom was largely responsible for the postponements of his crusade, which led in 1227 to his excommunication. Papal troops invaded the kingdom during his absence in Palestine (1228–29); but Frederick's return meant his im-

The Constitutions of Melfi (1231)

mediate victory, and the Peace of San Germano (1230) was followed by further internal reforms. The Constitutions of Melfi (1231), a legal code inspired by Roger II's Assizes, give a remarkable insight into the organization of the kingdom and also into the political ideas of its ruler. Frederick carried the evolution of the Sicilian administration considerably beyond what had been achieved by the Normans; on the other hand, the use that he made of assemblies of estates, including municipal representatives (from 1232), was a landmark in the history of the Sicilian parliament. In his highly centralized government Frederick was the heir to his Norman predecessors, whose work he continued. The same continuity can also be seen in Sicilian culture, but at Frederick's brilliant cosmopolitan court Italian poetry provided a new formative element.

The end of Hohenstaufen rule. Frederick's work was jeopardized and his dynasty ruined by the union of Sicily with the empire. The conflicts with the papacy and with its allies, the Lombard communes, meant a serious strain on the Sicilian economy and started a chain of events that was finally to lead to the establishment of the House of Anjou in the kingdom. Innocent IV, having declared Frederick deposed from the imperial throne (1245), wanted to deprive him and his descendants of the Sicilian crown as well. Frederick's death in 1250 and that of his son Conrad IV in 1254 provided the papacy with new opportunities, but the success of Manfred, one of Frederick's illegitimate sons, in establishing control over the kingdom even before Conrad's death complicated the situation and gave rise to lengthy and tortuous negotiations in which papal offers of the Sicilian crown to foreign princes alternated with rapprochements to Conrad and to Manfred. Manfred was crowned king at Palermo in 1258—a usurpation of the rights of Conradin, Conrad IV's young son and heir. In 1263, after negotiations with Manfred had broken down for the last time, Pope Urban IV announced the choice of Charles of Anjou, brother of Saint Louis, as king of Sicily (Edmund Lancaster, son of Henry III of England, who had been invested by the papacy with the kingdom in 1255, had failed to substantiate his claims). In 1265 Charles was invested with the kingdom by Pope Clement IV in Rome; near Benevento, in 1266, he defeated Manfred, who was killed.

Charles of Anjou. After Benevento there was no serious resistance to Charles in the kingdom. But his rule had still to stand its test when Conradin came to Italy in 1267 to seize his inheritance. A rising in his favour swept the kingdom and showed the fragility of Charles's position. In 1268, however, Conradin was defeated at Tagliacozzo and executed at Naples. There followed severe suppression of the revolt, especially in Sicily, where it had been more widespread than on the mainland. Charles tried to put his power on a firm foundation by a large-scale distribution of fiefs among his French nobility. At the same time he preserved, in its main outlines, Frederick's system of government. The "French colonization," however, and the sense of grievance that it caused among the Sicilians was partly responsible for the great revolt of 1282, known as the Sicilian Vespers, which severed Sicily once more from the mainland.

The Aragonese. The revolt of the Vespers precipitated the arrival in Sicily of Peter III of Aragon, who had a claim to the crown through his marriage (1262) to the Hohenstaufen heiress Constance, Manfred's daughter. The Sicilians, however, were jealous of their independence: it had been only after the papacy had rebuffed their appeal for recognition of a communal regime of their own that they had addressed themselves to Peter. While Peter was accepted as king (Peter I), Constance governed the island for him until his death in 1285; but then, whereas their eldest son succeeded to Aragon as Alfonso III, the Sicilian crown was transmitted to their second son, who became James I of Sicily. The termination of the personal union between Aragon and Sicily was gratifying to Sicilian separatist feeling, but its effect was prejudiced when Alfonso, having already deserted the Sicilian side in the continuing War of the Sicilian Vespers, died without issue in 1291. In his will he had stipulated that if his brother James were to succeed him in Aragon as James II, he should renounce Sicily to their youngest brother Frederick, in repetition of the previous arrangement; but James, heir to Aragon in his own right, would not at first be bound by this condition and simply nominated Frederick as his lieutenant in Sicily. Subsequently, however, he found that from Aragon's point of view it was indeed expedient to renounce the Sicilian connection; and by the Treaty of Anagni (1295) he agreed to surrender Sicily to the papacy.

The War of the Sicilian Vespers

The Sicilians reacted against this new betrayal by taking Frederick as their king. Though in fact he was Frederick II of Sicily, he took the style of Frederick III.

James II of Aragon allied himself with Charles II of Naples to execute the Treaty of Anagni against the Sicilians, but the allies' efforts came to nothing; and the Treaty of Caltabellotta (1302) brought the 20 years' War of the Vespers to an end: Frederick was to have the kingdom for his lifetime, after which it should revert to the Neapolitan Angevins. Pope Boniface VIII, whose predecessors had steadfastly opposed the Aragonese over Sicily, could only insist that Frederick should style himself "king of Trinacria" (*i.e.,* "of the Three-cornered Isle") instead of "king of the Island of Sicily."

The Angevin-Aragonese peace did not last long. Having taken the emperor Henry VII's side against Robert of Naples and the Guelphs, Frederick resumed the title "king of Sicily" and, in 1314, induced the Sicilian parliament to declare that on his death the crown should pass to his son Peter. Successive Neapolitan attacks on the island failed to reduce him, and on his death his son succeeded him as Peter II (1337).

Frederick's position had obliged him to rely on the support of the Sicilian parliament. He organized it as an assembly of three *bracci*, or houses, representing the feudatories, the clergy, and the towns of the royal domain, on the model of the three estates of the kingdom of Aragon; and he associated it with himself in the exercise of sovereignty. Himself a strong ruler, he was able to keep his barons in order. Under his successors, however, the barons began to assert themselves both in encroachment on the royal authority and in internal warfare of their own; *e.g.,* between the "Latin" faction (the older Sicilian nobility) and the "Catalan" (the Aragonese newcomers) during the reign of King Louis, who was only four years old when he succeeded his father, Peter II, in 1342.

The next king, Frederick III (properly numbered), who succeeded his brother Louis in 1355, managed to withstand renewed attacks by the Neapolitan Angevins (who occupied Messina for a short time in 1356) and finally came to terms with Joan I of Naples in 1373: she agreed that Sicily, officially called Trinacria again, should be a separate kingdom in vassalage both to the Holy See and to her own kingdom of Naples-Sicily.

Frederick III died in 1377, leaving a daughter, Mary, as his heiress. There ensued a long period of disorder. Peter IV of Aragon, on the grounds that the testament of Frederick III (II) precluded female succession to the Sicilian crown, claimed it for himself as the nearest male heir (he was also the father of Mary's mother and the husband of her aunt); and Mary underwent a series of abductions. Peter, however, in the face of objections from the papacy and the Angevins, in 1380 ceded his pretensions to his second son, Martin, duque de Montblanch, whose son Martin was to marry Mary. Peter IV died in 1387, leaving Aragon to his elder son John I; the Queen of Sicily was brought to Spain in 1388; and her marriage to the younger Martin took place in 1390. In 1392 the couple landed in Sicily with Martin of Montblanch and began to reign as queen and king-consort, despite strong local opposition. Mary died in 1401, leaving her widower to reign alone as Martin I of Sicily; but meanwhile the Duque de Montblanch had become king of Aragon as Martin I in 1395 through the death of John I. When Martin I of Sicily died without legitimate issue in 1409, he left his kingdom, with his second wife, Blanche of Navarre, as regent, to his father, who thus became Martin II of Sicily.

Martin II, who had no surviving children of his own, intended that Sicily at least, if not Aragon too, should go to his grandson Fadrique (Frederick) de Luna, a bastard of Martin I of Sicily. On Martin II's death, however, in

1410, this succession was contested; and Ferdinand of Antequera, son of Peter IV's daughter Leonor, having been chosen king of Aragon as Ferdinand I in 1412, defeated Fadrique's partisans and reestablished Blanche's authority as his regent in Sicily. Thenceforward the Aragonese (later the Spanish) and the Sicilian crowns were to remain united for nearly 300 years (until the War of the Spanish Succession).

Alfonso V of Aragon (I of Sicily), having succeeded his father, Ferdinand I, in 1416, used Sicily as his base for his expeditions to Naples during Joan II's precarious tenure of that kingdom; and the Aragonese conquest of Naples after Joan's death temporarily restored the personal union of the Neapolitan and Sicilian crowns. Alfonso, however, kept the two kingdoms theoretically separate, arranging as early as 1443 that his bastard son Ferdinand (Don Ferrante) was to inherit Naples, whereas Sicily passed at his death in 1458 to his brother John II of Aragon (John I of Sicily). John's son Ferdinand II finally recovered Naples for the legitimate line of Aragon (1503).

The dynastic troubles of the period 1337–1412 had further helped the Sicilian feudatories and the towns to extort privileges from the kings as the price of support. The royal lieutenants or viceroys who ruled for the Aragonese kings from 1415 had extensive powers but still had always to reckon with the Sicilian parliament. By the middle of the 15th century it had become an accepted principle of the constitution of the kingdom that no new taxes should be levied without the parliament's consent; and this power of veto over subsidies to the king made the *bracci* a formidable obstacle to royal or viceregal absolutism, as well as to any attempt to reduce feudal, ecclesiastical, or municipal privileges.

The Spanish Habsburgs. Ferdinand II, dying in 1516, left Sicily, with Naples and all his Spanish inheritance, to his grandson, the Austrian Habsburg Charles I of Spain, who three years later became Holy Roman emperor as Charles V. The reign began with a rising (1516) of the privileged orders against the viceroy, Hugo de Moncada, whom Charles recalled without cancelling the edicts that had provoked the trouble. A rising against the new viceroy, Ettore Pignatelli, duque de Monteleone, developed however into widespread outrages against the nobility, which reacted by co-operating with the viceroy. Thereafter the privileged orders in general professed themselves loyal to the distant monarchy, which stood as the guarantor of their privileges, while the viceroys had to bear the burden and the odium of actual government, raising the necessary taxes and trying to keep order in an island still prone to baronial vendetta and to jealous rivalries between towns. The arrival of the Ottoman Turks in the western Mediterranean exposed Sicily to danger from a new quarter, and it became a base for the Emperor's counterattacks on the North African coast.

Rising against the viceroy

In 1555 Charles V abdicated Naples to his son, the future Philip II of Spain. Pope Paul IV then promoted a scheme whereby the French should conquer Naples for a prince of their own house and seize Sicily for the Venetians; but Philip, to whom Charles V had abdicated Spain and Sicily in 1556, forced the Pope to recognize him in 1557 and secured peace from the French in 1559. The Battle of Lepanto (1571) checked the gravest menace from Turkey.

Contributions from Sicily were required for Philip II's enterprises in western Europe and for the defense of Spanish interests in the west and in Italy during Philip III's reign (1598–1621). With the growth of the French challenge to Spain during Philip IV's reign (1621–65) the viceroys had constantly to raise more troops and more money, and the measures to which they turned were resented as oppressive or contrary to the traditional privileges of Sicily. In May 1647, when the viceroy Pedro Fajardo, marquês de Los Vêlez, had put a tax on grain and at the same time tried to keep the old price of loaves by reducing their weight, the people of Palermo rose in revolt and forced him to repeal the tax. Then, in July, the Palermitans, led by Giuseppe Alessio, rose again to demand the repeal of all taxes imposed since Charles V's death and the reservation of the viceroyalty and other public offices to Sicilians; the viceroy withdrew from Palermo and, fearing lest the Sicilian barons and the other towns might join the rebellion, began negotiations with Alessio. The latter, however, lost control over his riotous followers and was assassinated; Messina, always the envious rival of Palermo, remained loyal; the French took no advantage of the situation; and in September the viceroy and his Spaniards returned to the capital, where he died a few weeks later. His successor, faced with another revolt, had to concede a general amnesty. The events of the early summer had largely inspired the parallel revolt of Masaniello in Naples.

The revolt of Messina (1674)

The revolt of Messina in 1674, during the reign of Charles II of Spain, was no less characteristic of Sicily under foreign rule than that of Palermo and was more serious because the French, at war with Spain again, were able to exploit it. The governor of Messina, Luis del Hoyo, wishing to break the stranglehold of the municipal oligarchy over the senate of Messina, tried to exploit popular discontent during a food shortage in order to introduce a popular element into the Senate. The oligarchy stirred up its own popular reaction; Del Hoyo was replaced by Diego de Soria as governor; but conflict between the Senate's and the governor's factions went on until, in August 1674, Soria was besieged in his palace and force to capitulate. The Messinese then appealed to Louis XIV of France, a French fleet arrived in September with supplies of food, and in April 1675 the Duc de Vivonne (Louis Victor de Rochechouart) was sworn in as Louis XIV's viceroy of Sicily. Though Palermo declared itself for Spain against its traditional rival, the whole island might have fallen to the French if the Dutch had not sent a fleet, under the great M.A. de Ruyter, to support their Spanish ally. The naval war around the island was indecisive, and Messina itself held out against the Spaniards until the French, making the Peace of Nijmegen (1678), deserted the Sicilian cause. A viceregal amnesty to the rebels was revoked, and Messina's Senate and privileges were abolished.

The Savoyard and Austrian Habsburg interval. Charles II's death (1700) was followed by the War of the Spanish Succession, during which Sicily's eventual fate was continually in discussion between the belligerent powers. The Austrians overran the Kingdom of Naples (1707); and the Bourbon Philip V of Spain, under pressure from the English, ceded Sicily to Victor Amadeus II of Savoy at the Peace of Utrecht (1713). The Franco-Austrian Peace of Rastatt (1714) left the Austrians in possession of Naples and also gave Sardinia to them, but Spain was not a party to this peace.

The Spaniards, having seized Sardinia in 1717, invaded Sicily in July 1718. Victor Amadeus was unable to defend his kingdom without Austrian help. Austria, however, wanted to exchange Sardinia with him for Sicily; and this exchange had been included in the plan on which the Quadruple Alliance of Great Britain, France, the United Provinces of the Netherlands, and Austria (August 1718) was formed for the settlement of the Austro-Spanish dispute.

The British admiral George Byng (later Viscount Torrington) destroyed the Spanish fleet off Cape Passero; the Austrians crossed the straits into Sicily; Victor Amadeus agreed to the plan for the exchange; and declarations of war against Spain were issued from Great Britain in December 1718 and from France in January 1719. By the Treaty of The Hague (February 1720) Spain accepted the terms of the alliance, and Sicily passed to Austria.

The Bourbons. Extremely disliked in Sicily, Austrian rule there came to an end during the War of the Polish Succession, when the Spanish Bourbon infante Don Carlos, having first conquered Naples and the mainland for himself, was crowned king at Palermo in July 1735. Having initiated his great program of enlightened reform for his two Italian kingdoms, he abdicated Naples and Sicily to his third son, Ferdinand (III of Sicily), when he himself became king of Spain as Charles III in 1759. The antifeudal policy of the viceroy Domenico Caracciolo in the 1780s was resented by the Sicilian baronage, but Ferdinand was able to take refuge in Sicily when he was expelled from Naples by the French Revolutionary army in 1799. On his second expulsion from Naples, during the Napoleonic period, Ferdinand established himself in Sicily again (1806), to remain there, under British protection, for a far longer time. The British occupation, with Lord William Bentinck as ambassador in practical control of affairs from 1811, was chiefly important to Sicily for the constitution of 1812. The old parliament, ever jealous of its right to bargain with the crown, had been obstructing the war effort and so was superseded by a bicameral organ on British lines. Though this "English" constitution went some way to satisfying Sicilian particularism, it offended not only the King but also the more reactionary elements in Sicily insofar as it curtailed their privileges, notwithstanding its aristocratic bias.

The restoration of the Bourbons to Naples (1815) put an end to the constitution; and in 1816 the unitary Kingdom of the Two Sicilies was proclaimed, subjecting Sicily to centralized government from Naples, with the King now styled Ferdinand I. The Sicilians hated this, and, in 1820, Palermo rose in revolt for their constitution of 1812. The contemporary Neapolitan revolution, on the other hand, was for the democratic Spanish constitution of 1812, and troops from Naples suppressed the Sicilian autonomist movement before Naples in turn was put down for Ferdinand by the Austrians.

Ferdinand I was succeeded by Francis I (1825–30), the latter by Ferdinand II (1830–59). When revolution broke out in Sicily in January 1848, the latter promised a constitution, but the Sicilians, acting again in disunion with the mainland, soon declared the deposition of the Bourbon dynasty and offered the crown to a prince of the House of Savoy. After the defeat of the Neapolitan revolution in 1849, Carlo Filangieri reduced Sicily to obedience again in April–May.

The last Bourbon king of Sicily

Francis II was the last Bourbon king of Sicily. Within a year of his accession, Giuseppe Garibaldi landed at Marsala with his "Thousand" on May 11, 1860. A provisional government was formed at Palermo in June, a constitution was proclaimed in August; but after Garibaldi's conquest of Naples and his junction with the Sardinian-Piedmontese army, a Sicilian plebiscite in October decided by an overwhelming majority under universal suffrage for annexation to Sardinia. Messina, which held out for Francis, fell to the Piedmontese in March 1861. Sicily thus became part of the new kingdom of Italy.

The Italian Regione. Despite the plebiscite, the House of Savoy found Sicily no more docile than earlier foreign dynasties had found it. Garibaldi's abortive *pronunciamiento* of 1862 was launched from Sicily and supported by the Sicilians, whom the royal government had to reduce by force. Brigandage and conspiracy continued into the 20th century, feeding on separatist tradition, on Catholic resentment at the government's treatment of the papacy before the Lateran Treaty, and on general discontent at the north's neglect of the backward south. Symptomatic of this state of affairs was the persistence of the Mafia.

After World War II, during which Sicily had become a battleground, the Italian government gave more sympathetic attention to the island's problem. (N.Ru./Ed.)

BIBLIOGRAPHY

General works. *Enciclopedia Italiana,* 36 vol. (1929–37), a basic, if somewhat dated, source; EDITORS OF HOLIDAY, *Italy* (1960), a good, general book mostly for the tourist; TOURING CLUB ITALIANO, *Conosci l'Italia* (1957–), a basic guidebook; see also DENIS MACK SMITH, *Italy: A Modern History* (1959); NINETTA JUCKER, *Italy* (1970); and LUIGI BARZINI, *The Italians* (1964).

Physical and human geography. *The land and people:* DONALD S. WALKER, *A Geography of Italy,* 2nd ed. (1967), a comprehensive text covering regional, economic, physical, and historical aspects; ROBERT E. DICKINSON, *The Population Problem of Southern Italy* (1955), a pioneer study of settlement patterns; CARLO BATTISTI, "La lingua nazionale e le minoranze linguistiche in Italia," in *Archivio per l'Alto Adige,* 54:155–208 (1960).

The economy: G. BARBERO, *Land Reform in Italy* (1961), is a study of economic achievements and future prospects. Italian statistical publications include: *Annuario statistico italiano, Annuario di statistica agraria,* and *Annuario di statistiche industriali*—all issued by the ISTITUTO CENTRALE DI STATISTICA. SHEPARD B. CLOUGH, *The Economic History of Modern Italy* (1964); and VERA C. LUTZ, *Italy: A Study in Economic Development* (1962), provide good background information. DONALD C. TEMPLEMAN, *The Italian Economy* (1981), a study of the 1970s economy in nontechnical language.

Administration and social conditions: Basic works include EMILIO CROSA (ed.), *La Constitution italienne de 1948* (1950); NORMAN KOGAN, *A Political History of Postwar Italy* (1981); ROBERT CHARLES FRIED, *The Italian Prefects: A Study in Administrative Politics* (1963); and ROBERTO ALMAGIA, *L'Italia,* 2 vol. (1959). See also JOSEPH LA PALOMBARA, *Interest Groups in Italian Politics* (1964); and SIDNEY G. TARROW, *Peasant Communism in Southern Italy* (1967).

Cultural life and institutions: CESARE CARAVAGLIOS, *Il folklore musicale in Italia* (1936); and PAOLO TOSCHI, *Arte popolare italiana* (1960), are two excellent Italian sources. Basic references are RENZO FRATTAROLO, *Introduzione bibliografica alla letteratura italiana* (1963); and MARIO PUPPO, *Manuale critico-bibliografico per lo studio della letteratura italiana* (revised annually).

History. *Italy in the early Middle Ages: The Cambridge Medieval History,* 8 vol. (1911–36; 2nd ed., 1966–), contains good coverage of medieval Italy. Primary documentary sources are published in the *Rerum Italicarum Scriptores* by L.A. MURATORI and in the *Monumenta Germaniae Historica.* Also important are the series of the *Historiae Patriae Monumenta,* the *Fonti della storia d'Italia* of the Italian Historical Institute for the Middle Ages, the *Regesta Chartarum Italiae,* and the *Biblioteca della Società storica subalpina.* Still of great use are THOMAS HODGKIN, *Italy and Her Invaders,* 2nd ed., 8 vol. (1892–1916, reprinted 1967); LUDO M. HARTMANN, *Geschichte Italiens im Mittelalter,* 4 vol. (1897–1915, successive editions); and PASQUALE VILLARI, *Le invasioni barbariche in Italia,* 2nd ed. (1905; Eng. trans. of 1st ed., *The Barbarian Invasions of Italy,* 2 vol., 1902). More recent works are ROMOLO CAGGESE, *L'alto Medioevo* (1937); *Il Medioevo,* by various authors, vol. 1 of the *Storia d'Italia,* ed. by the UTET of Turin, 2nd ed. (1965); and GABRIELE PEPE, *Il medio evo barbarico. d'Italia* (1963). Medieval Italy has been put into its European framework in the excellent works of GIOACCHINO VOLPE, *Il Medioevo,* 2nd ed. (1933); GIORGIO FALCO, *La Santa Romana Repubblica,* 5th ed. (1965; Eng. trans., *The Holy Roman Republic,* 2nd ed., 1964); and ROBERTO S. LOPEZ, *Naissance de l'Europe* (1962; Eng. trans., *The Birth of Europe,* 1967). For the social classes and daily life, see ANTONIO VISCARDI and GIANLUIGI BARNI, *L'Italia nell'età comunale* (1966); GIANLUIGI BARNI and GINA FASOLI, *L'Italia nell'alto Medioevo* (1971). For economic history, see WILHELM VON HEYD, ALOYS SCHULTE, and ADOLF SCHAUBE, see ALFRED J. DOREN, *Italienische Wirtschaftsgeschichte,* vol. 1 (1934); FILIPPO CARLI, *Storia del commercio italiano,* 2 vol. (1934–36); GINO LUZZATTO, *Storia economica d'Italia. Il Medioevo,* 2nd ed. (1963); ROBERT S. LOPEZ, "The Trade of Medieval Europe: The South," in *The Cambridge Economic History of Europe,* 2nd ed., vol. 2 (1965). For demography, see JULIUS BELOCH, *Bevölkerungsgeschichte Italiens,* 3 vol. (1937–61). (*Ostrogothic and Byzantine–Lombard periods*): ERNESTO SESTAN, *Stato e nazione nell'Alto Medioevo: Ricerche sulle origini nazionali in Francia, Italia, Germania* (1952); CHARLES H. DIEHL, *Études sur l'administration byzantine dans l'Exarchat de Ravenne, 568–751* (1888); GIAN PIETRO BOGNETTI, *L'età longobarda,* 4 vol. (1966–68); JULES GAY, *L'Italie méridionale et l'empire byzantin . . . , 867–1071,* 2 vol. (1904, reprinted 1960); ARCHIBALD R. LEWIS, *Naval Power and Trade in the Mediterranean, A.D. 500–1100* (1951, reprinted 1970); and NICOLA CILENTO, *Italia meridionale longobarda* (1966). (*Carolingian and post-Carolingian periods*): See the chapters on Italy in LOUIS HALPHEN, *Charlemagne et l'empire carolingien* (1947); and HEINRICH FICHTENAU, *Das Karolingische Imperium* (1949; Eng. trans., *The Carolingian Empire,* 1957); also LOUIS DUCHESNE, *Les Premiers temps de l'État pontifical* (1898; Eng. trans., *The Beginnings of the Temporal Sovereignty of the Popes, A.D. 754–1073,* 1908), a classic, now outdated; and PAOLO BREZZI, *Roma e l'impero medioevale, 774–1252* (1947); ROBERTO CESSI, *Venezia ducale,* vol. 1 (1963); MICHELANGELO SCHIPA, *Il Mezzogiorno d'Italia anteriormente alla monarchia. Ducato di Napoli e Principato di Salerno* (1923); and MICHELE AMARI, *Storia dei Musulmani di Sicilia,* 2nd ed., 3 vol. (1933–39). See also the section on Italy in ROBERT HOLTZMANN, *Geschichte der sächsischen Kaiserzeit, 900–1024,* 3rd ed. (1955); and CHRIS WICKHAM, *Early Medieval Italy: Central Power and Local Society, 400–1000* (1981), a political history.

The High Middle Ages: (*The 11th century*): AUGUSTIN FLICHE, *La Réforme grégorienne,* 3 vol. (1924–37); ERNST WERNER, *Die gesellschaftlichen Grundlagen der Klosterreform im 11. Jahrhundert* (1953). (*Norman Sicily*): JOHN JULIUS NORWICH, *The Normans in the South, 1016–1130* (1967); and *The Kingdom in the Sun, 1130–1194* (1970); FERDINAND CHALANDON, *Histoire de la domination normande en Italie et en Sicile,* 2 vol. (1907); M. CARAVALE, *Il regno normanno di Sicilia* (1966); S. TRAMON-

TANA, *I Normanni in Italia,* vol. 1 (1970). (*The 12th and 13th centuries*): EDOUARD JORDAN, *L'Allemagne et l'Italie aux XII^e^ et XIII^e^ siècles* (1939); PETER RASSOW, *Honor imperii* (1940). The principal contribution to the history of the formation of the autonomous communes is that given by GIOACCHINO VOLPE in *Medio Evo italiano,* 2nd ed. (1928, reprinted 1961) and a whole series of monographs on the Tuscan cities. On the rural communes, and in general on the evolution of society in the countryside, see FEDOR SCHNEIDER, *Die Entstehung von Burg und Landgemeinde in Italien* (1924); and GIAN PIETRO BOGNETTI, *Sulle origini dei comuni rurali del Medio Evo* (1927). Among important works on a single city or region are: VITO VITALE, *Breviario della storia di Genova,* 2 vol. (1955); FRANCESCO COGNASSO, *Storia di Torino* (1934); PIETRO VACCARI, *Pavia nell'alti Medioevo e nell'età comunale* (1956); vol. 3–5 of the *Storia di Milano* of the Fondazione Treccani, Milan (1954–55); PIETRO TORELLI, *Un Comune cittadino in territorio ad economia agricola . . . ,* 2 vol. (1930–52); L. SIMEONI, "Le origini del Comune di Verona," *Nuova Arch.,* vol. 21 (1913); J.K. HYDE, *Padua in the Age of Dante* (1966); GIORGIO CRACCO, *Società e Stato nel Medioevo Veneziano* (1967); DAVID HERLIHY, *Medieval and Renaissance Pistoia* (1967); GAETANO SALVEMINI, *Magnati e popolani in Firenze dal 1280 al 1295* (1899; new ed. 1960 and 1966); and NICOLA OTTOKAR, *Il Comune di Firenze alla fine del dugento* (1926), both on Florence; FERDINAND SCHEVILL, *Siena: The History of a Medieval Commune* (1909, reprinted 1964); ENRICO FIUMI, *Storia economica e sociale di San Gimignano* (1961); GIOACCHINO VOLPE, *Toscana medievale* (1964), including articles on Massa Marittima, Volterra, and Sarzana; and *Studi sulle istituzioni comunali a Pisa,* new ed. (1970); EMILIO CHRISTIANI, *Nobilità e popolo nel comune di Pisa . . .* (1962); DANIEL WALEY, *Mediaeval Orvieto* (1952); PAOLO BREZZI, *Roma e l'impero medioevale, 774–1252* (1947); EUGENIO DUPRE THESEIDER, *Roma dal comune di popolo alla signoria pontificia, 1252–1377* (1952); and ENRICO BESTA, *La Sardegna medievale,* 2 vol. (1908–09). An excellent discussion of the institutions and in general of the communal life is included in DANIEL WALEY, *The Italian City-Republics* (1969). (*The Papal State and Southern Italy*): JOACHIM SEEGER, *Die Reorganisation des Kirchenstaates unter Innocenz III* (1937); DANIEL WALEY, *The Papal State in the Thirteenth Century* (1961); ERNESTO PONTIERI, *Ricerche sulla crisi della monarchia siciliana nel secolo XIII,* 3rd ed. (1958); EDOUARD JORDAN, *Les Origines de la domination angevine en Italie* (1909); EMILE G. LEONARD, *Les Angevins de Naples* (1954); STEVEN RUNCIMAN, *The Sicilian Vespers* (1958); FRANCESCO GIUNTA, *Aragonesi e Catalani nel Mediterraneo,* 2 vol. (1953–59); and J.K. HYDE, *Society and Politics in Medieval Italy: The Evolution of Civil Life, 1000–1350* (1973).

Italy in the late Middle Ages and the Renaissance: Some of the best general accounts of the history of Italy in the 14th and 15th centuries are embedded in treatments of Renaissance culture as a whole. Among such works are DENYS HAY, *The Italian Renaissance in Its Historical Background* (1961); MYRON P. GILMORE, *The World of Humanism, 1453–1517* (1952); and PETER BURKE, *Culture and Society in Renaissance Italy, 1420–1540* (1972). The classic book of JACOB BURCKHARDT (Eng. trans., *The Civilization of the Renaissance in Italy . . . ,* many editions), also continues to be richly suggestive. The structure and the problems of the Italian communes are usefully described in DANIEL WALEY, *The Italian City-Republics* (1969); their revolutionary significance in the medieval political world is discussed in the last section of WALTER ULLMANN, *Principles of Government and Politics in the Middle Ages* (1961). GARRETT MATTINGLY, *Renaissance Diplomacy* (1955), studies both the diplomatic history of Italy in this period and the development of diplomatic institutions and practice. Stimulating on the confrontation between republicanism and despotism in Renaissance Italy and above all on its significance for cultural history and the emergence of the modern political consciousness is HANS BARON, *The Crisis of the Early Italian Renaissance,* rev. ed. (1966). For the crisis of 14th-century society and its effect on the religious life and the mood of Italy, see MILLARD MEISS, *Painting in Florence and Siena After the Black Death* (1951). On Milanese despotism there is a good biography by D.M. BUENO DE MESQUITA, *Giangaleazzo Visconti, Duke of Milan (1351–1402)* (1941). For Florence, FERDINAND SCHEVILL, *History of Florence from the Founding of the City Through the Renaissance* (1936, reprinted 1961), is solid and informative, but the best general introduction to all aspects of Florentine history and society in this period is GENE A. BRUCKER, *Renaissance Florence* (1969); the same author's *Florentine Politics and Society, 1343–1378* (1962) and *The Civic World of Early Renaissance Florence* (1977) are excellent examples of the best specialized studies of Florence in recent scholarship. For Florence under the Medici, see KURT S. GUTKIND, *Cosimo de' Medici, pater patriae, 1389–1464* (1938); and NICOLAI RUBINSTEIN, *The Government of Florence Under the Medici (1434–1494)* (1966). On Venice a good introduction is D.S. CHAMBERS, *The Imperial Age of Venice, 1380–1580* (1970); for the Venetian political tradition, see WILLIAM J. BOUWSMA, *Venice and the Defense of Republican Liberty* (1968); for Venetian society see ANGELO VENTURA, *Nobilità e popolo nella società veneta del '400 e '500* (1964). For Lucca see MARINO BERENGO, *Nobili e mercanti nella Lucca del Cinquecento* (1965); and for Naples see ALAN RYDER, *The Kingdom of Naples Under Alfonso the Magnanimous: The Making of a Modern State* (1976). Some sense of the problems of the Papal State may be derived from PETER PARTNER, *The Papal State Under Martin V* (1958); these problems are vividly illustrated in the memoirs of PIUS II, available in English under the title *Memoirs of a Renaissance Pope* (1959). For Sicily see DENIS MACK SMITH, *A History of Sicily,* 2 vol. (1968). On the events at the end of the 15th century and their impact on Machiavelli and Guicciardini, see FEDERICO CHABOD, *Machiavelli and the Renaissance* (1958), which also contains a useful bibliography for Italy in the age of the Renaissance; and FELIX GILBERT, *Machiavelli and Guicciardini: Politics and History in Sixteenth-Century Florence* (1965). ERIC W. COCHRANE, *Historians and Historiography in the Italian Renaissance* (1981), an excellent overview of historical writing.

Italy in the 16th–18th centuries: An exhaustive bibliography may be found in the appendix to ERNESTO PONTIERI, *Le lotte per il predominio in Europe tra la Spagna e la Potenza ispanoasburgica,* included in vol. 5, pt. 2, of the *Storia Universale* (1971). See also LUIGI SIMEONI, *La Signorie,* vol. 2 (1950); ROMOLO QUASSA, *Preponderanza Spagnuola, 1559–1700,* 2nd ed. (1950); *Storia d'Italia* (*op. cit.*), vol. 2, with an updated bibliography and listing of specific monographs; and LUIGI BULFERETTI, "La decadenza italiana nel seicento," *Cultura e scuola,* 1:98–104 (1962), a review of recent studies. See also the relevant volumes of *The Cambridge Modern History* and *The New Cambridge Modern History.* For a general view of Italy in the 18th century, beyond what may be found in various histories of Italy, see GIORGIO CANDELORO, *Storia dell'Italia moderna,* vol. 1 (1956); GUIDO QUAZZA, *Il problema italiano e l'equilibrio europeo, 1720–1738* (1965); FRANCO VENTURI, *Settecento riformatore* (1969); and GIUSEPPE GALASSO, *Potere e istituzioni in Italia: Dalla caduta dell'Impero romano a oggi* (1974). For Lombardy in particular, see FRANCO VALSECCHI, *L'assolutismo illuminato in Austria e in Lombardia,* 2 vol. (1931–34); S. PUGLIESE, *Condizioni economiche e finanziarie della Lombardia nella prima metà del secolo XVIII* (1924); and FEDERICO CHABOD, *Lo stato e la vita religiosa a Milano nell'epoca di Carlo V* (1971). A basic text is vol. 12 of the *Storia di Milano,* of the FONDAZIONE TRECCANI, Milan, *L'età delle reforme, 1706–1796* (1959). For Tuscany, see NICCOLO RODOLICO, "La Toscana alla vigilia delle riforme," "Emanuele di Richecourt iniziatore delle riforme lorenesi in Toscana" and "I primi provvedimenti contro la manumorta ecclesiastica in Toscana," all included in the collection *Saggi di storia medievale e moderna* (1963), with numerous bibliographical references and references to works by the same author on the Lorraine period. An excellent biography of Peter Leopold is ADAM WANDRUSZKA, *Leopold II* (1963, in German; Italian trans., 1968). For Naples, see HEINRICH BENEDIKT, *Das Königreich Neapel unter Kaiser Karl VI* (1927); RAFFAELE COLAPIETRA, *Vita pubblica e classi politiche del viceregno napoletano, 1656–1734* (1961); RAFFAELE AJELLO, "Il banco di San Carlo: Organi di governo e opinione pubblica nel Regno di Napoli di fronte al problema della ricompra dei diritti fiscali," in *Rivista Storica Italiana,* vol. 81, fascicle 4, pp. 812–881 (1969); LINO MARINI, *Il mezzogiorno d'Italia di fronte a Vienna e a Roma e altri studi di storia meridionale* (1970); GIUSEPPE RECUPERATI, "Napoli e i viceréo austrici, 1707–1784," in *Storia di Napoli,* vol. 7 (n.d.); and HAROLD M. ACTON, *The Bourbons of Naples, 1734–1825* (1956). For Sicily, see RAFF MARTINI, *La Sicilia sotto gli Austriaci (1719–1734)* (1907); FRANCESCO DE STEFANO, *Storia della Sicilia dal secolo XI al XIX,* pp. 231–235 (1948), with extensive bibliography; DENIS MACK SMITH, *Storia della Sicilia medioevale e moderna,* pp. 316–332 (1970); VIRGILIO TITONE, *La Sicilia dalla dominazione spagnola all'unità d'Italia* (1955); GIUSEPPE GALASSO, *Mezzogiorno medievale e moderno* (1965) and *Dal comune medievale all'unità: Linee di storia meridionale* (1969); and the relevant volumes of *The Cambridge Modern History* and *The New Cambridge Modern History.*

Italy from 1789 to 1871: The two most important comprehensive studies are those by CESARE SPELLANZON, *Storia del Risorgimento e dell'Unità d'Italia,* vol. 1–4 (1933–50), which covers the period from the 18th century to 1849 and continues down to the Crimean War with vol. 6–8 by E. DI NOLFO (1959–65); and by GIORGIO CANDELORO, *Storia dell'Italia moderna,* 6 vol. (1956–1970). Spellanzon and Di Nolfo pay particular attention to the political currents and diplomatic history, Candeloro to socio-economic structures. On the Jacobin movement, reference should be made to the interpretations by DELIO CANTIMORI, in *Studi di storia,* pp. 629–638 (1959); and by ARMANDO SAITTA, "La questione del giacobinismo italiano," in *Critica storica,* vol. 4, pp. 204–249 (1965). For the Republic and the Kingdom of Italy, see CARLO ZAGHI, *Napoleone e l'Europa* (1969); and, on the South, PASQUALE VILLANI, "Il Regno di Napoli nel decennio francese (1806–1815)," in *Studi storici in onore di Gabriele Pepe,* pp. 689–702 (1969), which sets forth the plan of a future work. On the various Italian states during the Restoration, the best studies are AUGUSTO SANDONA, *Il Regno Lombardo-Veneto, 1814–1859. La costituzione e l'amministrazione* (1912); KENT R. GREENFIELD, *Economics and Liberalism in the Risorgimento: A Study of Nationalism in Lombardy, 1814–1848,* rev. ed. (1965); REUBEN J. RATH, *The Provisional Austrian Regime in Lombardy-Venetia, 1814–1815* (1969); GAETANO CINGARI, *Mezzogiorno e Risorgimento: La Restaurazione a Napoli dal 1821 al 1830* (1970); and ROSARIO ROMEO, *Il Risorgimento in Sicilia* (1950). On Cavour and his milieu, see ADOLFO OMODEO, *L'opera politica del conte di Cavour (1848–1857)* (1968); ROSARIO ROMEO, *Cavour e il suo tempo, 1810–1842* (1969); and, related to Pius IX, ROGER AUBERT, *Le Pontificat de Pie IX* (1952). On the unification movement, see RAYMOND GREW, *A Sterner Plan for Italian Unity: The Italian National Society in the Risorgimento* (1963); and DENIS MACK SMITH, *Cavour and Garibaldi, 1860: A Study in Political Conflict* (1954). Studies of the formation of the Italian state and the policy of the Right are: CLAUDIO PAVONE, *Amministrazione centrale e amministrazione periferica da Rattazzi a Ricasoli* (1964); ERNESTO RAGIONIERI, *Politica e amministrazione nella storia dell'Italia unita* (1967); ALDO BERSELLI, *La destra storica dopo l'unità,* 2 vol. (1963–65); and ARNALDO SALVESTRINI, *I moderati toscani e la classe dirigente italiana (1859–1876)* (1965).

Italy from 1871 to the present: Comprehensive works or works on individual periods include BENEDETTO CROCE, *Storia d'Italia dal 1871 al 1915,* 2nd ed. (1928; Eng. trans., *A History of Italy, 1871–1915,* 1929, reprinted 1963), a classic of liberal ethico-political historiography; GIOACCHINO VOLPE, *Italia moderna 1815–1915,* 3 vol. (1946–52); DENIS MACK SMITH, *Italy: A Modern History* (1959), on the years 1861–1958; CHRISTOPHER SETON-WATSON, *Italy from Liberalism to Fascism, 1870–1925* (1967), an accurate and concise reconstruction; GIORGIO CANDELORO, *Storia dell'Italia moderna,* vol. 6 (1970), which follows economic and social developments closely for the years 1871–96; FEDERICO CHABOD, *L'Italie contemporaine* (1950); LUIGI SALVATORELLI and GIOVANNI MIRA, *Storia d'Italia nel periodo fascista* (1956); FRANCO CATALANO, *L'Italia dalla dittatura alla democrazia, 1919–1948,* 2nd ed. (1965); NORMAN KOGAN, *A Political History of Postwar Italy* (1966); and GIUSEPPE MAMMARELLA, *L'Italia dopo il fascismo, 1943–1968* (1970). For special topics, see RICHARD A. WEBSTER, *The Cross and the Fasces: Christian Democracy and Fascism in Italy* (1960); RICHARD HOSTETTER, *The Italian Socialist Movement,* vol. 1, *Origins (1860–1882)* (1958); and GAETANO ARFE, *Storia del socialismo italiano, 1892–1926,* 2nd ed. (1965). On the Communist Party, see PAOLO SPRIANO, *Storia del Partito Comunista Italiano,* 3 vol. (1967–70), up to 1939. For the Fascist period, see ANGELO ROSSI, Eng. trans., *The Rise of Italian Fascism 1918–1922* (1938); ADRIAN LYTTELTON, *The Seizure of Power: Fascism in Italy, 1919–1929* (1973); F. CHABOD, *L'Italie contemporaine* (1950; Eng. trans., *A History of Italian Fascism,* 1963); RENZO DE FELICE, *Mussolini,* 4 vol. (1965–74); and F.W.D. DEAKIN, *The Brutal Friendship: Mussolini, Hitler, and the Fall of Italian Fascism,* rev. ed. (1966). For anti-Fascism and the Resistance, see NORMAN KOGAN, *Italy and the Allies* (1956); C.R.S. HARRIS, *Allied Military Administration of Italy, 1943–45* (1957); and CHARLES F. DELZELL, *Mussolini's Enemies: The Italian Anti-Fascist Resistance* (1961). For foreign and colonial policy, see C.J. LOWE and F. MARZARI, *Italian Foreign Policy, 1870–1940* (1975); RENE ALBRECHT-CARRIE, *Italy at the Paris Peace Conference* (1938, reprinted 1966); and J.L. MIEGE, *L'Imperialisme colonial italien de 1870 à nos jours* (1968). On the relations of church and state, see DANIEL A. BINCHY, *Church and State in Fascist Italy* (1941); and A.C. JEMOLO, *Chiesa e Stato in Italia negli ultimi cento anni* (1948; Eng. trans. of the abridged 1955 ed., *Church and State in Italy [1850–1950],* 1960). For economic history, see SHEPARD B. CLOUGH, *The Economic History of Modern Italy* (1964); BRUNO CAIZZI, *Storia dell'industria italiana dal XVIII secolo ai giorni nostri* (1965). On the trade-union movement, see DANIEL L. HOROWITZ, *The Italian Labor Movement* (1963), and M.F. NEUFELD, *Italy: School for Awakening Countries* (1961). See also VICTORIA DE GRAZIA, *The Culture of Consent: Mass Organization of Leisure in Fascist Italy* (1981).

Traditional regions. *Liguria:* M. ALMAGRO, "Ligures en España," in *Rivista de studi liguri,* XV, pp. 195–208 (1949), and XVI, pp. 37–56 (1950); A. BERTHELOT, "Les Ligures," in *Revue archéologique,* pp. 72 ff., 245 ff. (1933); and J. JANNORAY, *Ensérune* (1955). There is no good modern general history of Genoa, since N. LAMBOGLIA *et al., Storia di Genova,* 3 vol. (1941–42), reached only to *c.* 1200. V. VITALE, *Breviario della storia di Genova,* 2 vol. (1955–56), provides an indispensable outline and a guide to the voluminous specialist literature for the period to 1815; for the subsequent period, see his "Genova," in *Enciclopedia italiana,* vol. 16 (1932). Many useful articles and monographs appear in the *Atti della Società Ligure di Storia Patria,* vol. 1 (1861–). See also R.S. LOPEZ, "Market Expansion: The Case of Genoa," *Journal of Economic History,* 24:445–464 (1964); E. BACH, *La Cité de Gênes au XII^e^ siècle* (1955); and J. HEERS, *Gênes au XV^e^ siècle: activité économique et problèmes sociaux* (1961).

Sardinia: MARGARET GUIDO, *Sardinia* (1963), is a scholarly book on Sardinian prehistory, with important illustrations. The section on Sardinia by the ECONOMIST INTELLIGENCE UNIT in *The Mezzogiorno: Investment Prospects for the Seventies* (1971) deals with economics but is rather technical. T. and B. HOLME and B. GHIRADELLI, *Travellers' Guide to Sardinia* (1967), is a short guide for the interested tourist, covering a wide range of subjects; MARY DELANE, *Sardinia: The Undefeated Island* (1968), an account of a journey, includes descriptions of terrain, as well as history and folklore. JOHN WARRE TYNDALE, *The Island of Sardinia,* 3 vol. (1849), is a survey by a barrister dealing in great detail with the history and customs of Sardinians at that time, as does ALBERT DE LA MARMORA, *Voyage en Sardaigne,* 3 pt. (1839–60). D.H. LAWRENCE, *Sea and Sardinia* (1921), is a highly personalized report of people encountered on a brief visit. MARCELLO SERRA, *Mal di Sardegna* (1955), is considered as something of a classic in its analysis of the effects of history and the conditions of life on the island. G.M. TREVELYAN, *Garibaldi and the Thousand* (1909), is of special note for the description of the island of Caprera, where Garibaldi lived and is buried, his tomb guarded always by a sailor. GRAZIA DELEDDA, *Canne al vento* (1958), is a novel demonstrating the continuing belief in witchcraft, as well as a remarkable understanding of the attitudes of the Sards to each other; the author received the Nobel Prize for Literature in 1926 for her descriptions of Sardinian life.

Venetia: The fullest guide-history is G. LORENZETTI, *Venezia e il suo estuario,* 2nd ed. (1956). See also M. MURARO and A. GRABAR, *Treasures of Venice* (1963). Complete accounts covering the whole of the history of Venice include S. ROMANIN, *Storia documentata della Repubblica di Venezia* (1853); H. KRETSCHMAYR, *Geschichte von Venedig* (1905–34); P.G. MOLMENTI, *La storia de Venezia nella vita privata* (1925); and R. CESSI, *Storia della Repubblica di Venezia* (1944); see also the series of lectures relating to the various centuries of the history of Venice in the series *La civiltà veneziana,* ed. by FONDAZIONE "GIORGIO CINI," vol. 1–7 (1955–62).

For studies of individual periods or aspects, see R. CESSI, *Venezia ducale,* 2nd ed. (1940), *Le origini del ducato veneziano* (1952), *La Repubblica di Venezia e il problema adriatico* (1953); C. MANFRONI, *Storia della marina italiana* (1893–1904); F. THIRIET, *La Romanie vénitienne au moyen âge* (1959); P. PINTON, "Veneziani e Longobardi a Ravenna," *Archivio veneto,* vol. 38 (1889); W. LENEL, *Die Entstehung der Vorherrschaft von Venedig an der Adria* (1897); L.M. HARTMANN, "Die wirtschaftliche Anfange Vendigs," *Vierteljahrsschrift für Sozial und Wirtschaftsgeschichte,* vol. 2 (1904), R. HEYNEN, *Zur Entstehung des Kapitalismus in Venedig* (1905); G. LUZZATTO, "I piu antichi trattati tra Venezia e le città marchigiane," *Nuovo archivio veneto,* new series, vol. 11 (1906); D. GHETTI, *I patti tra Venezia e Ferrara dal 1191 al 1313* (1907); V. BRUNELLI, *Storia della città di Zara* (1913); G. PRAGA, *Storia della Dalmazia* (1911); E. BESTA, "La cattura dei venziani in Oriente," *Antologia veneta,* vol. 1 (1900); R. CESSI, "Venezia e la quarta crociata," *Archivio veneto,* vol. 48–49 (1951); G. SORANZO, *La guerra tra Venezia e la S. Sede per il dominio di Ferrara (1308–1318)* (1905); M. BRUNETTI, "Contributo alla storia delle relazioni veneto-genovesi dal 1348 al 1350," *Miscellanea deputazione di storia veneta,* series 3, vol. 10 (1916); V. MARCHESI, *Il patriarcato di Aquileia dal 1394 al 1412* (1894); R. CESSI, *Politica ed economia di Venezia nel trecento* (1952), "La lega italica," *Atti Istituto Veneto,* vol. 102 (1944); B. BELLOTTI, *Bartolomeo Colleoni* (1922); A. LUZIO, *I primordi della lega di Cambrai* (1913); P.G. MOLMENTI, *Sebastiano Venier e la battaglia di Lepanto* (1899); I. RAULICH, *La caduta dei carraresi* (1894); A. LUZIO, "La congiura spagnola contro Venezia nel 1618," *Miscellanea deputazione storia veneta,* series 3, vol. 12 (1918); A. BATTISTELLA, "Una campagna veneto-ispana in Adriatico," *Archivio veneto,* series 5, vol. 2–3 (1928); M. SCIPA, "La pretesa fellonia del duca d'Ossuna," *Archivio storico per le province napoletane,* vol. 35–36 (1910–11); L. DAMERINI, *Francesco Morosini* (1938); G.C. ZIMOLO, "Tre campagne di guerra (1701–1703) e la Repubblica di Venezia," *Archivio veneto,* series 5, vol. 3 (1928); M.M. KOVALEWSKI, *La Fin d'une aristocratie* (1901); E. PESENTI, *Angelo Emo e la marina veneta del suo tempo* (1899); M. PETROCCHI, "Il tramonto della Repubblica di Venezia e l'assolutismo illuminato," *Miscellanea deputazione veneta di storia patria,* vol. 3 (1956); M. BERENGO,

La società veneta alla fine del settecento (1956); G. TABACCO, *Andrea Tron, 1712–1785, e la crisi dell'aristocrazia senatoria a Venezia* (1957); J.C. DAVIS, *The Decline of the Venetian Nobility as a Ruling Class* (1962).

Mezzogiorno: Classic histories of the Kingdom of Naples are: PANDOLFO COLLENUCCIO, *Compendio delle historie del regno di Napoli* (1543; new ed. by A. SAVIOTTI, 1929); ANGELO DI COSTANZO, *Istoria del regno di Napoli* (1572–81; new ed., 3 vol., 1805); G.A. SUMMONTE, *Historie della città e del regno di Napoli* (1602–43); P. GIANNONE, *Istoria civile del regno di Napoli,* 4 vol. (1723; new ed., 8 vol., 1821; Eng. trans. 1729–31); P. COLLETTA, *Storia del reame di Napoli dal 1734 al 1825* (1834; new ed. 1957; Eng. trans. 1860); B. CROCE, *Storia del regno di Napoli* (1925). The periodical *Archivo storico per le provincie napoletane,* 80 vol. (1876–1961), sponsored by the Società Napoletana di Storia Patria, provides bibliographies in its volumes for 1910–14, 1930, 1932, 1938, 1950, and 1960. For medieval sources see B. CAPASSO, *Le fonti della storia delle provincie napoletane dal 568 al 1500,* ed. by E.O. MASTROJANNI (1902). Useful works in English are S. RUNCIMAN, *The Sicilian Vespers* (1958); H. ACTON, *The Bourbons of Naples* (1956) and *The Last Bourbons of Naples* (1961); R.M. JOHNSTON, *The Napoleonic Empire in Southern Italy* . . . , 2 vol. (1904).

Sicily: ALDO PECORA, *Sicilia* (1968), is an informative, general work. Geography and physical description may be found in OLINTO MARINELLI, *Sicilia* (1968); GUSTAVO CUMIN, *La Sicilia* (1944); and FRANCIS M. GUERCIO, *Sicily: The Garden of the Mediterranean, the Country and Its People,* 2nd ed. (1954). Land use and economics are covered in FERDINANDO MILONE, *Sicilia* (1960), on the interaction of nature and man; JOHN P. COLE, *Italy* (1964); VERA C. LUTZ, *Italy: A Study in Economic Development* (1962), with an acute understanding of the south; MARGARET CARLYLE, *The Awakening of Southern Italy* (1962); and NUNZIO PRESTIANNI, *L'economia agraria della Sicilia* (1946), which is technical but interesting. On archaeology, see MARGARET GUIDO, *Sicily: An Archaeological Guide* (1967); L. BERNABO BREA, *La Sicilia prima dei Greci,* 4th ed. (1966; Eng. trans., *Sicily Before the Greeks,* rev. ed., 1966); and the various writings of PAOLO ORSI. See also periodicals and other publications of the Instituto Centrale di Statistica, the Banco di Sicilia, and organs of the Sicilian regional government and of the University of Sicily.

Jainism

Jainism—along with Hinduism and Buddhism—is one of the three most ancient of India's religious traditions still in existence. Its name derives from the Sanskrit verb root *ji,* "to conquer." The name refers to the ascetic battle that the Jaina monks must fight against the passions and bodily senses in order to gain omniscience and the complete purity of soul that represents the highest religious goal in the Jaina system. The monk-ascetic who achieves this omniscience and purity is called a Jina (literally, "Conqueror," or "Victor"), and adherents to the tradition are called Jainas, or Jains. Although Jainism has a much smaller number of adherents than do Hinduism and Sikhism, its influence on India's culture has been considerable, including significant contributions in philosophy and logic, art and architecture, grammar, mathematics, astronomy and astrology, and literature.

Jainism has largely been confined to India, although the migration of Indians to other, predominantly English-speaking countries has spread its practice to many Commonwealth nations and to the United States. Its continuous existence in India for some 2,500 years is in sharp contrast to Buddhism, which is widespread in Asia but no longer widely practiced in the land of its origin. This gives Jainism a unique status as the only Sanskritic non-Hindu religious tradition to have survived in India to the present.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 823.

This article is divided into the following sections:

HISTORY

Mahāvīra

Early history (6th century BC–c. 5th century AD). Jaina history began in the 6th century BC with Vardhamāna, who is known as Mahāvīra ("Great Hero"). Mahāvīra was the 24th and last Tirthankara (literally, "Ford-maker") of the current age (*kalpa*) of the world. (Tirthankaras, also called Jinas, are revealers of the Jaina religious path [*dharma*] who have crossed over life's stream of rebirths and have set the example that all Jainas must follow.) Mahāvīra was a contemporary of Siddhārtha Gautama (the Buddha) and was born in the same area, the lower Gangetic Plain. Although Mahāvīra was a historical figure, all of the accounts of his life are legendary and serve the ritual life of the Jaina community better than they do the historian. However, a little of the historical circumstances of Mahāvīra and the early Jaina community can be pieced together from a variety of sources.

The 6th century BC was a period of intense religious activity in the lower Gangetic Plain. In addition to Buddhism, the Ājīvika sect, founded by Gośāla Maskarīputra, appeared; and at about this time, probably in the same region, the two great "forest" Upanishad texts of early Hinduism, the *Brihadāranyaka* and the *Chāndogya,* came into existence. The prevailing ethos common to all these religious perspectives was asceticism, which stood in contrast to the ritualistic Brahmanic schools associated with the earliest period of classical Hinduism.

Mahāvīra, like the Buddha, was the son of a chieftain of the Kshatriya (military or ruling) class. At age 30 he renounced his princely status to take up the ascetic life. It is likely that he pursued the discipline of a preestablished ascetic tradition and had a reforming influence on it. His acknowledged status as the 24th Tirthankara (or Jina) means that Jainas perceive him as the last revealer in this cosmic age of the Jaina *dharma.* Mahāvīra had 11 disciples (called *gaṇadharas*), all of whom were Brahman converts to Jainism; all founded monastic lineages, but only two—Indrabhūti Gautama and Sudharman, the disciples who survived Mahāvīra—served as the points of origin for the historical Jaina monastic community.

The community appears to have grown quickly—Jaina tradition states that it numbered 14,000 monks and 36,000 nuns at the time of Mahāvīra's death. From the beginning the community was subject to a number of schismatic movements. Jamāli, Mahāvīra's son-in-law, led the first of seven schisms that occurred during the Jina's lifetime. None of these had a significant effect on the Jaina community. The only schism to have a lasting effect was that between the Svetambaras (literally, "White-robed") and the Digambaras ("Sky-clad"; *i.e.,* naked); this division still exists. The major points of difference between the two concern the question of proper monastic attire and whether or not a soul can attain liberation from a female body (a possibility the Digambaras deny).

Each sect has its own account of how the schism arose. The separation appears to have begun as a physical split of the community during the 3rd century BC. According to Digambara tradition, Bhadrabāhu I (whom the Digambaras regard as their founder) foresaw a 12-year famine in the Mauryan kingdom of Candra Gupta and took half of the monastic community south with him to Śravaṇa Beḷgoḷa (near modern Hassan, in Karnātaka state). Digambara tradition also states that Candra Gupta accompanied Bhadrabāhu as his disciple. Svetambara tradition, however, states that Bhadrabāhu went to Nepal and that the Svetambara–Digambara split was led by a monk named Śivabhuti in the last half of the 1st century AD. All differences of doctrine and praxis that developed between the two sects appear to have arisen from this geographical separation.

The four councils

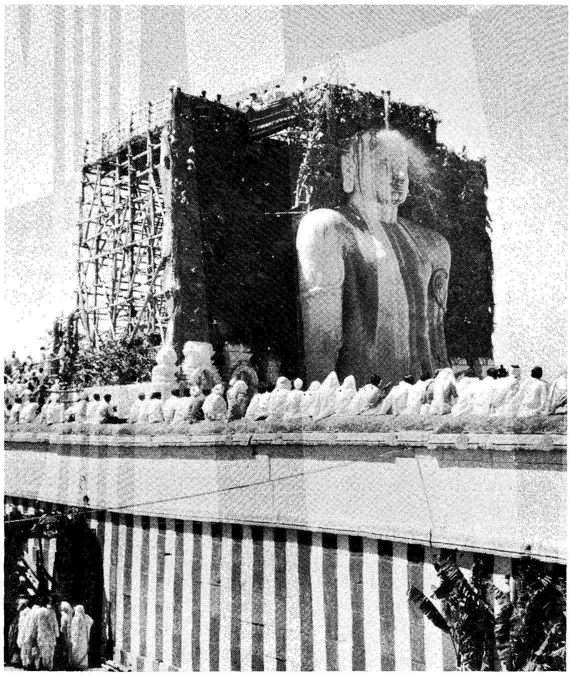
These differences were formalized through a series of councils that met to preserve and codify the teachings of Mahāvīra in written form. It was felt that the teachings, preserved orally since his death, were in danger of being lost. Four councils were held between the 4th century BC and the 5th century AD. The last one, held at Valabhī in Saurāṣṭra (modern Gujarāt state) in either AD 453 or 456, codified the Svetambara canon that is still in use. The Digambara monastic community considered this redaction too corrupt to be normative, and the schism between the two communities became irrevocable.

During this period, Jainism spread from its place of origin westward to Ujjain, where it gained the patronage of Candra Gupta, the grandfather of Aśoka (the last great Mauryan emperor), and later Samprati, the grandson of Aśoka. Later, in the 1st century BC, a monk named Kālakācārya seems to have caused the overthrow of King Gardabhilla of Ujjain and his replacement with the Śāhi kings, who were probably of Scythian or Persian origin. By the time of the Gupta dynasty (AD 320–*c.* 600), Jainas were retaining the patronage of the Gupta emperors of Magadha, but they had become stronger in central and western India than in their homeland.

Early medieval developments (500–1100). The early medieval period was the time of Jainism's greatest flowering, particularly for the Digambara community in the south. The Digambaras gained the patronage of three major dynasties during these centuries—the Gaṅgas in Karnātaka (3rd–11th century); the Rāṣṭrakūṭas, whose kingdom was just north of the Gaṅga realm (8th–12th century); and the Hoysaḷas in Karnātaka (11th–14th century). Digambara monks are reputed to have engineered the succession of the Gaṅga and the Hoysaḷa dynasties, thus stabilizing uncertain political situations and guaranteeing Jaina political protection and support.

Digambara political activity

This involvement in politics on the part of the Digambaras allowed Jainism to prosper in Karnātaka and the Deccan. An abundance of epigraphical evidence details an elaborate patronage system through which kings, queens, state ministers, and military generals endowed the Jaina community with tax revenues and with direct grants for the construction and upkeep of temples. In addition, many of these political figures had Jaina monks as spiritual teachers and advisers. Two notable examples are Śāntala Devī, the queen of the Hoysaḷa king Viṣṇuvardhana, and the Gaṅga general Chāmuṇḍarāya, who in the 10th century oversaw the creation of a colossal statue of Bāhubali (locally called Gomateśvara; son of Ṛṣabha, the first Tirthankara) at Śravaṇa Beḷgoḷa.



Ceremony of anointing the colossal image of the Jaina saint Bāhubali (called locally Gommateśvara) at Śravaṇa Beḷgoḷa, Karnātaka state, India.

James Burke—LIFE Magazine © 1972 Time Inc.

During this period Digambara writers produced a large amount of philosophical treatises, commentaries, and poetry, which was written in Prakrit, Kannada, and Sanskrit. Much of this literary activity had royal patronage and participation. Noteworthy was the monk Jinasena, whose Sanskrit philosophical and poetic writing had the support of the Rāṣṭrakūṭa king, Amoghavarṣa I. Himself an author in Kannada and Sanskrit, Amoghavarṣa seems to have renounced his throne and become a disciple of Jinasena in the early 9th century. This privileged position allowed the Digambara Jainas to engage in sectarian debate from a position of strength. Inscriptions and epigraphs describe many of the most important monks of this period as victors over the Buddhists, Brahmans, Vaiṣṇavites, and Śaivites in philosophical and religious debate.

The Svetambaras seem to have been less flamboyantly embroiled in dynastic politics than their southern counterparts, though there is evidence of such activity in Gujarāt and Rājasthān that helped establish sympathetic kings in the 8th century (Vānarāja, 716–806) and the 12th century (Kumārapāla, whose reign ended with the Muslim invasions). Kumārapāla's accession was masterminded by the great Svetambara scholar and minister of state Hemacandra. The Svetambaras were no less productive in literary output than their Digambara counterparts at this period.

Beginning in the early centuries AD, the role of the Jaina layman was articulated with a detail and precision not seen up to that point. The process began for the Digambaras as early as the 2nd to 3rd century; with the Svetambaras it seems to have begun in the 5th to 6th century. The early medieval period was a time of particularly intense reflection for both groups on the role of the laity. A large Āvaśyaka literature, discussing the layman's religious behaviour and vows, poured forth from these beginnings and lasted until the 17th century. A formalized caste system appeared among the Jaina laity. This was depicted and given authority by Jinasena in his *Ādipurāṇa,* a hagiography of Ṛṣabha and his two sons Bāhubali and Bharata. It differed from the Hindu system in that the Kshatriyas were given a place of prominence over the Brahmans; in addition, the Jainas did not see the caste system as an inherent part of the structure of a created universe. There also were differences in the organization of the caste system between the Digambaras and the Svetambaras.

Late medieval–early modern developments (1100–1800). In the period of their greatest influence (6th–

late 12th century), Jaina monks ceased being wandering ascetics and tended to become dwellers at temples or monastic residences, surrounded by the comforts that their calling demanded they forego. In addition, the Digambara monks' active involvement in dynastic politics undoubtedly earned them enemies. These two factors led to a decline of Jaina influence in ensuing centuries.

Effect of the Muslim invasion

The Svetambara community's eclipse was greatly accelerated by the successful invasion of Muslim forces into western and northern India in the 12th century. With this sudden shift of political control from indigenous to foreign hands, the Svetambara community concentrated on stabilizing itself in the new circumstances. At about this time, the monastic libraries were put underground in Rājasthān to keep the manuscripts from being destroyed and to preserve them better from the elements. There is evidence of Jaina laymen serving as ministers to Muslim rulers, which surely benefited the community.

Reform movements appeared within the community at various times, often stressing the inappropriateness of image worship, especially for monks. This was likely a response to strong Muslim religious values. The most successful of these reform movements was that of the mid-15th-century layman Lonkā Sāha, which led ultimately to the founding of the Sthānakavāsī sect in the 18th century.

By the advent of the Vijayanagar Empire in the 14th century, Digambara Jainism had lost all significant royal support and survived largely by keeping to itself. At this time elaborate temple rituals with Tantric overtones developed within the Digambara community. This, plus the lax attitudes of the administrators of Digambara temple complexes, helped fuel resentment both within and outside the community. This situation made the Digambaras susceptible to attacks by renascent Hindu devotional movements. These movements began in Tamil Nādu as early as the 6th century and in Karnātaka in the 12th century. One of the most vigorous of these Hindu movements was that of the Lingayats, or Vīraśaivas, which arose in full force in the 12th century in northern Karnātaka, a stronghold of Digambara Jainism. The Lingayats gained royal support, and many Jainas themselves converted to the Lingayat religion in ensuing centuries.

Digambara laity were among the foremost critics of their community's deteriorating situation. The most significant Digambara reform movement occurred in the late 16th century, led by a layman and poet named Bānārasīdās. This movement attacked the elaborateness of Digambara ritualism and the cavalier behaviour of its religious leaders.

Recent Jaina history. In modern times, Svetambara Jainism has maintained a more effective organization and has a larger monastic community than its Digambara counterpart. Both communities devote much energy to maintaining temples and publishing critical editions of their religious texts.

In addition, the Jainas stress publicly their deep and longstanding commitment to *ahiṃsā* ("nonviolence"). Notable in this connection is the friendship and exchange of letters between Mohandas Gandhi and the Svetambara layman Raychandrabhai Mehta. Gandhi considered his interactions with Mehta to be important in formulating his own ideas on the use of nonviolence as a political tactic.

Jainas have traditionally been professional and mercantile people. These trades have made them adaptable to other environments and societies besides those of India. Many Jainas have emigrated overseas, and this has had the result of increasing international awareness of Jainism. (G.R.S.)

IMPORTANT FIGURES OF JAINA LEGEND

Sixty-three significant figures form the focus of Jaina legend and story. The most important of these are the 24 Tirthankaras, perfected human beings who appear from time to time to preach and embody the Jaina religious path; they represent the highest religious attainment for the Jaina. The Tirthankaras, along with 12 *cakravartin*s ("world conquerors"), nine *vāsudeva*s (counterparts of Vāsudeva, the patronymic of Krishna), and nine *baladeva*s (counterparts of Balarāma, the elder half-brother of Krishna), constitute a list of 54 *mahāpuruṣa*s ("great souls"), to which were later added nine *prativāsudeva*s (enemies of the *vāsudeva*s). Other, more minor, figures include nine *nārada*s (counterparts of the deity Nārada, the messenger between gods and humans), 11 *rudra*s (counterparts of the Vedic god Rudra, from whom Śiva is said to have evolved), and 24 *kāmadeva*s (gods of love), all of which show Hindu influences. Bāhubali is said to be the first *kāmadeva.* (See also HINDUISM: *Hindu mythology.*)

Subordinated to these figures are the gods, classified into four groups: *bhavanavāsī*s (gods of the house), *vyantara*s (intermediaries), *jyotiṣka*s (luminaries), and *vaimānika*s (astral gods). These, in turn, are divided into several subgroups. Other gods and goddesses also occur in various Jaina texts, such as the 64 *dikkumārī*s (maidens of the directions), who act as nurses to a new-born Tirthankara. Such deities played an important role in ancient Indian folk religion, and the Jainas, Buddhists, and Hindus all assimilated them into their pantheons and rituals.

DOCTRINES OF JAINISM

The Jaina's religious goal is the complete perfection and purification of the soul. This can occur only when the soul is in a state of eternal liberation from and nonattachment to corporeal bodies. Liberation is impeded by the accumulation of *karman*s (see below *Karman*), bits of material, generated by a person's actions, that bind themselves to the soul and consequently bind the soul to material bodies through many births; this has the effect of thwarting the full self-realization and freedom of the soul. To understand how the Jainas perceive and address this problem, however, it is first necessary to explain the Jaina conception of reality.

The notion of the wheel of time

Time and the universe. Time, according to the Jainas, is eternal and formless. It is conceived as a wheel with 12 spokes called *ārā*s ("ages"), six making an ascending arc and six a descending one. In the ascending arc (*utsarpiṇī*), man progresses in knowledge, age, stature, and happiness, while in the descending arc (*avasarpiṇī*) he deteriorates. The two cycles joined together make one rotation of the wheel of time, which is called a *kalpa.*

The world is eternal and uncreated. Its constituent elements, the six substances (*dravya*s), are soul, matter, time, space, and the principles of motion and the arrest of motion. These are eternal and indestructible, but their conditions change constantly.

Jainas divide the inhabited universe into five parts. The lower world (*adholoka*) is subdivided into seven tiers, each one darker and more tortuous than the one above it. The middle world (*madhyaloka*) consists of numberless concentric continents separated by seas, the centre continent of which is called Jambudvīpa. Human beings occupy Jambudvīpa, the second continent, and half of the third; the focus of Jaina activity, however, is Jambudvīpa, the only continent on which it is possible for the soul to achieve liberation. The celestial world (*ūrdhvaloka*) consists of two categories of heaven: one for the souls of those who may or may not have entered the Jaina path, and one for the souls of those who are far along on the path and are close to the time of their emancipation. At the apex of the occupied universe is the *siddhaśīlā,* the crescent-shaped abode of liberated souls (*siddha*s). Finally, there are some areas inhabited solely by *ekendriya*s, organisms that have only a single sense. Although *ekendriya*s permeate all parts of the occupied universe, there are places where they are the only living beings.

Jīva and ajīva. Jaina reality is constituted by *jīva* ("soul," or "living substance") and *ajīva* ("non-soul," or "inanimate substance"). *Ajīva* is divided into two categories: non-sentient and material, and non-sentient and nonmaterial. All but *jīva* are without life.

Characteristics of *jīva*

The essential characteristics of *jīva* are consciousness (*cetanā*), bliss (*sukha*), and energy (*vīrya*). In its pure state, *jīva* possesses these qualities in infinite measure. The souls, infinite in number, are divisible in their embodied state into two main classes, immobile and mobile, according to the number of sense organs possessed by the body they inhabit. The first group consists of souls inhabiting immeasurably small particles of earth, water, fire, and air, plus the vegetable kingdom, which possess only the sense of touch. The second group comprises souls that inhabit bodies that have between two and five sense organs. The Jainas believe that the four elements (earth, water, fire, and air) also are animated by souls. Moreover, the universe is full of an infinite number of minute beings, *nigoda*s, which are slowly evolving.

A *jīva* is formless and genderless and cannot be perceived by the senses. A soul is not all-pervasive, but can, by contraction or expansion, occupy various amounts of space. Like the light of a lamp in a small or a large room, it can fill both the smaller and larger bodies it occupies. While the soul assumes the exact dimensions of the body it occupies, it is not identical with that body.

Matter (*pudgala*) has the characteristics of touch, taste, smell, and colour. Its essential characteristic is lack of consciousness. The smallest unit of matter is the atom (*paramāṇu*). Heat, light, and shade are forms of fine matter.

The non-sentient, nonmaterial substances are the principles of motion and its arrest, space, and time. They are always pure and are not subject to defilement. The principles of motion and its arrest permeate the universe; they do not exist independently but, rather, form a necessary precondition for any object's movement or coming to rest. Space is infinite, all-pervasive, and formless and provides accommodation for the entire universe. It is divided into occupied (*i.e.,* the universe) and unoccupied portions. Time is said to consist of innumerable eternal and indivisible particles of "non-corporeal substance" that never mix with one another but that fill the entire universe. Thus, the non-sentient, nonmaterial substances form the context within which the drama of a *jīva*'s struggle to extricate itself from involvement with matter occurs.

Karman. The fundamental tenet of Jaina doctrine is that all phenomena are linked together in a universal chain of cause and effect. Every event has a definite cause behind it. By nature each soul is pure, possessing infinite knowledge, bliss, and power; however, these faculties are restricted from beginning-less time by foreign matter coming in contact with the soul. Fine foreign matter producing the chain of cause and effect, of birth and death, is *karman,* a fine atomic substance and not a process as in Hinduism. To be free from the shackles of *karman,* a person must stop the influx of new *karman*s and eliminate the acquired ones.

Karmic particles are acquired as the result of intentional action tinged with passionate expression. Acquired *karman*s can be annihilated through a process called *nirjarā,* which consists of fasting, not eating certain kinds of food, control over taste, resorting to lonely places, mortifications of the body, atonement and expiation for sins, modesty, service, study, meditation, and renunciation of the ego. *Nirjarā* is, thus, the calculated cessation of passionate action.

Stages of the soul's growth

A soul passes through various stages of spiritual development before becoming free from all karmic bondages. These stages of development (*guṇasthāna*s) involve progressive manifestations of the innate faculties of knowledge and power and are accompanied by decreasing sinfulness and increasing purity.

*Jīva*s become imprisoned in a succession of bodies due to their connection with karmic matter. These embodied souls bear different colours or tints (*leśyā*), varying according to the merits or demerits of the particular being. This doctrine of *leśyā,* peculiar to Jainism, seems to have been borrowed from the Ājīvika doctrine of six classes of bodies, expounded by Gośāla Maskarīputra. The six *leśyā*s in Jainism are, in the ascending order of man's spiritual progress, black, blue, gray, fiery red, lotus-pink (or yellow), and white.

Theories of knowledge as applied to liberation. In Jaina thought, four stages of perception—observation, will to recognize, determination, and impression—lead to a subjective cognition (*matijñāna*), the first of five kinds of knowledge (*jñāna*). The second kind of knowledge is *śrutajñāna,* derived from the scriptures and general information. Both of these are mediated cognition, based on external conditions perceived by the senses. There are three kinds of immediate knowledge—*avadhi* (supersensory perception), *manaḥparyāya* (reading the thoughts of others), and *kevala,* which is the stage of omniscience. *Kevala* is necessarily accompanied by freedom from karmic obstruction and by direct experience of the soul's pure form unblemished by its attachment to matter. Omniscience is the foremost attribute of a liberated *jīva,* the emblem of its purity; thus, a liberated soul, such as a Tirthankara, is called a *kevalin* ("possessor of omniscience").

According to Jainism, *yoga,* the ascetic physical and meditative discipline of the monk, is the means to the attainment of omniscience, and thus to *mokṣa,* or liberation. *Yoga* is the cultivation of true knowledge of reality, faith in the teachings of the Tirthankaras, and pure conduct; it is, thus, intimately connected to the three jewels (*ratnatraya*) of right knowledge, right belief, and right conduct (respectively, *samyagjñāna, samyagdarśaṇa,* and *samyakcāritra*). (See INDIAN PHILOSOPHY; LOGIC, THE HISTORY AND KINDS OF.)

Jaina ethics. The *ratnatraya* constitute the basis of Jaina ethics. Right knowledge, faith, and conduct must be cultivated together; none of them can be achieved in the absence of the others. Right faith leads to calmness or tranquillity, detachment, kindness, and the renunciation of pride of birth, beauty of form, wealth, scholarship, prowess, and fame. Right faith leads to perfection only when followed by right conduct. Yet, there can be no virtuous conduct without right knowledge, which consists of clear distinction between the self and the nonself. Knowledge of scriptures is distinguished from inner knowledge. Knowledge without faith and conduct is futile. Without purification of mind, all austerities are mere bodily torture. Right conduct is thus spontaneous, not a forced mechanical quality. Attainment of right conduct is a gradual process, and a householder can observe only partial self-control; when he becomes a monk, he is further able to observe more comprehensive rules of conduct.

Concept of *ahiṃsā*

Two separate courses of conduct are laid down for the ascetics and the laity. In both cases, the code of morals is based on the doctrine of *ahiṃsā,* or nonviolence. Since thought gives rise to action, violence in thought merely precedes violent behaviour. Violence in thought, then, is the greater and subtler form of violence, because it arises from ideas of attachment and aversion, grounded in passionate states, which result from negligence or lack of care in behaviour. Jainism enjoins avoidance of all forms of injury, whether committed by body, mind, or speech.

RITUAL PRACTICES AND RELIGIOUS AND SOCIAL INSTITUTIONS

The monks and their practices. Svetambaras acknowledge two classes of monks: *jinakalpin*s, who wander naked and use the hollows of their palms as alms bowls; and *sthavirakalpin*s, who retain minimal possessions such as a robe, an alms bowl, a whisk broom, and a *mukhavastrikā* (a piece of cloth held over the mouth to protect against the ingestion of small insects). A monk must obey the "great vows" (*mahāvrata*s) to avoid injuring any life form, lying, stealing, having sexual intercourse, or accepting personal possessions. To help him live out his vows, a monk's life is carefully regulated in all details by specific ordinances and by the oversight of his superiors. For example, to help him observe the vow of noninjury, a monk may not take meals after dark, since to do so would increase the possibility that he would harm any insects that might be attracted to the food. Monks are expected to suffer with equanimity such hardships as those imposed by the weather, geographical terrain, travel, or physical abuse. Exceptions are allowed in emergencies, since a monk who survives a calamity can purify himself by confession and by practicing even more rigorous austerities.

Among the Digambaras, a full-fledged monk remains naked, though there are lower-grade monks who wear a loincloth and keep with them one piece of cloth not more than one and one-half yards long. Digambara monks use a peacock-feather duster and water gourd, live apart from human habitations, and beg and eat only once a day, using the palm of one hand as an alms bowl.

Eight essentials noted for the conduct of monks include the three *gupti*s (care in thought, speech, and action) and the five *samiti*s (kinds of vigilance over conduct). The six *āvaśyakas*, or obligations, are equanimity; praise of the Tirthankaras (Jinas); obeisance to the Jinas, teachers, and scriptures; atonement; resolution to avoid sinful activities; and meditation.

The act of *sallekhanā*

The type of austerities in which a monk engages, the length of time he engages in them, and their severity are carefully regulated by his preceptor, who takes into account the monk's spiritual development, his capacity to withstand the austerities, and his ability to understand how they help further his spiritual progress at a given time. The culmination of a monk's ascetic rigours is the act of *sallekhanā,* in which a monk lies on one side on a bed of thorny grass and ceases to move or take food. This act of ritual starvation is the monk's ultimate act of nonattachment, in which he lets go of the body for the sake of his soul. The ascetic's preparatory rigours, which point to and culminate in this act, generally take 30 years or more. While it is a tenet of Jaina doctrine that no one can achieve liberation in this corrupt time, it is thought that the act of *sallekhanā* still has value since it can improve a soul's spiritual situation in the next birth.

Religious disciplines of the laity. The life of a lay votary is a preparatory stage to the rigours of ascetic life. The lay votary is enjoined to observe eight primary behavioral qualities (which vary but usually include the avoidance of meat, wine, honey, fruits, roots, and night eating) and 12 vows: five *aṇuvratas* ("little vows"), three *guṇavratas*, and four *śikṣāvratas*. The *aṇuvratas* are vows to abstain from gross violence, falsehood, and stealing; to be content with one's own wife; and to limit one's possessions. The other sets of vows are supplementary in nature, meant to strengthen and protect the *aṇuvratas*. They involve avoidance of unnecessary travel, harmful activities, and the pursuit of pleasure; fasting and control of diet; offering of gifts and service to monks, the poor, and fellow believers; and voluntary death if the observance of vows proves impossible.

The *sāmāyika,* a lay meditative and renunciatory ritual of limited duration, aims at strengthening equanimity of mind and resolve to pursue the spiritual discipline of the Jaina *dharma.* This ritual brings the lay votary close to the demands required of an ascetic for a limited time. It may be performed in a person's own house, in a temple, in a fasting hall, or before a monk.

Eleven *pratimās*, or stages of a householder's spiritual progress, are listed. Medieval writers conceived *pratimā* (literally, "statue") as a regular progressing series, a ladder leading to higher stages of spiritual development. The last two stages lead logically to renunciation of the world and assumption of the ascetic life.

The disciplines to which Jaina laity must adhere have influenced significantly the types of vocations that they pursue. Since all of their actions should minimize acts of violence to other living creatures, Jainas tend to pursue commercial and professional enterprises and to avoid such careers as military service. This has created an ironic situation in which many adherents to a highly austere and ascetic religion are wealthy.

Sacred times and places. *Festivals and fairs.* The principal Jaina festivals are connected with the five major events in the life of each Tirthankara. These mark the occasions of the Tirthankara's descent into his mother's womb, birth, renunciation, attainment of omniscience, and final emancipation.

The festival of Paryuṣaṇa

The most popular Jaina festival is Paryuṣaṇa, or Paijusaṇa, which occurs in the month of Bhādrapada (August–September). Paryuṣaṇa literally means (1) pacification by forgiving and service with wholehearted effort and devotion and (2) staying at one place for the monsoon season. On the last day of the festival, Jainas distribute alms to the poor and take a Jina image in procession through the streets. Confession is performed during the festival to remove all ill feelings about conscious or unconscious misdeeds during the past year.

Twice a year, for nine days (March–April and September–October), a fasting ceremony known as *olī* is observed. These are also the eight-day festivals corresponding to the mythical celestial worship of images of the Jinas.

On the full-moon day of the month of Kārttika (October–November), at the same time that Hindus celebrate Dewali (festival of lights), Jainas commemorate the Nirvana of Mahāvīra by lighting lamps. Five days later is Jñānapañcamī (literally, "The Fifth Level of Knowledge," *i.e., kevala*), which the Jainas celebrate with temple worship and with worship of the scriptures. Mahāvīra Jayanti, the birth date of Mahāvīra, is celebrated in early April.

The Jainas also celebrate a number of festivals in common with Hindus, such as Holi (spring festival), Navaratra (nine nights festival), and Poṅgal (a South Indian spring festival).

Pilgrimages and shrines. The erection of shrines and the donation of religious manuscripts are regarded as pious acts. Most villages or towns inhabited by Jainas have at least one Jaina shrine; some have become pilgrimage sites. Lists of these shrines have been composed, and the most noteworthy shrines are offered adoration in daily worship.

Places of pilgrimage were created at sites marking the principal events in the lives of Tirthankaras. Parasnāth Hill and Rājgīr in Bihār and Śatruñjaya and Girnār hills on the Kāthiāwār Peninsula are among such important ancient pilgrimage sites. Other shrines that have become pilgrimage destinations are Śravaṇa Beḷgoḷa in Karnātaka, Mounts Abu and Kesariajī in Rājasthān, and Antarikśa Pārśvanātha in Akola district, Mahārāshtra.

Cave temples

Several Jaina cave temples, dating from as early as the 2nd century BC, have been discovered and excavated. Cave temples are found at Udayagiri and Khandagiri, in Orissa; Rājgīr, in Bihār; Aihole, in Karnātaka; Ellora, in Mahārāshtra; and Sittānnavāsal in Tamil Nādu.

Temple worship and observance. Temple worship is mentioned in early texts that describe gods worshiping Jina images and relics in heavenly eternal shrines. Worship, closely associated with the obligatory rites of the laity, is offered to all liberated souls, to monks, and to the scriptures. Though Tirthankaras remain unaffected by offerings and worship, such actions serve as a form of meditative discipline for the votary offering them. Daily worship includes recitation of the names of the Jinas and idol worship by bathing the image and making offerings to it. Svetambaras decorate images with clothing and ornaments. The worshiper also chants hymns of praise and prayers and mutters sacred formulas. Such Jaina rituals show considerable similarity in form to Hindu rituals. A long-standing debate within both Jaina communities over the centuries has concerned the relative value of external acts of worship and internalized acts of mental discipline and meditation.

Domestic rites and rites of passage. Early Jaina literature is silent about domestic rites and rites of passage marking the main events in a person's life. These rituals are modeled mainly on the 16 Hindu samskaras, which include conception, birth, naming, first meal, tonsure, investiture with the sacred thread, beginning of study, marriage, and death. They are first discussed in Jinasena's 9th-century work, *Ādipurāṇa.*

Welfare institutions. Jainas are renowned for various types of munificence, such as sponsoring pilgrimages, famine relief, relief to Jaina widows and the poor, and maintaining shelters for old animals to save them from slaughter (an act of *ahiṃsā*). In addition, Jainas have encouraged research in and publication of editions of Jaina canonical and commentarial texts. Noteworthy in this connection are the Bhāratīya Jñānapīṭha publishing house in Vārānasi, Uttar Pradesh, and Lalbhai Dalpatbhai Institute for Indological Research at Ahmadābād, Gujarāt.

JAINA LITERATURE

Canonical and commentarial literature. Jaina canonical scriptures do not belong to a single period, nor is any text free from later revision or additions. The sacred literature, preserved orally from the time of Mahāvīra, was first systematized in a council at Patna about the end of the 4th century BC, and again in two later councils at Mathura and Valabhī in the early 3rd century AD. The fourth and last council, at Valabhī in the mid-4th century, is considered the source of the existing Svetambara canon, though some commentators insist that the present reading is in accordance with the Mathura council.

The original, unadulterated teachings of the Jinas are said to be contained in 14 texts, called the Pūrvas ("Foundation"), which are now lost. Svetambaras and Digambaras agree that a time will come when the teachings of the Jinas will be completely lost; Jainism will then disappear from the earth and reappear at an appropriate point in the next time cycle (*kalpa*). The two sects disagree, however, about the extent to which the corruption and loss of the Jinas' teachings has already occurred. Consequently, the texts for each sect differ.

Svetambara canon

The Svetambaras follow an extensive canon (*āgama*) as the repository of their tradition, which they believe is based upon compilations of Mahāvīra's discourses by his disciples. This canon preserves the teachings of Mahāvīra in an imperfect way, as it is thought to be mixed with much that was not said by the Jina. The number of texts considered to make up the Svetambara canon has varied over time and by monastic group. Largely through the influence of the 19th-century German scholar Johann Georg Bühler, however, Western scholars have fixed the number of texts in this canon at 45, divided into six groups: the 11 Aṅgas ("Parts"; originally there were 12, but one, the Dṛṣṭivāda, has been lost), 12 Upāṅgas (subsidiary texts), four Mūla-sutras (basic texts), six Cheda-sutras (concerned with discipline), two Cūlikā-sutras (appendix texts), and 10 Prakīrṇakas (mixed, assorted texts). The Aṅgas contain several dialogues, mainly between Mahāvīra and his disciple Indrabhūti Gautama, presumably recorded by the disciple Sudharman, who transmitted the teachings to his own disciples.

According to modern scholars, the Ācārāṅga and the Sūtrakṛtaṅga, among the Aṅgas, and the Uttarādhyayana, among the Mūla-sutras, are among the oldest parts of the canon. The Cheda-sutra text, Daśāśrutaskandha, concludes with the Kalpa-sutra, which recounts the lives of the Jinas and includes an appendix of rules for monastic life and a list of eminent monks.

Bhadrabāhu, whom tradition credits with being the last Jaina sage to know the contents of the Pūrvas, is asserted to be the author of the Niryuktis, the earliest commentaries on the Jaina canonical texts. These concise, metrical commentaries, written in Prakrit, gave rise to an expanded corpus comprising texts called Bhāṣyas and Cūrṇis. These were composed between the 4th and 7th centuries and contain many ancient Jaina historical and legendary traditions, along with a large number of popular stories brought into the service of Jaina doctrine. The Bhāṣyas and Cūrṇis, in turn, gave rise in the medieval period to a large collection of Sanskrit commentaries. Haribhadra, Sīlāṅka, Abhayadeva, and Malayagiri are the best-known authors of such commentaries.

Digambara canon

Digambaras give canonical status to two works in Prakrit: the Karmaprābhṛta ("Chapters on Karman," also called Ṣhaṭkhaṇḍāgama) and the Kaṣāyaprābhṛta ("Chapters on the Kaṣāyas"). The Karmaprābhṛta, based on the now-lost Dṛṣṭivāda text, deals with the doctrine of *karman* and was committed to writing by Pushpadanta and Bhūtabalin in the mid-2nd century; the Kaṣāyaprābhṛta, compiled by Guṇadhara from the same source at about the same time, deals with the passions that defile and bind the soul. Later commentaries by Vīrasena (8th century) and his disciple Jinasena (9th century) on the Kaṣāyaprābhṛta are also highly respected by Digambaras.

Philosophical and other literature. In addition to the canons and commentaries, the Svetambara and Digambara traditions have produced a voluminous corpus of literature, written in several languages, in the areas of philosophy, poetry, drama, grammar, music, mathematics, medicine, astronomy, astrology, and architecture. In Tamil, the epics *Cilappatikāram* and *Jīvikaciṇṭāmaṇi*, which are written from a Jaina perspective, are important works of early postclassical Tamil literature. Jaina authors were also an important formative influence on Kannada literature. The *Ādipurāṇa* of the Jaina lay poet Pampa (another text dealing with the lives of Ṛṣabha, Bāhubali, and Bharata) is the earliest extant piece of *mahākāvya* ("high poetic") Kannada literature. Jainas were similarly influential in the Prakrit languages, Apabhramsa, Old Gujarati, and, later, Sanskrit. A particularly important literary figure in Prakrit and Sanskrit was the Svetambara monk Hemacandra (12th century), who composed an important Prakrit grammar, as well as poetry, philosophical treatises, and a mammoth epic poem on the lives of the 63 Jaina *mahāpuruṣa*s, entitled *Triṣaṣṭiśalākāpuruṣacaritra.*

Other noncanonical Jaina writers on philosophy include Mallavādin I (4th century), Siddhasena Divākara (*c.* 5th century), Haribhadra Sūri (*c.* 8th century), Samantabhadra (before the 5th century), Akalaṅka (*c.* 8th century), Siddharṣi Gaṇin (10th century), Śāntisūri (11th century), Vidyānandin (*c.* 8th–9th century), Anantakīrti (10th century), Māṇikyanandin (11th century), Prabhācandra (11th century), and Vādi Deva Sūri (12th century). Among later authors, Upādhyāya Yaśovijaya (*c.* 17th century), a versatile scholar, is especially noteworthy.

Digambaras also value the Prakrit works of Kuṇḍakuṇḍa (*c.* 2nd century), including the *Pravacanasāra* (on ethics), the *Samayasāra* (on fine entities), the *Niyamasāra* (on Jaina monastic discipline), and the six *Prābhṛta*s ("Chapters") on various religious topics. Of similar importance is the *Tattvārthādhigama-sutra* of Umāsvāmin (or Umāsvāti), whose work is claimed by both communities. Composed early in the Christian Era, the *Tattvārthādhigama-sutra* was the first work in Sanskrit on Jaina philosophy dealing with such subjects as logic, epistemology, ontology, ethics, cosmography, and cosmogony; it generated numerous commentaries, including one by Umāsvāti himself.

RELIGIOUS SYMBOLISM AND ICONOGRAPHY

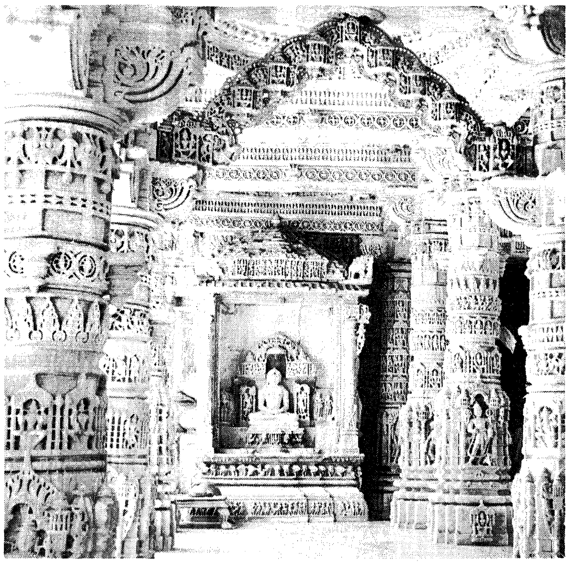
Image worship was introduced at an early stage, perhaps even during the century immediately following the death of Mahāvīra. The Jina himself appears to have made no statement regarding the worship of images. Descriptions of *stūpa*s (reliquaries for the bones and ashes of saints), commemorative pillars, and tree shrines appear in early Jaina texts, which also refer to the worship in the heavens by gods of images of the four legendary Śāśvata Jinas ("Eternal Victors") and of costly relic boxes. Mention is made of *śilāpaṭa*s, which apparently were stone plaques or reliefs placed on lion thrones underneath trees, such as those associated with the worship of Yakshas (mythical nature spirits), and also depicted on Buddhist reliefs from Bharhut (2nd century BC). The *śilāpaṭa*s appear to be the prototypes of the later Jaina *āyāgapaṭa*s (tablets of homage) from Mathura (Uttar Pradesh state), which show representations of *stūpa*s, *caitya* pillars surmounted by elephants, *dharmacakra*s (wheels of the law), and the *aṣṭamaṅgala*s (eight auspicious symbols). Later *āyāgapaṭa*s show a Jina attended by two nude disciples and the figure of the monk Kaṇha Samaṇa with his disciples, or they depict the figure of a noblewoman with attendants.

*Āyāgapaṭa*s

The earliest extant Tirthankara image is possibly the highly-polished Mauryan period torso from Lohanipur, near Patna. Numerous Tirthankara images in the sitting and standing postures dating from the early Christian Era have been uncovered in excavations of a Jaina *stūpa* at Mathura. The earliest images of Tirthankaras are all nude. The various Jinas are distinguished by inscriptions giving their names carved on the pedestals, but later iconographic devices such as symbols specific to each Jina did not evolve until about the 5th century.

Worship of the 16 principal Jaina Tantric goddesses, the Mahāvidyās, was probably introduced in the Gupta age. From the 6th to the 11th century a common pair of attendants was employed in sculpture for all the Tirthankaras, but from about the 9th century 24 *śāsanadevatā*s were evolved, each one to attend a different Tirthankara. The names of many of the attendants suggest Hindu or Buddhist influence.

The religious merit that accrues from hearing and reading Jaina texts encouraged the careful and loving preservation of illustrated manuscripts. The miniature paintings on palm-leaf and paper manuscripts preserved in the Jaina monastic libraries provide a continuous history of the art of painting in western India from the 11th century to the present. The lives of the Jinas and legends of Jaina saints provide a framework for the artists to depict gods and goddesses, throne rooms and village interiors, gardens, and temples. Religious symbols such as the *ashṭamaṅgala*s



Interior of a Jaina temple, showing the Tirthankara image enshrined. Dilwāra Temple, Mount Abu, India, 13th century AD.

By courtesy of the Government of India Tourist Office, London

and the 14 dreams of the mothers of the Tirthankaras frequently appear in paintings.

In addition to the miniatures and to painted wooden book covers that often show mythological scenes, paintings on cloth are also known. Wall paintings are found on cave shrines at Sittānavāsal (Tamil Nādu state) and at Ellora.

Jaina temples generally contain a number of metal images of various types and metal plaques showing auspicious symbols. Metal images of the Jinas are also kept by pious Jainas for home devotion. Among the earliest known bronzes are one of Pārśvanātha in the Prince of Wales Museum of Western India in Bombay, which may date from the 1st century BC, and a group of bronzes (1st–3rd century AD) from Chausa in Bihār in the Patna Museum.

JAINISM AND OTHER RELIGIONS

Shared key concepts

Jainism, Hinduism, and Buddhism. Jainism, Hinduism, and Buddhism share a discourse made available through the Sanskrit language and the dialects (Prakrits) derived from it. Having a set of key concepts in common has enabled these traditions to finely hone their religious debates. For example, all three traditions share a notion of *karman* as the actions of individuals that determine their future births; yet, each has attached connotations to the concept that are uniquely its own. This is also true with terms such as *dharma* (often translated "duty," "righteousness," or "religious path"), *yoga* ("ascetic discipline"), and *yajña* ("sacrifice," or "worship"). This Sanskritic discourse has been brought into the service of the religious and philosophical speculations, as well as the polemics, of each of these traditions.

The same circumstance occurs in the ritual life and literature of each religion. In the ritual sphere, for example, the *abhiṣeka,* or head-anointing ritual, has had great significance among all three, especially in royal contexts. The best-known example of this ritual is the one performed every 12 to 14 years on the statue of Bāhubali at the Jaina pilgrimage site at Śravaṇa Beḷgoḷa. The structure of this ritual is similar in each religious context; in each case, however, it has specific meanings peculiar to that context.

In the literary sphere, each tradition developed an extensive corpus of canonical and commentarial literature, and each has developed a body of narrative literature. For example, so great was the influence of the story of Rāma in the classical Hindu *Rāmāyaṇa,* that the Buddhists and Jainas felt obliged to retell the story in their own terms. Jaina literature includes 16 different tellings of this story in Sanskrit and Prakrit.

Finally, each tradition shares a similar understanding of the ascetic life, though each understands it as functioning properly only within the context of its own religious system. Many of the terms applied to figures in each monastic organization are the same (though not necessarily the same in meaning), and several of the monastic ritual and meditative activities are similar in structure.

Jainism and Islām. In reference to Muslim influence on Jainism, it has been suggested that the concept of *āśātanās*—activities that are unfitting or indecent in a temple—reveals a notion of the sanctity of the temple that is more evocative of Muslim *barakah* ("holiness") than of any traditional Jaina attitude. The most obvious influence of Islām is seen, however, in the repudiation by the Svetambara Loṅkāsāha sect of image worship as something without canonical support. A parallel sect, the Terāpanthin, also arose among the Digambaras.

Jaina influence at the Mughal court of Akbar is a bright chapter in Jaina history. Akbar honoured Hīravijaya Sūri, then the leader of the Svetambara Tapā *gaccha* (subgroup). His disciples and other monks gained the respect of the Mughal emperors Jahāngīr, Shāh Jahān, and even the Muslim chauvinist Aurangzeb. Akbar issued a decree prohibiting animal slaughter near important Jaina sites during the Paryuṣaṇa festival. Jahāngīr also issued decrees for the protection of Śatruñjaya, and Aurangzeb issued a decree favouring the Jainas with respect to proprietary rights over Mount Śatruñjaya. Mughal painting, influential in different schools of Indian painting, also influenced Jaina miniature painting.

BIBLIOGRAPHY

General sources: Good introductions are HERMANN JACOBI, "Jainism," in *Encyclopædia of Religion and Ethics,* vol. 7, pp. 465–474 (1928); and COLETTE CAILLAT, "Jainism," in *The Encyclopedia of Religion,* ed. by MIRCEA ELIADE, vol. 7, pp. 507–514 (1987). Standard works include HERMANN JACOBI (trans.), *Gaina Sūtras,* 2 vol. (1884–95, reissued as *Jaina Sutras,* 1968), with noteworthy introductions by Jacobi to each volume; JOHANN GEORGE BUHLER, *On the Indian Sect of the Jainas,* trans. from German, 2nd ed. (1963); HELMUTH VON GLASENAPP, *Der Jainismus* (1925, reprinted 1964), the most comprehensive text on Jainism, and *The Doctrine of Karman in Jain Philosophy,* trans. from German (1942); and WALTHER SCHUBRING, *The Doctrine of the Jainas* (1962; originally published in German, 1935), a scholarly work, and *The Religion of the Jainas,* trans. from German (1966). See also CHHOTELAL JAIN, *Chhotelal Jain's Jaina Bibliography,* 2nd. rev. ed., edited by SATYA RANJAN BANNERJEE, 2 vol. (1982); AMULYACHANDRA SEN, *Schools and Sects in Jaina Literature* (1931); JAGMANDERLAL JAINI, *Outlines of Jainism* (1916, reprinted 1982); A.L. BASHAM, *History and Doctrine of the Ājīvikas* (1951, reprinted 1981), a discussion of the Ājīvika influence on early Jainism; CHHOGMAL CHOPRHA, *A Short History of the Terapanthi Sect of the Swetamber Jains and Its Tenets,* 4th ed. (1950); BIMALA CHURN LAW, *Mahavira: His Life and Teachings* (1937), a good introduction to the subject; and PADMANABH S. JAINI, *The Jaina Path of Purification* (1979), a survey that discusses the Jaina understanding of karmic bondage and the path to liberation.

Special studies: NATHMAL TATIA, *Studies in Jaina Philosophy* (1951, reprinted 1980), especially the discussion on the problem of *ajñāna,* or false sense of reality, in various systems; NARENDRA NATH BHATTACHARYYA, *Jain Philosophy: Historical Outline* (1976); SATKARI MOOKERJEE, *The Jaina Philosophy of Non-Absolutism: A Critical Study of Anekāntavāda,* 2nd ed. (1978), a standard work by an authority on Indian philosophy; MOHANLAL MEHTA, *Jaina Philosophy,* new ed. (1971), *Jaina Culture* (1969), and *Jaina Psychology: A Psychological Analysis of the Jaina Doctrine of Karma* (1957); SHANTARAM B. DEO, *History of Jaina Monachism from Inscriptions and Literature* (1956); R. WILLIAMS, *Jaina Yoga* (1963, reprinted 1983), a masterly analysis of the Jaina ethics concerning the laity, with critical notes on authors of different sourcebooks; DAYANAND BHARGAVA, *Jaina Ethics* (1968); HARI SATYA BHATTACHARYA, *Jain Moral Doctrine* (1976); T.K. TUKOL, *Sallekhanā Is Not Suicide* (1976), a treatise on the monastic ritual of self-starvation; KAMAL C. SOGANI, *Ethical Doctrines in Jainism* (1967); VILAS ADINATH SANGAVE, *Jaina Community: A Social Survey,* 2nd ed. (1980); CHAMPAT R. JAIN, *Jaina Law* (1926); COLETTE CAILLAT, *Attonements in the Ancient Ritual of the Jaina Monks* (1975; originally published in French, 1965); COLETTE CAILLAT and RAVI KUMAR, *The Jain Cosmology* (1981); and A.N. UPADHYE, *Upadhye Papers* (1983), a collection of essays on Jaina history and literature by an eminent Jaina scholar.

Literature and art: M. WINTERNITZ, *History of Indian Literature,* vol. 2 (1933, reprinted 1971; originally published in German, 1920); H.R. KAPADIA, *A History of the Canonical Literature of the Jainas* (1941), a good description of the Jaina canon; A.

CHAKRAVARTY, *Jaina Literature in Tamil* (1974), a survey of Jaina works in this South Indian language and Jaina influence on Tamil literature; JAGDISHCHANDRA JAIN, *Prakrit Narrative Literature: Origin and Growth* (1981); B.C. BHATTACHARYA, *The Jaina Iconography,* 2nd rev. ed. (1974), a brief outline of the subject; JYOTINDRA JAIN and EBERHARD FISCHER, *Jaina Iconography,* 2 vol. (1978), a later work; UMAKANT P. SHAH, *Studies in Jaina Art* (1955), a review of Jaina art in North India, with a discussion of various symbols in Jaina worship and a good bibliography, and *Akota Bronzes* (1959), a description of rare Jaina bronzes from a site in Gujarāt; A. GHOSH (ed.), *Jaina Art and Architecture,* 3 vol. (1974–75); KLAUS FISCHER, *Caves and Temples of the Jains* (1956); MOTI CHANDRA, *Jain Miniature Paintings from Western India* (1949), a standard textbook; W. NORMAN BROWN, *The Story of Kālaka* (1933), a well-known work on Kālakācārya and miniature Jaina paintings; UMAKANT P. SHAH (ed.), *Treasures of Jaina Bhaṇḍāras* (1978); and P.B. DESAI, *Jainism in South India and Some Jaina Epigraphs* (1957), a useful compilation.

(U.P.S./G.R.S.)

Jakarta

Jakarta is the capital of the Republic of Indonesia and one of the largest and most consistently growing cities in that country. Until 1949 the city was called Batavia; its present name was adopted in that year but was spelled Djakarta until 1972. Coextensive with the metropolitan district of Jakarta Raya, it has an area of 228 square miles (590 square kilometres) and lies at the mouth of the Ciliwung (Liwung River) on the northwest coast of Java.

In 1966 the city was declared to be a special metropolitan district (*daerah khusus ibukota*), thus gaining a status approximately equivalent to that of a state or province. The city has long been a major trade and financial centre; it has also become an important industrial city and an important centre for education.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. Jakarta lies on a low, flat alluvial plain, with extensive swampy areas; it is easily flooded during the rainy seasons. The parts of the city farther inland are slightly higher. The draining of swamps for building purposes and the continuous decrease of upland forest vegetation have increased the danger of floods. With an excess of water in the soil, Jakarta still has a shortage of clean drinking water, for which there is an increasing demand. The area is quite fertile for fruit and other horticulture, as most of the soil is of old volcanic origin.

Climate. Jakarta is a tropical, humid city, with temperatures ranging between the extremes of 75° and 93° F (24° and 34° C) and a relative humidity between 75 and 85 percent. The average mean temperatures are 79° F (26° C) in January and 82° F (28° C) in October. The annual rainfall is more than 67 inches (1,700 millimetres). Temperatures are often modified by sea winds. Jakarta, like any other large city, now also has its share of air and noise pollution.

The city layout. Although the Dutch were the first to attempt to plan the city, the city layout is probably more British than Dutch in character, as can be seen from such large squares as the Medan Merdeka (Freedom Field) and Lapangan Banteng (Place of the Gaur [large wild ox]). The Oriental style, or "indische" style, as the Dutch call it, is, however, not only apparent in the city's way of life but also in the types of houses, the wide, tree-lined streets, and the original spacious gardens and house lots. In Kebayoran, a satellite town built since World War II on the southwestern side of the city, and in other modern developments, the houses and garden lots are much smaller than in the older colonial districts.

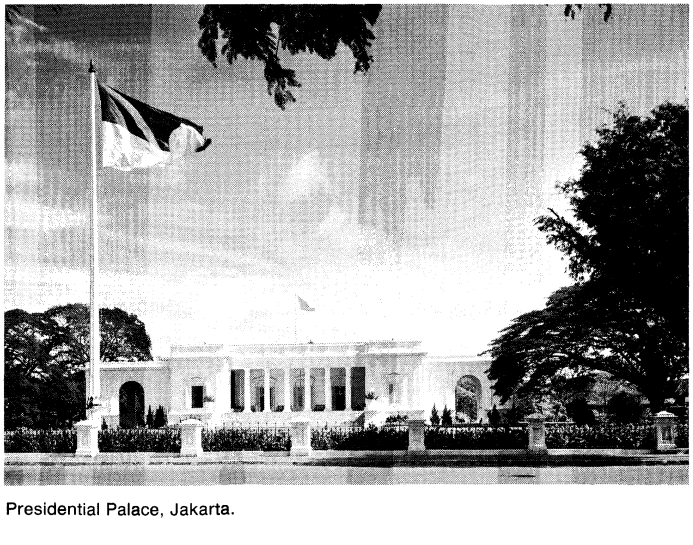
Traditional neighbourhoods

Jakarta has always been a city of new settlers who assimilated local ways and became Jakartans themselves. Some traditional neighbourhoods can, however, be identified. The Kota (Fort), or Old City, for example, sometimes also called the downtown section, is the central business district and also the financial capital of Indonesia. It houses a significant part of the Chinese population. The area of Kemayoran (Progress) and Senen, originally on the eastern fringe of the city, is now almost central in its location and increasingly has become the major retail area of the city.

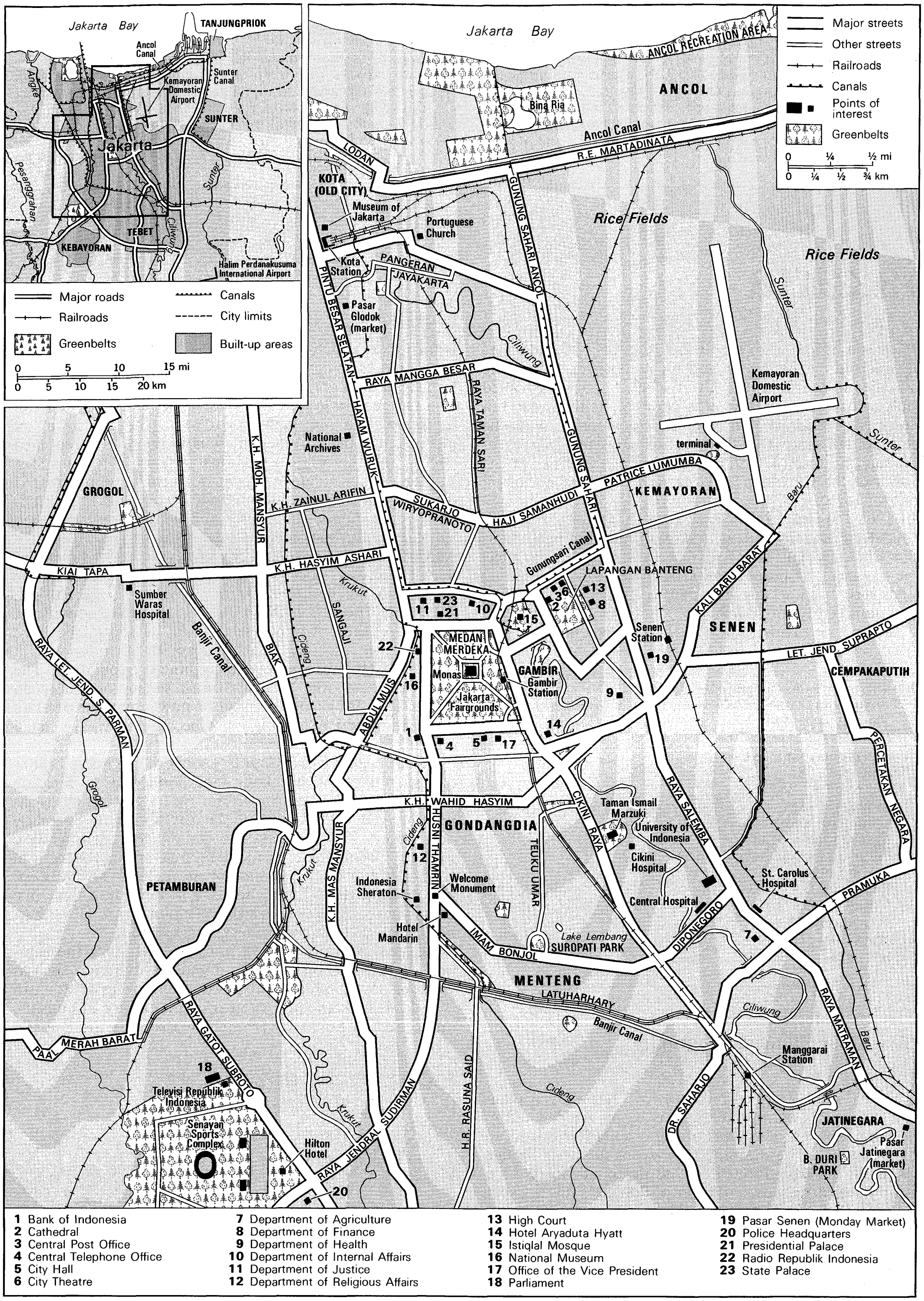
The Jatinegara (Real Country) section, originally a Sundanese settlement but later incorporated as a separate town, then a Dutch army camp (Meester Cornelis), has now merged with the rest of Jakarta and includes many new settlers. The Menteng and Gondangdia sections were formerly fashionable residential areas near the central Medan Merdeka (then called Weltevreden). To the west, Tanahabang (Red Earth) and Jati Petamburan, are, like Kemayoran, densely developed. Tanjungpriok is the harbour, with its own community attached to it.

The most common type of house in the city is the kampong, or village house. Most are built of materials such as wood or bamboo mats, but this does not necessarily mean that this is substandard. Another common type of housing, often used to house government workers, is the colonial

H. Armstrong Roberts



Presidential Palace, Jakarta.



Jakarta and (inset) its metropolitan area.

urban house, or *rumah gedongan;* these are mostly single-family detached or semidetached houses, each standing on a separate lot. Apartment buildings constitute a more modern category; although they are more economical in the use of land than single-family types, their architectural and construction costs often make them fairly expensive. Housing is generally overcrowded.

Some of Jakarta's buildings, such as the Portuguese Church (1695) in the Old City, are of architectural or historical interest. Some of the buildings around the city square in the Kota also date from colonial times, including the old city hall (1710), which has been restored and now serves as the municipal museum. The National Archives building was originally the palace of a Dutch governor general, Abraham van Riebeeck. The Ministry of Finance building, facing Banteng Square, also was designed as

Buildings from colonial times

a governor's palace (Herman Willem Daendels, one of Napoleon's marshals). The Presidential Palace, north of Medan Merdeka, faces the Monas, or Monumen Nasional (National Monument), at 360 feet (110 metres) the highest building in Jakarta. The Istiqlal Mosque, in the northeast corner of Medan Merdeka opposite Banteng Square, is one of the largest mosques in Southeast Asia. The National Museum (formerly the Central Museum), on the west side of the Medan Merdeka, houses a collection of historical, cultural, and artistic artifacts.

After World War II Jakarta underwent a building boom. The Hotel Indonesia (the city's first high-rise building) and the Senayan Sports Complex were built for the Asian Games in 1962. Most high-rise buildings are located along Husni Thamrin and Jendral Sudirman roads, connecting Jakarta with Kebayoran.

THE PEOPLE

The population of Jakarta has increased more than 100 percent since 1940. Much of this increase is attributed to immigration. Although government regulations close the city to unemployed new settlers, better economic conditions inevitably attract new people. In addition, much of the population is young and fertile, resulting in a very high natural increase potential. Analysis of the immigrant stream shows that after the West Javanese, the largest groups represented are the Central and East Javanese; a sizable number also are from Sumatra. Other population groups—Arabs, Indians, Europeans, and Americans—are present in small numbers.

THE ECONOMY

Economically, Jakarta plays several roles. It can be identified first as the national capital and a central place of control for the national economy, then as an administrative centre in its own right, and, finally, as a significant industrial hub. In addition, its location as a port makes it an important centre for trade.

Industry. Jakarta has some manufacturing industries, including several iron foundries and repair shops, margarine and soap factories, breweries, and printing works. Machinery, cigarettes, paper, glassware, wire cable, and aluminum and asbestos, and more recently also automotive, products are manufactured. There are also tanneries, sawmills, textile mills, food-processing plants, and a film industry.

Commerce and trade. The cost of living in the city continues to rise. Land is expensive, and rents are high, so that industrial development and the construction of new housing usually are undertaken on the outskirts, while commerce and banking remain concentrated in the city centre. The Indonesian Chamber of Commerce is active in promoting trade with other countries; the annual Jakarta Fair (usually held from July to August) also serves to promote trade.

City-operated markets

To meet the needs of the local city population, the municipality operates several markets. The central city markets (Pasar Kota), like the markets of Pasar Senen to the east of the central city and Pasar Glodok in the Kota area, are major retail centres. The Pasar Jatinegara is primarily a food supply centre. The district markets are fairly large, with each one catering to a whole section of the city. There are also small neighbourhood markets, each serving only a limited area. Special markets include one selling fish, one selling used and new automobile parts, the Pasar Rumput flea market, and the Jalan Surabaya souvenir and antique market. Jakarta also has several general neighbourhood markets.

Transportation. Major road arteries lead west from the centre of the Kota Old City and east and south from the administrative centre in Gambir. To the east, a major railroad connects the city with all of the island of Java. There is also a highway, primarily a regional supply road, running between Jakarta and the agriculturally productive areas of East and Central Java. To the south, a road and railroad connect Jakarta with Bogor, Sukabumi, and Bandung. To the west, a railroad and road run to Banten and to the harbour in Merak, which is connected by ferry to Lampung in Sumatra.

The port of Tanjungpriok in Jakarta is the largest in Indonesia, handling exports from West Java and a large proportion of Indonesia's import trade; many goods are transshipped to other islands or harbours.

Jakarta is served by several international airlines, by Garuda Indonesian Airways (the national airline, with international and domestic service), and by other domestic airlines. The city has airports at Kemayoran (relatively close to the city centres) and Halim Perdanakusuma and the new international airport in Cengkareng.

The *becak,* or tricycle taxi

The central bus terminal, located on Lapangan Banteng, serves all the city, intercity, and regional bus lines; there are also suburban bus terminals in Jatinegara, Kebayoran, Grogol, Kota, and Tanjungpriok. The major railroad stations are at Kota in the old city, Gambir on Medan Merdeka, Pasar Senen on the east, and Manggarai and Jatinegara on the south. Tanahabang serves the west and traffic to Merak. Traffic jams occur particularly during the morning and afternoon rush hours. Public transportation in the city is by bus or minibus. The *becak*, or tricycle taxi, is used only for local neighbourhood transportation, and regular taxis now operate throughout the metropolitan area.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The mayor of the city has the same status as the governor of a province. The city government is composed of two branches, the executive and the electorate. The executive consists of a governor assisted by four vice governors, an executive staff, and a regional secretary; there are also a number of city directorates, bureaus, and agencies attached to the executive. The electorate consists of 35 to 40 members, including representatives of eight political parties, four representatives of the armed forces, and nine representatives of the so-called functional groups. It is headed by a council of five members, one chairman, and four vice chairmen.

Public utilities. Public utilities are usually operated or owned by the Indonesian government. The State Electricity Company and the subsidiary State Gas Company both supply Jakarta. Postal and cable services and telephone services are supplied by state companies working under the aegis of the Department of Communications. Jakarta's electricity comes from several sources; these include the thermal plant in Ancol, close to the port of Tanjungpriok, smaller diesel plants in various parts of the city, and the Jatiluhur hydroelectricity project located close to Purwakarta, about 70 miles (110 kilometres) southeast of Jakarta. A new thermal power plant is under construction in Surabaya.

The city government is responsible for the water supply. The city water is obtained in part from freshwater springs in the Bogor area, but most of the supply comes from the Pejompongan water treatment plant. Water is required mainly for domestic purposes but is also needed for industry and to supply ships. The removal of garbage and the provision of other sanitation services are also the responsibility of the municipality.

Hospitals and clinics

Health. There are three major hospitals—one operated directly by the Department of Health, one by the Roman Catholic Church, and one by a Protestant mission. Three municipal hospitals each serve a separate area of the city—the Sumber Waras Hospital in western Jakarta, the Fatmawati Hospital in Kebayoran (southern Jakarta), and the Persahabatan (Friends) Hospital in eastern Jakarta. Altogether Jakarta has about 40 general or special hospitals. In addition, there are several hundred general clinics or polyclinics located throughout the city. A quarantine hospital is in operation in Tanjungpriok. The city also operates a hospital and rehabilitation centre for the mentally ill and destitute, and there are many family-planning and child care clinics.

Education. To meet the needs for primary education, many new elementary schools and secondary schools have been built, and a number of old school buildings have been renovated. There is a well-developed system of kindergartens, elementary schools, *madrasah*s (religious schools), secondary schools, and high schools. There are also many vocational and special schools and more than 100 universities, academies, and institutes for higher learning. The largest and best known university is the Universitas Indonesia (founded 1950).

CULTURAL LIFE

Among other cultural activities, the Taman Ismail Marzuki centre has facilities for traditional or classical art performances as well as theatres for presenting modern plays and concerts; the centre also has a planetarium. Traditional performances include *wayang* dance and drama, gamelan music, and *wayang* puppet shows. Traditional performances representing the culture of other parts of Indonesia are included in the programs presented at the annual Jakarta Fair.

Extensive public recreation areas in and around Jakarta include the Bina Ria seaside recreation area at Ancol and the Ragunan zoo, near Pasarminggu. Playgrounds include, among others, the Taman Ria complex at the Jakarta Fair grounds, located just south of the Monas. The 250-acre (100-hectare) Taman Mini Indonesia Indah (Beautiful Indonesia in Miniature) park, just southeast of the city, contains exhibits of traditional houses representing each of Indonesia's 27 provinces. The city also provides public recreation facilities.

History

The first settlements on the site of Jakarta were established at the mouth of the Ciliwung, perhaps as early as the 5th century AD. The city's official history, however, starts in 1527, when the Sultan of Bantam defeated the Portuguese there and called the place Jayakerta (Sundanese: Glorious Fortress).

The Dutch, under the leadership of Jan Pieterszoon Coen, captured and razed the city in 1619, after which the capital of the Dutch East Indies—a walled township named Batavia—was established on the site.

The colonial history of the city can be divided into three major periods. First was that of the Dutch East India Company, when most of the activities of the city centred around the fortress and the company warehouses. At that time the city somewhat resembled a typical Dutch town, complete with canals. The second period began in the early 1800s, when the city was extended to include higher and more healthful areas to the south, which would later become the seat of the new colonial government. A brief interval of British control during the Napoleonic Wars, ending in 1815, interrupted the second period. During the third period, which lasted from about the 1920s to 1941, the city became modernized.

The colonial era

The colonial era ended with the entry of Japan into World War II, when Indonesia was occupied by Japanese forces. After the war the city was briefly occupied by the Allies and then was returned to the Dutch. During the Japanese occupation and again after Indonesian nationalists declared independence on August 17, 1945, the city was renamed Djakarta. The Dutch name Batavia remained the internationally recognized name until full Indonesian independence was achieved and Djakarta was officially proclaimed the national capital (and its present name recognized) on December 27, 1949.

Jakarta has undergone tremendous growth and development since Indonesia's independence. Despite its problems, the city has become one of the largest metropolises of tropical Asia.

BIBLIOGRAPHY. On the history of Jakarta, see F. DE HAAN, *Uit Oud-Batavia* (1898); and DJAKARTA, KOTAPRADJA, *Sedjarah pemerintahan kota Djakarta* (1958). A contemporary description is given in the guidebook, *Djakarta,* issued by the PETUNDJUK DCI (1969). Census and population information may be found in the Jakarta volumes issued by the BIRO PUSAT STATISTIK in its *Sensus penduduk* (Census of Population) reports for the censuses of 1961, 1971, and 1980; and in H.J. HEEREN (ed.), *The Urbanisation of Djakarta* (1955). See also *Djakarta: Its Rehabilitation and Development* (n.d.), issued by the Badan Perentjana Pembangunan (Development Planning Body) of the City Government; and PAULINE D. MILONE, *Urban Areas in Indonesia* (1966).

(W.J.W.)

Japan

The island country of Japan (Japanese: Nihon or Nippon) lies off the east coast of Asia. It consists of four main islands—Hokkaido (Hokkaidō), Honshu (Honshū), Shikoku, and Kyushu (Kyūshū)—that, along with numerous smaller islands, trend in a great northeast-southwest arc for some 1,500 miles (2,400 kilometres); the average width of the chain is about 130 miles, and Japan has a total land area of 145,870 square miles (377,801 square kilometres). Honshu is the largest of the four main islands, followed in size by Hokkaido, Kyushu, and Shikoku. The national capital, Tokyo (Tōkyō), is one of the most populous cities in the world.

Japan is bounded to the west by the Sea of Japan, which separates it from the eastern shores of the Soviet Union and North and South Korea; to the north by La Perouse (Sōya) Strait (separating it from the Soviet island of Sakhalin) and the Sea of Okhotsk; to the northeast by the southern Kuril Islands (since World War II under Soviet administration); to the east and south by the Pacific Ocean; and to the southwest by the East China Sea, which separates it from the People's Republic of China.

The Japanese landscape is rugged, with more than four-fifths of the land surface consisting of mountains. The abundant rainfall and generally mild temperatures throughout most of the country have produced a lush vegetation and, despite the mountainous terrain and generally poor soils, have made it possible to raise a variety of crops. Japan has a large population, which is heavily concentrated in the low-lying areas along the Pacific coast of Honshu.

Complexity and contrast are the keynotes of life in Japan—a nation possessing an intricate and ancient cultural tradition, yet one that, especially since World War II, has emerged as a modern industrial giant. Tension between old and new is apparent in all phases of Japanese life. A characteristic sensitivity to natural beauty and a concern with form and balance are evident in such cities as Kyōto and Nara, as well as in Japan's ubiquitous gardens. Even in the countryside, however, the impact of rapid westernization is evident upon many aspects of Japanese life. The agricultural regions are characterized by low population densities and well-ordered rice fields and fruit orchards, whereas the industrial and urbanized belt along the Pacific coast of Honshu is noted for its highly concentrated population and for the despoilment of the environment by airborne pollutants and other wastes. Heavy emphasis is placed on education.

Japan's spectacular economic growth—the greatest of any nation since the 1940s—has brought the country to the forefront of the world economy. It is one of the world's principal shipbuilders and is a major producer and exporter of manufactured goods, including automobiles, electrical products, chemicals, and steel. An important feature of the burgeoning economy is the prevalence of large, quasi-monopolistic industrial companies. (Ak.W./Ed.)

This article is divided into the following sections:

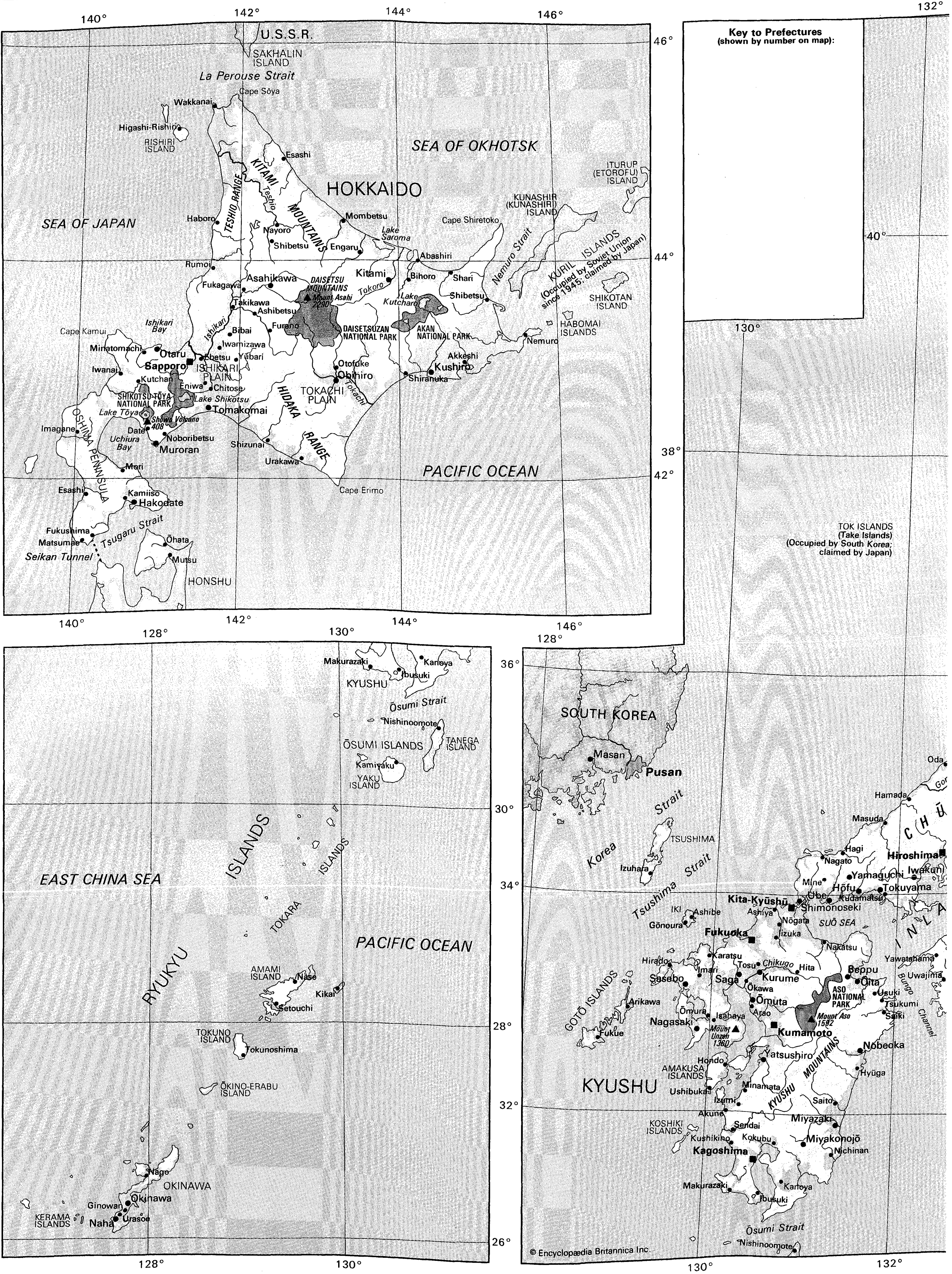
Physical and human geography

THE LAND

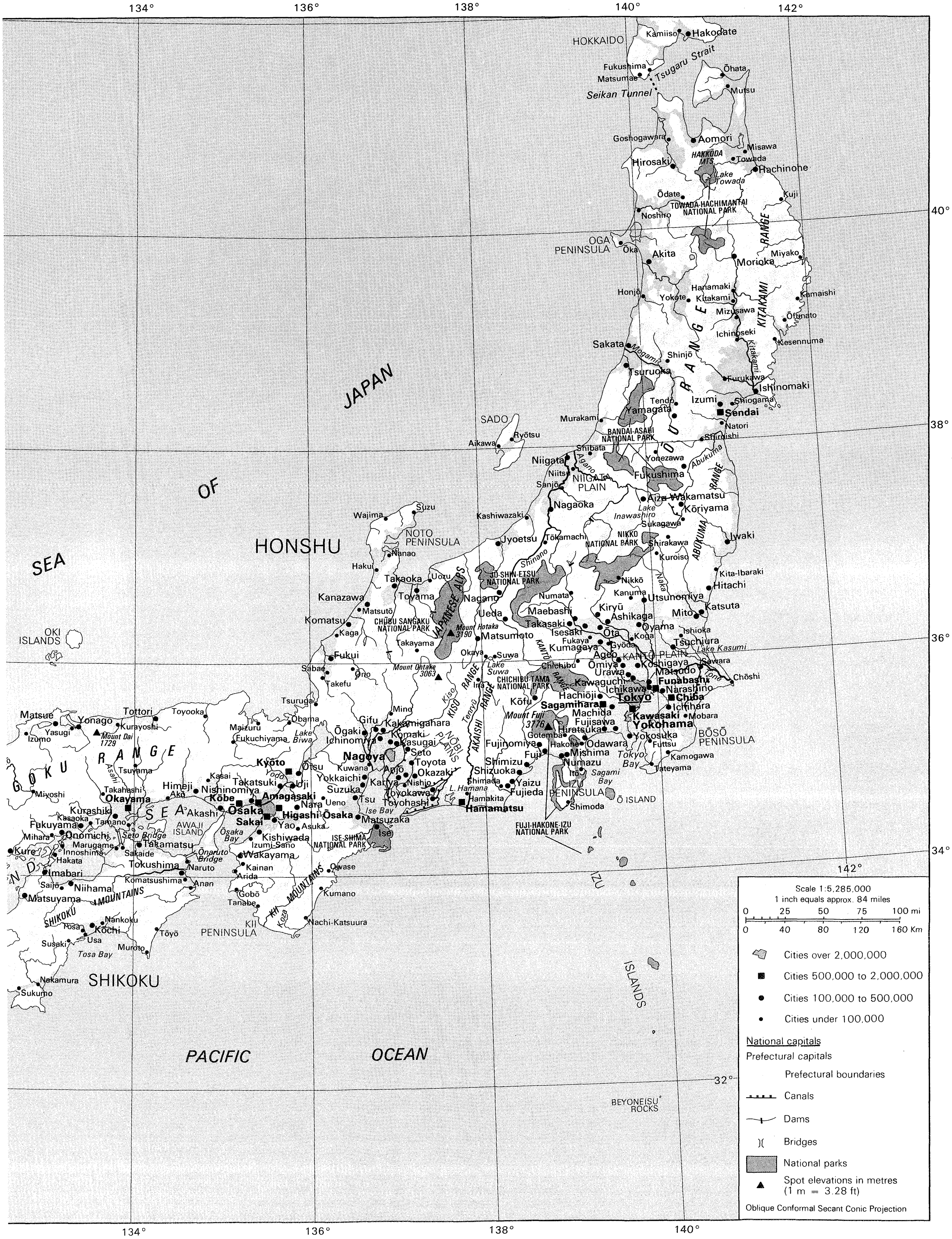
Mountainous character of Japan

Relief. The mountainous character of the country is the outcome of geologically recent orogenic (mountain-building) forces, as evidenced by the frequent occurrence of violent earthquakes, volcanic activity, and signs of change in levels along the coast. There are no sizable structural plains and peneplains (large land areas levelled by erosion), which usually occur in more stable regions of the Earth. The mountains are for the most part in a youthful stage of dissection in which steep slopes are incised by dense river-valley networks. Rivers are mostly torrential, and their valleys are accompanied by series of river terraces that are the result of movements in the Earth's crust. Recent volcanoes are juxtaposed with old and highly dissected ones. The shores are characterized by elevated and depressed features such as headlands and bays, which display an incipient stage of development.

The mountains are divided into many small land blocks that are separated by lowlands or deep saddles; there is no long or continuous mountain range. These land blocks are the result of intense faulting (movement of adjacent rock masses along a fracture) and warping (bending of the Earth's crust), the former process being regarded as dominant. One consequence is that mountain blocks are often bounded by fault scarps and flexure slopes that descend in step formation to the adjacent lowlands.



140°
142°
144°
146°
46°
44°
42°
U.S.S.R.
SAKHALIN ISLAND
La Perouse Strait
Cape Sōya
Wakkanai
Higashi-Rishiri
RISHIRI ISLAND
Esashi
SEA OF OKHOTSK
HOKKAIDO
KITAMI MOUNTAINS
TESHIO RANGE
Teshio
ITURUP (ETOROFU) ISLAND
KUNASHIR (KUNASHIRI) ISLAND
Haboro
SEA OF JAPAN
Nayoro
Mombetsu
Lake Saroma
Cape Shiretoko
Shibetsu
Engaru
Nemuro Strait
KURIL ISLANDS (Occupied by Soviet Union since 1945; claimed by Japan)
Abashiri
Rumoi
Kitami
Bihoro
Shari
Asahikawa
Fukagawa
DAISETSU MOUNTAINS
Mount Asahi 2290
Tokoro
Lake Kutcharo
Shibetsu
SHIKOTAN ISLAND
Takikawa
Ashibetsu
Ishikari Bay
Ishikari
Bibai
Furano
DAISETSUZAN NATIONAL PARK
AKAN NATIONAL PARK
HABOMAI ISLANDS
Cape Kamui
Iwamizawa
Nemuro
Minatomachi
Otaru
Ebetsu
Yūbari
Akkeshi
Sapporo
ISHIKARI PLAIN
Otofuke
Iwanai
Kutchan
Obihiro
Kushiro
Eniwa
Chitose
TOKACHI PLAIN
Shiranuka
SHIKOTSU-TŌYA NATIONAL PARK
Lake Shikotsu
HIDAKA RANGE
Tokachi
Lake Tōya
Tomakomai
Shōwa Volcano 408
Imagane
Date
Noboribetsu
Uchiura Bay
Muroran
Shizunai
OSHIMA PENINSULA
Urakawa
PACIFIC OCEAN
Mori
Cape Erimo
Esashi
Kamiiso
Hakodate
Fukushima
Matsumae
Tsugaru Strait
Ōhata
Seikan Tunnel
Mutsu
HONSHU
Key to Prefectures (shown by number on map):
132°
130°
40°
38°
TOK ISLANDS (Take Islands) (Occupied by South Korea; claimed by Japan)
128°
130°
30°
28°
26°
Makurazaki
Kanoya
Ibusuki
KYUSHU
Ōsumi Strait
Nishinoomote
ŌSUMI ISLANDS
TANEGA ISLAND
Kamiyaku
YAKU ISLAND
RYUKYU ISLANDS
TOKARA ISLANDS
EAST CHINA SEA
PACIFIC OCEAN
AMAMI ISLAND
Nase
Kikai
Setouchi
TOKUNO ISLAND
Tokunoshima
ŌKINO-ERABU ISLAND
Nago
OKINAWA
Okinawa
Ginowan
KERAMA ISLANDS
Naha
Urasoe
36°
34°
32°
SOUTH KOREA
Masan
Pusan
Korea Strait
TSUSHIMA
Izuhara
Tsushima Strait
Oda
Hamada
Masuda
CHŪ
Hagi
Nagato
Hiroshima
Yamaguchi
Iwakuni
Mine
Hōfu
Tokuyama
Ube
Kudamatsu
Kita-Kyūshū
Shimonoseki
IKI
Ashibe
Ashiya
Nōgata
Gōnoura
SUŌ SEA
INLA
Fukuoka
Iizuka
Nakatsu
Hirado
Karatsu
Chikugo
Hita
Yawatahama
Imari
Tosu
Kurume
Beppu
Uwajima
Sasebo
Saga
Ōita
Bungo Channel
Ōkawa
ASO NATIONAL PARK
Usuki
Tsukumi
GOTŌ ISLANDS
Arikawa
Ōmuta
Ōmura
Isahaya
Arao
Mount Aso 1592
Saiki
Nagasaki
Mount Unzen 1360
Kumamoto
Fukue
Nobeoka
Hondo
Yatsushiro
KYUSHU MOUNTAINS
AMAKUSA ISLANDS
Hyūga
Minamata
KYUSHU
Ushibuka
Izumi
Saito
Akune
Miyazaki
KOSHIKI ISLANDS
Sendai
Kushikino
Kokubu
Miyakonojō
Kagoshima
Nichinan
Makurazaki
Kanoya
Ibusuki
Ōsumi Strait
Nishinoomote
© Encyclopædia Britannica Inc.


134°
136°
138°
140°
142°
40°
38°
36°
34°
32°
HOKKAIDO
Kamiiso
Hakodate
Tsugaru Strait
Fukushima
Matsumae
Seikan Tunnel
Ōhata
Mutsu
Goshogawara
Aomori
HAKKODA MTS
Misawa
Towada
Hirosaki
Lake Towada
Hachinohe
Ōdate
Kuji
Noshiro
TOWADA-HACHIMANTAI NATIONAL PARK
OGA PENINSULA
Oka
Akita
Morioka
Miyako
KITAKAMI RANGE
Honjō
Yokote
Hanamaki
Kitakami
Kamaishi
Mizusawa
Ōfunato
Ichinoseki
Kesennuma
Sakata
Mogami
Shinjō
Kitakami
Tsuruoka
Furukawa
Ishinomaki
Tendo
Izumi
Shiogama
Yamagata
Sendai
Murakami
Natori
BANDAI-ASAHI NATIONAL PARK
Shiroishi
SADO
Ryōtsu
Aikawa
Shibata
Niigata
Agano
Yonezawa
Abukuma
Niitsu
NIIGATA PLAIN
Fukushima
Sanjō
Aizu-Wakamatsu
Nagaoka
Kōriyama
Lake Inawashiro
Sukagawa
Kashiwazaki
NIKKO NATIONAL PARK
ABUKUMA RANGE
Iwaki
Jyoetsu
Tōkamachi
Shirakawa
Kuroiso
OU RANGE
Shinano
Nikkō
Kita-Ibaraki
JO-SHIN-ETSU NATIONAL PARK
Hitachi
Kanuma
Naka
Numata
Utsunomiya
Nagano
Maebashi
Kiryū
Ashikaga
Mito
Katsuta
Ueda
Takasaki
Ōta
Oyama
Isesaki
Ishioka
Fukaya
Koga
Kumagaya
Gyōda
Tsuchiura
Lake Kasumi
KANTŌ PLAIN
Ageo
Sawara
Ōmiya
Koshigaya
Urawa
Matsudo
Tone
Chōshi
Kawaguchi
Funabashi
Ichikawa
Narashino
Tokyo
Chiba
Ichihara
Kawasaki
Mobara
Yokohama
Yokosuka
BŌSŌ PENINSULA
Futtsu
Tokyo Bay
Kamogawa
Tateyama
Sagami Bay
Ō ISLAND
IZU
IZU ISLANDS
JAPAN
OF
SEA
HONSHU
Suzu
Wajima
NOTO PENINSULA
Nanao
Hakui
Takaoka
Uozu
Toyama
Kanazawa
Matsutō
JAPANESE ALPS
Komatsu
CHŪBU SANGAKU NATIONAL PARK
Mount Hotaka 3190
Matsumoto
Kaga
Takayama
OKI ISLANDS
Okaya
Suwa
Lake Suwa
Fukui
Sabae
Ono
Mount Ontake 3063
Takefu
Kiso
KISO RANGE
Ina
CHICHIBU-TAMA NATIONAL PARK
Chichibu
KANTŌ RANGE
Hachiōji
Kōfu
Sagamihara
Machida
Tenryū
AKAISHI RANGE
Mount Fuji 3776
Fujisawa
Hiratsuka
Gotemba
Odawara
Hakone
Mishima
Numazu
Fuji
Fujinomiya
Ito
Shimizu
Shizuoka
Shimada
Yaizu
Fujieda
IZU PENINSULA
Shimoda
FUJI-HAKONE-IZU NATIONAL PARK
L. Hamana
Hamakita
Hamamatsu
Toyohashi
Toyokawa
Nishio
Okazaki
Toyota
Anjō
Kariya
Seto
Kasugai
Nagoya
Komaki
Kakamigahara
Gifu
Ōgaki
Ichinomiya
Mino
NŌBI PLAIN
Tsuruga
Obama
Lake Biwa
Maizuru
Fukuchiyama
Ōtsu
Kyoto
Yodo
Uji
Kuwana
Yokkaichi
Suzuka
Tsu
Ueno
Ise Bay
Matsuzaka
Ise
ISE-SHIMA NATIONAL PARK
Owase
KII MOUNTAINS
Koza
Kumano
Nachi-Katsuura
KII PENINSULA
Takatsuki
Kasai
Himeji
Nishinomiya
Amagasaki
Kōbe
Ōsaka
Higashi-Ōsaka
Nara
Sakai
Yao
Asuka
Kishiwada
Izumi-Sano
Ōsaka Bay
Wakayama
Kainan
Arida
Gobō
Tanabe
Akashi
AWAJI ISLAND
Ōnaruto Bridge
Naruto
Tokushima
Anan
Komatsushima
Tottori
Toyooka
Matsue
Yonago
Kurayoshi
Yasugi
Izumo
Mount Dai 1729
CHUGOKU RANGE
Asahi
Tsuyama
Miyoshi
Takahashi
Akō
Okayama
Kurashiki
Kasaoka
Tamano
Fukuyama
Onomichi
Mihara
Seto Bridge
SEA
Marugame
Takamatsu
Sakaide
Kure
Innoshima
Hakata
Imabari
Saijō
Niihama
Matsuyama
SHIKOKU MOUNTAINS
Nankoku
Kōchi
Tosa
Usa
Susaki
Tosa Bay
Tōyō
Muroto
Nakamura
Sukumo
SHIKOKU
PACIFIC
OCEAN
BEYONEISU ROCKS
Scale 1:5,285,000
1 inch equals approx. 84 miles
0 25 50 75 100 mi
0 40 80 120 160 Km
Cities over 2,000,000
Cities 500,000 to 2,000,000
Cities 100,000 to 500,000
Cities under 100,000
National capitals
Prefectural capitals
Prefectural boundaries
Canals
Dams
Bridges
National parks
Spot elevations in metres (1 m = 3.28 ft)
Oblique Conformal Secant Conic Projection

MAP INDEX

Political subdivisions

Aichi, *prefecture*	35 00 N	137 15 E
Akita, *prefecture*	40 00 N	140 30 E
Aomori, *prefecture*	40 45 N	141 00 E
Chiba, *prefecture*	35 30 N	140 20 E
Ehime, *prefecture*	33 45 N	132 45 E
Fukui, *prefecture*	36 00 N	136 15 E
Fukuoka, *prefecture*	33 30 N	130 30 E
Fukushima, *prefecture*	37 30 N	140 15 E
Gifu, *prefecture*	35 45 N	137 00 E
Gumma, *prefecture*	36 30 N	139 00 E
Hiroshima, *prefecture*	34 30 N	133 00 E
Hokkaido, *prefecture*	44 00 N	143 00 E
Hyōgo, *prefecture*	35 00 N	135 00 E
Ibaraki, *prefecture*	36 30 N	140 30 E
Ishikawa, *prefecture*	36 45 N	136 45 E
Iwate, *prefecture*	39 30 N	141 30 E
Kagawa, *prefecture*	34 15 N	134 00 E
Kagoshima, *prefecture*	30 00 N	130 00 E
Kanagawa, *prefecture*	35 30 N	139 30 E
Kōchi, *prefecture*	33 40 N	133 30 E
Kumamoto, *prefecture*	32 30 N	130 40 E
Kyōto, *urban prefecture*	35 10 N	135 30 E
Mie, *prefecture*	34 30 N	136 30 E
Miyagi, *prefecture*	38 30 N	141 00 E
Miyazaki, *prefecture*	32 15 N	131 30 E
Nagano, *prefecture*	36 00 N	138 00 E
Nagasaki, *prefecture*	33 00 N	129 30 E
Nara, *prefecture*	34 30 N	136 00 E
Niigata, *prefecture*	37 30 N	139 00 E
Ōita, *prefecture*	33 15 N	131 30 E
Okayama, *prefecture*	35 00 N	134 00 E
Okinawa, *prefecture*	26 42 N	128 11 E
Ōsaka, *urban prefecture*	34 30 N	135 30 E
Saga, *prefecture*	33 20 N	130 00 E
Saitama, *prefecture*	36 00 N	139 30 E
Shiga, *prefecture*	35 15 N	136 00 E
Shimane, *prefecture*	35 00 N	132 30 E
Shizuoka, *prefecture*	35 00 N	138 00 E
Tochigi, *prefecture*	36 45 N	139 45 E
Tokushima, *prefecture*	34 15 N	134 00 E
Tokyo, *metropolis*	35 45 N	139 30 E
Tottori, *prefecture*	35 30 N	134 00 E
Toyama, *prefecture*	36 30 N	137 30 E
Wakayama, *prefecture*	34 00 N	135 30 E
Yamagata, *prefecture*	38 30 N	140 15 E
Yamaguchi, *prefecture*	34 20 N	131 30 E
Yamanashi, *prefecture*	35 30 N	138 30 E

Cities and towns

Abashiri	44 01 N	144 17 E
Ageo	35 58 N	139 36 E
Aikawa	38 02 N	138 15 E
Aizu-Wakamatsu	37 30 N	139 55 E
Akashi	34 38 N	134 59 E
Akita	39 43 N	140 07 E
Akkeshi	43 02 N	144 51 E
Akō	34 45 N	134 24 E
Akune	32 01 N	130 12 E
Amagasaki	34 43 N	135 25 E
Anan	33 55 N	134 40 E
Anjō	34 57 N	137 05 E
Aomori	40 49 N	140 45 E
Arao	32 59 N	130 29 E
Arida	34 05 N	135 07 E
Arikawa	32 59 N	129 07 E
Asahikawa	43 46 N	142 22 E
Ashibe	33 48 N	129 46 E
Ashibetsu	43 31 N	142 11 E
Ashikaga	36 20 N	139 27 E
Ashiya	33 53 N	130 40 E
Asuka	34 28 N	135 49 E
Beppu	33 17 N	131 30 E
Bibai	43 19 N	141 52 E
Bihoro	43 49 N	144 07 E
Chiba	35 36 N	140 07 E
Chichibu	35 59 N	139 05 E
Chitose	42 49 N	141 39 E
Chōshi	35 44 N	140 50 E
Date	42 27 N	140 51 E
Ebetsu	43 07 N	141 34 E
Engaru	44 03 N	143 31 E
Eniwa	42 53 N	141 34 E
Esashi	41 52 N	140 07 E
Esashi	44 56 N	142 35 E
Fuji	35 12 N	138 43 E
Fujieda	34 52 N	138 16 E
Fujinomiya	35 18 N	138 37 E
Fujisawa	35 21 N	139 29 E
Fukagawa	43 43 N	142 03 E
Fukaya	36 12 N	139 17 E
Fukuchiyama	35 18 N	135 07 E
Fukue	32 41 N	128 50 E
Fukui	36 04 N	136 13 E
Fukuoka	33 35 N	130 24 E
Fukushima	37 45 N	140 28 E
Fukushima	41 29 N	140 15 E
Fukuyama	34 29 N	133 22 E
Funabashi	35 42 N	139 59 E
Furano	43 15 N	142 22 E
Furukawa	38 34 N	140 58 E
Futtsu	35 19 N	139 49 E
Gifu	35 25 N	136 45 E
Ginowan	26 16 N	127 45 E
Gobō	33 53 N	135 10 E
Gōnoura	33 45 N	129 41 E
Goshogawara	40 48 N	140 27 E
Gotemba	35 18 N	138 55 E
Gyōda	36 08 N	139 28 E
Haboro	44 22 N	141 42 E
Hachinohe	40 30 N	141 29 E
Hachiōji	35 39 N	139 20 E
Hagi	34 24 N	131 25 E
Hakata	34 07 N	132 57 E
Hakodate	41 45 N	140 43 E
Hakone	35 11 N	139 02 E
Hakui	36 53 N	136 47 E
Hamada	34 53 N	132 05 E
Hamakita	34 49 N	137 47 E
Hamamatsu	34 42 N	137 44 E
Hanamaki	39 23 N	141 07 E
Higashi-Ōsaka	34 40 N	135 36 E
Higashir Rishiri	45 12 N	141 07 E
Himeji	34 49 N	134 42 E
Hirado	33 22 N	129 33 E
Hiratsuka	35 19 N	139 21 E
Hirosaki	40 35 N	140 28 E
Hiroshima	34 24 N	132 27 E
Hita	33 19 N	130 56 E
Hitachi	36 36 N	140 39 E
Hōfu	34 03 N	131 34 E
Hondo	32 27 N	130 12 E
Honjō	39 23 N	140 03 E
Hyūga	32 18 N	131 45 E
Ibusuki	31 13 N	130 36 E
Ichihara	35 25 N	140 07 E
Ichikawa	35 44 N	139 55 E
Ichinomiya	35 18 N	136 48 E
Ichinoseki	38 55 N	141 08 E
Iizuka	33 38 N	130 41 E
Imabari	34 03 N	133 00 E
Imagane	42 25 N	140 01 E
Imari	33 16 N	129 53 E
Ina	35 50 N	137 57 E
Innoshima	34 19 N	133 10 E
Isahaya	32 50 N	130 03 E
Ise	34 29 N	136 42 E
Isesaki	36 19 N	139 12 E
Ishinomaki	38 25 N	141 18 E
Ishioka	36 11 N	140 16 E
Itō	34 58 N	139 05 E
Iwaki	37 05 N	140 50 E
Iwakuni	34 09 N	132 11 E
Iwamizawa	43 12 N	141 46 E
Iwanai	42 58 N	140 30 E
Izuhara	34 12 N	129 17 E
Izumi	38 21 N	140 48 E
Izumi	32 04 N	130 25 E
Izumi-Sano	34 25 N	135 19 E
Izumo	35 22 N	132 46 E
Jyoetsu	37 07 N	138 15 E
Kaga	36 18 N	136 20 E
Kagoshima	31 36 N	130 33 E
Kainan	34 09 N	135 12 E
Kakamigahara	35 24 N	136 53 E
Kamaishi	39 16 N	141 53 E
Kamiiso	41 49 N	140 39 E
Kamiyaku	30 23 N	130 30 E
Kamogawa	35 06 N	140 06 E
Kanazawa	36 34 N	136 39 E
Kanoya	31 23 N	130 51 E
Kanuma	36 33 N	139 44 E
Karatsu	33 26 N	129 58 E
Kariya	34 59 N	136 59 E
Kasai	34 46 N	137 48 E
Kasaoka	34 30 N	133 30 E
Kashiwazaki	37 22 N	138 33 E
Kasugai	35 15 N	137 00 E
Katsuta	36 24 N	140 33 E
Kawaguchi	35 48 N	139 43 E
Kawasaki	35 32 N	139 43 E
Kesennuma	38 54 N	141 35 E
Kikai	28 20 N	129 50 E
Kiryū	36 24 N	139 20 E
Kishiwada	34 28 N	135 22 E
Kita-Ibaraki	36 50 N	140 40 E
Kita-Kyūshū	33 50 N	130 50 E
Kitakami	39 18 N	141 07 E
Kitami	43 48 N	143 54 E
Kōbe	34 41 N	135 10 E
Kōchi	33 33 N	133 33 E
Kōfu	35 39 N	138 35 E
Koga	36 11 N	139 43 E
Kokubu	31 44 N	130 46 E
Komaki	35 17 N	136 55 E
Komatsu	36 24 N	136 27 E
Komatsushima	34 00 N	134 35 E
Kōriyama	37 24 N	140 23 E
Koshigaya	35 54 N	139 47 E
Koza, see Okinawa		
Kudamatsu	34 00 N	131 52 E
Kuji	40 11 N	141 46 E
Kumagaya	36 08 N	139 23 E
Kumamoto	32 48 N	130 43 E
Kumano	33 57 N	136 05 E
Kurashiki	34 35 N	133 46 E
Kurayoshi	35 26 N	133 49 E
Kure	34 14 N	132 34 E
Kuroiso	36 58 N	140 03 E
Kurume	33 19 N	130 31 E
Kushikino	31 44 N	130 16 E
Kushiro	42 58 N	144 23 E
Kutchan	42 54 N	140 45 E
Kuwana	35 04 N	136 42 E
Kyōto	35 00 N	135 45 E
Machida	35 34 N	139 25 E
Maebashi	36 23 N	139 04 E
Maizuru	35 27 N	135 20 E
Makurazaki	31 16 N	130 19 E
Marugame	34 17 N	133 47 E
Masuda	34 40 N	131 51 E
Matsudo	35 47 N	139 54 E
Matsue	35 28 N	133 04 E
Matsumae	41 26 N	140 07 E
Matsumoto	36 14 N	137 58 E
Matsutō	36 31 N	136 33 E
Matsuyama	33 50 N	132 45 E
Matsuzaka	34 34 N	136 32 E
Mihara	34 24 N	133 05 E
Minamata	32 13 N	130 24 E
Minatomachi	43 12 N	140 46 E
Mine	34 11 N	131 12 E
Mino	35 32 N	136 55 E
Misawa	40 41 N	141 24 E
Mishima	35 07 N	138 55 E
Mito	36 22 N	140 28 E
Miyako	39 38 N	141 57 E
Miyakonojō	31 44 N	131 04 E
Miyazaki	31 52 N	131 25 E
Miyoshi	34 48 N	132 51 E
Mizusawa	39 08 N	141 08 E
Mobara	35 25 N	140 18 E
Mombetsu	44 21 N	143 22 E
Mori	42 06 N	140 35 E
Morioka	39 42 N	141 09 E
Murakami	38 14 N	139 29 E
Muroran	42 18 N	140 59 E
Muroto	33 22 N	134 10 E
Mutsu	41 17 N	141 13 E
Nachi-Katsuura	33 39 N	135 53 E
Nagano	36 39 N	138 11 E
Nagaoka	37 27 N	138 51 E
Nagasaki	32 48 N	129 55 E
Nagato	34 23 N	131 12 E
Nago	26 35 N	127 59 E
Nagoya	35 10 N	136 55 E
Naha	26 13 N	127 40 E
Nakamura	32 59 N	132 56 E
Nakatsu	33 34 N	131 13 E
Nanao	37 03 N	136 58 E
Nankoku	33 35 N	133 39 E
Nara	34 41 N	135 50 E
Narashino	35 41 N	140 03 E
Naruto	34 11 N	134 37 E
Nase	28 23 N	129 30 E
Natori	38 10 N	140 53 E
Nayoro	44 21 N	142 28 E
Nemuro	43 20 N	145 35 E
Nichinan	31 39 N	131 19 E
Niigata	37 55 N	139 03 E
Niihama	33 58 N	133 16 E
Niitsu	37 48 N	139 07 E
Nikkō	36 45 N	139 37 E
Nishinomiya	34 43 N	135 20 E
Nishinoomote	30 44 N	131 00 E
Nishio	34 52 N	137 03 E
Nobeoka	32 35 N	131 40 E
Noboribetsu	42 27 N	141 11 E
Nōgata	33 44 N	130 44 E
Noshiro	40 12 N	140 02 E
Numata	36 38 N	139 03 E
Numazu	35 06 N	138 52 E
Obama	35 30 N	135 45 E
Obihiro	42 55 N	143 12 E
Oda	35 17 N	132 37 E
Ōdate	40 18 N	140 33 E
Odawara	35 15 N	139 10 E
Ōfunato	39 04 N	141 43 E
Ōgaki	35 21 N	136 37 E
Ōhata	41 24 N	141 10 E
Ōita	33 14 N	131 36 E
Ōka	34 06 N	136 13 E
Ōkawa	33 12 N	130 22 E
Okaya	36 03 N	138 03 E
Okayama	34 39 N	133 55 E
Okazaki	34 57 N	137 10 E
Okinawa (Koza)	26 21 N	127 48 E
Ōmiya	35 54 N	139 38 E
Ōmura	32 54 N	129 57 E
Ōmuta	33 02 N	130 27 E
Ōno	35 59 N	136 29 E
Onomichi	34 25 N	133 12 E
Ōsaka	34 40 N	135 30 E
Ōta	36 18 N	139 22 E
Otaru	43 13 N	141 00 E
Otofuke	42 59 N	143 12 E
Ōtsu	35 00 N	135 52 E
Owase	34 02 N	136 11 E
Oyama	36 18 N	139 48 E
Rumoi	43 56 N	141 39 E
Ryōtsu	38 05 N	138 26 E
Sabae	35 57 N	136 11 E
Saga	33 15 N	130 18 E
Sagamihara	35 33 N	139 21 E
Saijō	33 55 N	133 11 E
Saiki	32 58 N	131 53 E
Saito	32 10 N	131 21 E
Sakai	34 35 N	135 28 E
Sakaide	34 19 N	133 54 E
Sakata	38 55 N	139 50 E
Sanjō	37 37 N	138 57 E
Sapporo	43 03 N	141 21 E
Sasebo	33 10 N	129 43 E
Sawara	35 53 N	140 30 E
Sendai	31 49 N	130 18 E

MAP INDEX (continued)

Sendai	38 15 N 140 53 E
Seto	35 14 N 137 06 E
Setouchi	28 08 N 129 15 E
Shari	43 55 N 144 40 E
Shibata	37 57 N 139 20 E
Shibetsu	43 40 N 145 08 E
Shibetsu	44 10 N 142 23 E
Shimada	34 49 N 138 11 E
Shimizu	35 01 N 138 29 E
Shimoda	34 40 N 138 57 E
Shimonoseki	33 57 N 130 57 E
Shinjō	38 46 N 140 18 E
Shiogama	38 19 N 141 01 E
Shirakawa	37 07 N 140 13 E
Shiranuka	42 57 N 144 05 E
Shiroishi	38 00 N 140 37 E
Shizunai	42 20 N 142 22 E
Shizuoka	34 58 N 138 23 E
Sukagawa	37 17 N 140 23 E
Sukumo	32 56 N 132 44 E
Susaki	33 22 N 133 17 E
Suwa	36 02 N 138 08 E
Suzu	37 25 N 137 15 E
Suzuka	34 52 N 136 32 E
Takahashi	34 40 N 133 35 E
Takamatsu	34 20 N 134 03 E
Takaoka	36 45 N 137 01 E
Takasaki	36 20 N 139 01 E
Takatsuki	34 51 N 135 37 E
Takayama	36 08 N 137 15 E
Takefu	35 54 N 136 10 E
Takikawa	43 33 N 141 54 E
Tamano	34 28 N 133 56 E
Tanabe	33 44 N 135 22 E
Tateyama	34 59 N 139 52 E
Tendō	38 21 N 140 22 E
Tōkamachi	37 08 N 138 46 E
Tokunoshima	27 48 N 128 59 E
Tokushima	34 04 N 134 34 E
Tokuyama	34 03 N 131 49 E
Tokyo	35 42 N 139 46 E
Tomakomai	42 38 N 141 36 E
Tosa	34 26 N 135 48 E
Tosu	33 22 N 130 31 E
Tottori	35 30 N 134 14 E
Towada	40 18 N 140 52 E
Toyama	36 41 N 137 13 E
Tōyō	33 22 N 134 18 E
Toyohashi	34 46 N 137 23 E
Toyokawa	34 49 N 137 24 E
Toyooka	35 32 N 134 50 E
Toyota	35 05 N 137 10 E
Tsu	34 43 N 136 31 E
Tsuchiura	36 05 N 140 12 E
Tsukumi	33 04 N 131 52 E
Tsuruga	35 39 N 136 04 E
Tsuruoka	38 44 N 139 50 E
Tsuyama	35 03 N 134 00 E
Ube	33 56 N 131 15 E
Ueda	36 24 N 138 16 E
Ueno	34 45 N 136 08 E
Uji	34 53 N 135 48 E
Uozu	36 48 N 137 24 E
Urakawa	42 09 N 142 47 E
Urasoe	26 15 N 127 43 E
Urawa	35 51 N 139 39 E
Usa	33 27 N 133 27 E
Ushibuka	32 11 N 130 01 E
Usuki	33 08 N 131 49 E
Utsunomiya	36 33 N 139 52 E
Uwajima	33 13 N 132 34 E
Wajima	37 24 N 136 54 E
Wakayama	34 13 N 135 11 E
Wakkanai	45 25 N 141 40 E
Yaizu	34 52 N 138 20 E
Yamagata	38 15 N 140 20 E
Yamaguchi	34 10 N 131 29 E
Yao	34 37 N 135 36 E
Yasugi	35 26 N 133 15 E
Yatsushiro	32 30 N 130 36 E
Yawatahama	33 27 N 132 25 E
Yokkaichi	34 58 N 136 37 E
Yokohama	35 27 N 139 39 E
Yokosuka	35 18 N 139 40 E
Yokote	39 18 N 140 34 E
Yonago	35 26 N 133 20 E
Yonezawa	37 55 N 140 07 E
Yūbari	43 04 N 141 59 E

Physical features and points of interest

Abukuma, *river*	37 50 N 140 52 E
Abukuma Range	37 30 N 140 45 E
Agano, *river*	37 57 N 139 08 E
Akaishi Range	35 25 N 138 10 E
Akan National Park	43 30 N 144 15 E
Amakusa Islands	32 20 N 130 05 E
Amami Island	28 15 N 129 20 E
Asahi, *river*	35 00 N 133 58 E
Asahi, Mount, *volcano*	43 40 N 142 51 E
Aso, Mount, *volcano*	32 53 N 131 06 E
Aso National Park	33 00 N 131 07 E
Awaji Island	34 20 N 134 50 E
Bandai-Asahi National Park	38 16 N 139 57 E
Beyoneisu Rocks	31 55 N 139 55 E
Biwa, Lake	35 15 N 136 05 E
Bōsō Peninsula	35 20 N 140 10 E
Bungo Channel	33 00 N 132 13 E
Chichibu-Tama National Park	35 52 N 139 00 E
Chikugo, *river*	33 09 N 130 21 E
Chūbu Sangaku National Park	36 30 N 137 41 E
Chūgoku Range	35 15 N 134 00 E
Dai, Mount, *volcano*	35 22 N 133 32 E
Daisetsu Mountains	43 41 N 142 53 E
Daisetsuzan National Park	43 30 N 142 57 E
East China Sea	29 00 N 127 00 E
Erimo, Cape	41 55 N 143 15 E
Fuji, Mount, *volcano*	35 22 N 138 44 E
Fuji-Hakone-Izu National Park	35 21 N 138 44 E
Gonō, *river*	35 02 N 132 13 E
Gotō Islands	32 50 N 129 00 E
Hakkōda Mountains	40 40 N 140 53 E
Hamana, Lake	34 45 N 137 34 E
Hidaka Range	42 35 N 142 45 E
Hokkaido, *island*	44 00 N 143 00 E
Honshu, *island*	36 00 N 138 00 E
Hotaka, Mount	36 17 N 137 39 E
Iki, *island*	33 47 N 129 43 E
Inawashiro, Lake	37 29 N 140 06 E
Inland Sea	34 20 N 133 30 E
Ise Bay	34 43 N 136 43 E
Ise-Shima National Park	34 23 N 136 48 E
Ishikari, *river*	43 15 N 141 23 E
Ishikari Bay	43 25 N 141 01 E
Ishikari Plain	43 00 N 141 40 E
Izu Islands	34 20 N 139 40 E
Izu Peninsula	34 45 N 139 00 E
Japan, Sea of	37 00 N 135 45 E
Japanese Alps, *mountains*	36 00 N 137 00 E
Jō-shin-etsu National Park	36 46 N 138 40 E
Kamui, Cape	43 20 N 140 21 E
Kantō Plain	36 00 N 139 30 E
Kantō Range	35 50 N 138 50 E
Kasumi, Lake	36 00 N 140 25 E
Kerama Islands	26 05 N 127 15 E
Kii Mountains	34 00 N 135 50 E
Kii Peninsula	34 00 N 135 45 E
Kiso, *river*	35 02 N 137 30 E
Kiso Range	35 40 N 137 45 E
Kitakami, *river*	38 40 N 141 19 E
Kitakami Range	39 30 N 141 30 E
Kitami Mountains	44 22 N 142 43 E
Korea Strait	34 00 N 129 00 E
Koshiki Islands	31 46 N 129 48 E
Koza, *river*	33 31 N 135 50 E
Kutcharo, Lake	43 38 N 144 21 E
Kyushu, *island*	33 00 N 131 00 E
Kyushu Mountains	32 35 N 131 05 E
La Perouse (Sōya) Strait	45 45 N 142 00 E
Mogami, *river*	38 55 N 139 48 E
Naka, *river*	36 20 N 140 36 E
Nemuro Strait	44 00 N 145 20 E
Niigata Plain	37 45 N 139 00 E
Nikkō National Park	36 49 N 139 33 E
Nōbi Plains	35 15 N 137 30 E
Noto Peninsula	37 20 N 137 00 E
Ō Island	34 44 N 139 24 E
Oga Peninsula	39 55 N 139 50 E
Okhotsk, Sea of	45 00 N 145 00 E
Oki Islands	36 15 N 133 15 E
Okinawa, *island*	26 42 N 128 11 E
Ōkino-Erabu Island	27 22 N 128 35 E
Ōnaruto Bridge	34 13 N 134 39 E
Ontake, Mount, *volcano*	35 53 N 137 39 E
Ōsaka Bay	34 30 N 135 18 E
Oshima Peninsula	42 00 N 140 30 E
Ōsumi Islands	30 30 N 130 00 E
Ōsumi Strait	31 00 N 131 00 E
Ōu Range	39 00 N 140 45 E
Pacific Ocean	32 00 N 138 00 E
Rishiri Island	45 11 N 141 15 E
Ryukyu Islands	29 00 N 129 00 E
Sado, *island*	38 00 N 138 25 E
Sagami Bay	35 15 N 139 25 E
Saroma, Lake	44 08 N 143 50 E
Seikan Tunnel	41 17 N 140 49 E
Seto Bridge	34 23 N 133 40 E
Shikoku, *island*	33 45 N 133 30 E
Shikoku Mountains	33 50 N 133 26 E
Shikotsu, Lake	42 45 N 141 20 E
Shikotsu-Tōya National Park	42 47 N 141 10 E
Shinano, *river*	37 00 N 138 45 E
Shiretoke, Cape	44 21 N 145 20 E
Shōwa Volcano	42 32 N 140 52 E
Sōya, Cape	45 31 N 141 56 E
Sōya Strait, see La Perouse Strait	
Suō Sea	33 50 N 131 30 E
Suwa, Lake	36 03 N 138 05 E
Tanega Island	30 40 N 131 00 E
Tenryū, *river*	34 39 N 137 48 E
Teshio, *river*	44 40 N 142 20 E
Teshio Range	44 15 N 142 05 E
Tokachi, *river*	42 44 N 143 42 E
Tokachi Plain	43 00 N 143 30 E
Tokara Islands	29 37 N 129 43 E
Tokoro, *river*	43 40 N 143 45 E
Tokuno Island	27 45 N 128 58 E
Tokyo Bay	35 25 N 139 47 E
Tone, *river*	35 44 N 140 51 E
Tosa Bay	33 20 N 133 40 E
Towada, Lake	40 28 N 140 55 E
Towada-Hachimantai National Park	40 35 N 140 53 E
Tōya, Lake	42 35 N 140 51 E
Tsugaru Strait	41 35 N 141 00 E
Tsushima, *islands*	34 20 N 129 20 E
Tsushima Strait	34 40 N 129 00 E
Uchiura Bay	42 20 N 140 40 E
Unzen, Mount, *volcano*	32 45 N 130 17 E
Yaku Island	30 20 N 130 30 E
Yodo, *river*	34 41 N 135 31 E

Coalescing alluvial fans—cone-shaped deposits of alluvium that run together—are formed where rivers emerge from the mountains. When the rivers are large enough to extend their courses to the sea, low deltaic plains develop in front of the fans; this occurs most frequently where the rivers empty into shallow and sheltered bays, as in the deltas of Kantō (Kwanto), Nōbi, and Ōsaka. In most places, however, fan surfaces plunge directly into the sea and are separated by low, sandy beach ridges.

Dissected plains are common. Intense disturbances have caused many former alluvial fans, deltas, and sea bottoms to be substantially uplifted to form flat-topped uplands such as those found in the Kantō Plain (Kantō-heiya). Frequently the uplands have been overlain with volcanic ash, as in the Kantō and Tokachi plains.

In addition, Japan is truly a land of volcanoes. Violent eruptions are frequently experienced, and at least 60 volcanoes have been active within historical time. New volcanoes born during the 20th century include Shōwa-shinzan (Shōwa Volcano) on Hokkaido and Myōjin-shō off the Beyoneisu (or Bayonnaise) Rocks in the Pacific. The abundant hot springs are mostly of volcanic origin. Many of the gigantic volcanoes are conical in shape (*e.g.,* Fuji-san [Mount Fuji]), while others form steep lava domes (*e.g.,* Dai-sen and Unzen). Conspicuous shield volcanoes (broad, gently sloping volcanic cones such as Mauna Loa on Hawaii Island) are rare, and extensive lava plateaus are lacking. One of the characteristics of the volcanic areas is the prevalence of calderas (large, circular, basin-shaped volcanic depressions), especially in the northeast and southwest, many of which are occupied by lakes, such as Kutcharo, Towada, and Ashi.

Japan's mountains have been influenced by the orogenic formation of six mountain arcs off the Pacific coast of Asia. They are, from northeast to southwest, the Chishima Range of the Kuril Islands; the Karafuto (Sakhalin) Mountain system of Hokkaido; the Northeast, Southwest, and Shichito-Mariana ranges of Japan; and the Ryukyu (Nansei) Island formations. The four major landform areas of Japan—the Hokkaido, Northeastern, Central, and Southwestern regions—have developed as a result of the formation of these arcs.

Four major landform regions

The Hokkaido region was formed by the coalescence of the Chishima and Karafuto arcs. The backbone of the region runs from north to south. The Chishima arc enters Hokkaido as three volcanic chains with elevations of more than 6,000 feet (1,800 metres); these are arranged in ladder formation and terminate in the heart of the region.

Chief components of the mountain system are the Kitami-sanchi (Kitami Mountains) in the north and the Hidaka-sammyaku (Hidaka Range) in the south.

The Northeastern Region nearly coincides with the northeastern mountain arc and stretches from southwest Hokkaido to central Honshu. Several rows of mountains, lowlands, and volcanic zones are well oriented to the general trend of the insular arc of this region, which is convex toward the Pacific Ocean. The Kitakami and Abukuma ranges on the east coast are somewhat oblique to the general trend; they are chiefly composed of older rocks, and plateau-like landforms survive in the centre. In the western zone, the formations conform to the general trend and are composed of a basement complex overlain by thick accumulations of young rocks that have been subjected to mild folding. The Ōu-sammyaku, capped with towering volcanoes that form the Nasu-kazantai (Nasu Volcanic Zone), is separated from the coastal ranges by the Kitakami-Abukuma lowlands to the east and by a row of basins in the west.

The Central Region of Honshu is dominated by the overlapping of the Northeast, Southwest, and Shichito-Mariana mountain arcs. It contains Japan's highest mountain, Fuji-san, which rises to 12,388 feet (3,776 metres), and its broadest width of 168 miles (270 kilometres). The trend of the mountains, lowlands, and volcanic zones intersects the island almost at right angles. The most notable physical feature is the Fossa Magna, a great rift lowland that traverses the widest portion of Honshu from the Sea of Japan to the Pacific. It is partially occupied by mountains and volcanoes of the Fuji-kazantai. Intermontane basins are sandwiched between the lofty, partially glaciated central mountain knots of the Akaishi, Kiso, and Hida ranges (which together form the Japanese Alps) to the west and the Kantō-sanchi to the east. The shallow structural basin of the Kantō Plain, which stretches to the east of the Kantō-sanchi, is the most extensive lowland of Japan; the immense metropolis of Tokyo spreads out from its centre, covering a vast area of the plain.

The Southwestern Region of southern Honshu, Shikoku, and northern Kyushu generally coincides with the southwest mountain arc, and the general trend of highlands and lowlands is roughly convex toward the Sea of Japan. The region is divided into the Inner Zone, formed by complex faulting, and the Outer Zone, formed by warping. The Inner Zone is chiefly composed of granite, Paleozoic (225,000,000–570,000,000 years old), and volcanic rocks, which are arranged in complicated juxtaposition. The Outer Zone, consisting of the Akaishi, Kii, Shikoku, and Kyushu mountain groups, in contrast, is characterized by a regular zonal arrangement from north to south of crystalline schists, Paleozoic, Mesozoic (65,000,000–225,-000,000 years old), and Tertiary (2,500,000–65,000,000 years old) formations. The outstanding surface features of the Inner Zone present a highly complex mosaic of numerous fault blocks, while those of the Outer Zone are continuous except where the sea straits separate them into the four independent groups. The Inland Sea (Seto-naikai) is the region where the greater amount of depression has resulted in the invasion of sea waters. The northern edge of the Inner Zone is studded with the gigantic lava domes of the Daisen-kazantai, which, together with volcanic Aso-san, bury a considerable part of the western extension of the Inland Sea in central Kyushu.

T. Okuda—Aspect Picture Library, London

Taishō Pond in Kamikōchi Valley, Chūbu Region. Beyond are the peaks of Hotaka-dake, highest mountain in the Hida Range, which is the northernmost range in the Japanese Alps.

The Ryukyu Islands Region constitutes the main portion of the Ryukyu arc, which penetrates into Kyushu as the Kirishima-kazantai and terminates at Aso-san (Mount Aso). The influence of the arc is also seen in the trend of the many elongated islands off western Kyushu, including the Koshiki, Gotō, and Tsushima islands. The islands of the Shichito-Ogasawara region, to the east of the Ryukyu arc, consist of a number of volcanoes on the submarine ridge of the Shichito-Marina arc and the Bonin Islands (in Japanese Ogasawara-guntō), which include Peel Island and Iwo Jima (Iō-jima).

Drainage and soils. The increasing demand for fresh water because of the rice culture, industrialization, and urbanization is a serious problem. Difficulties of supply lie in the paucity of natural water reservoirs, the swift runoff of the rivers, and the engineering difficulties of constructing large-scale dams in the rugged mountains.

Japan's rivers are generally short and swift-running and are supplied by small drainage basins. The most significant rivers are the Teshio and Ishikari rivers of Hokkaido; the Kitakami, Tone, Shinano, Kiso, and Tenryū rivers of Honshu; and the Chikugo of Kyushu. Some of the rivers from the volcanic areas of northeastern Honshu are acidic and are useless for irrigation and other purposes.

Major rivers

Biwa-ko (Lake Biwa), the largest in Japan, covers 260 square miles of central Honshu. All other major lakes are in the northeast. Most of the coastal lakes, such as Kasumiga-ura (Lake Kasumi) and Hamana-ko of Honshu, are drowned former valleys, the bay mouths of which have been dammed by sandbars. Inland lakes such as Biwa, Suwa, and Inawashiro of Honshu occupy tectonic depressions of recent fault origin. Lakes of volcanic origin (*e.g.,* Kutcharo of Hokkaido and Towada and Ashi of Honshu) outnumber all other types.

The soils of Japan are customarily divided from northeast to southwest into a weak podzolic (soils with a thin organic mineral layer over a gray leached layer) zone, a brown-earth zone, and a red-earth zone. There are some local variations. The northern half of the Tōhoku area of northern Honshu is included in the brown forest soil area. The northern tip of Hokkaido is classed as a subzone of the podzolic soils; the remainder of the island is included in the subzone of the acidic brown forest soils; and most of western Honshu is a transitional zone. Yellow-brown forest soils extend along the Pacific coast from southern Tōhoku to southern Kyushu, while red and yellow soils are confined to the Ryukyu Islands.

Kuroboku soils (black soils rich in humus content) are found on terraces, hills, and gentle slopes throughout Japan, while gley (sticky, blue-gray compact) soils are found in the poorly drained lowlands. Peat soils occupy the moors in Hokkaido and Tōhoku. Muck (dark soil, containing a high percentage of organic matter) and gley paddy soils are the products of years of rice culture. Polder soils (those reclaimed from the sea) are widely distributed, and immature volcanic ash soils are found on the uplands. The widespread reddish soils are generally regarded as the products of a former warm, humid climate.

Climate. Japan's present climate is influenced by the country's latitudinal extent, the surrounding oceans, and the neighbouring Asian landmass, whereas local climatic variations are the result of relief features. In winter, the high pressure zone over eastern Siberia and the low pressure zone over the western Pacific result in an eastward flow of cold air (the winter monsoon) from late September to late March that picks up moisture over the Sea of Japan. The winter monsoon deposits its moisture as rain or snow on the side of Japan facing the Sea of Japan and brings dry, windy weather to the Pacific side. The pressure systems are reversed during the summer, and air movements from the east and south (the summer monsoon) from mid-April to early September bring warmer temperatures and rain. Cyclonic storms and frequent and destructive typhoons occur during the summer and early fall, especially in the southwest.

The effects of ocean currents

The warm waters of the Kuroshio (Japan) Current, which corresponds to the Gulf Stream of the Atlantic, flow northward along Japan's Pacific coast as far as latitude 35° N. The Tsushima Current branches westward from the Kuroshio Current off southern Kyushu and washes the coasts of Honshu and Hokkaido along the Sea of Japan; it is this current that lends moisture to the winter monsoon. The counterpart of the Labrador Current, the cold Oyashio (Chishima) Current, flows southeastward from the Bering Sea along the east coast of Hokkaido and northeastern Honshu. Its waters meet those of the Kuroshio Current, causing dense sea fogs in summer, especially off Hokkaido.

The chief physical feature to affect climate is the mountainous backbone of the islands. The ranges interrupt air flow from the northwest and southeast and cause the gloomy weather and heavy snows of winter along the Sea of Japan coast and the bright and windy winter weather along the Pacific. Temperatures and annual precipitation are about the same on both coasts, but they drop noticeably in the mountainous interior.

Temperatures are generally warmer in the south than in the north, and the transitional seasons of spring and fall are shorter in the north. At Asahikawa, in central Hokkaido, the mean temperature in January, the coldest month, is 16° F (−9° C), and the mean temperature in August, the hottest month, is 70° F (21° C), with an annual average temperature of 43° F (6° C). At Tokyo, the mean temperature for January is 39° F (4° C), the mean for August 81° F (27° C), and the annual average 59° F (15° C). Inland from Tokyo, Matsumoto is cooler, with an annual average temperature of 52° F (11° C); whereas an annual average of 57° F (14° C) occurs on the Sea of Japan coast at Kanazawa. The warmest temperatures occur on Kyushu and the southern islands; at Kagoshima, the mean temperature for January is 45° F (7° C), the mean for August is 81° F (27° C), and the average is 63° F (17° C).

Precipitation in the form of rain and snow is plentiful throughout the islands. Maximum precipitation falls in the early summer, and the minimum occurs in winter except on the Sea of Japan coast, which receives the country's highest snowfall. The summer rainy season occurs through June and July; it is known as the *baiu* ("plum rain") because it begins when the plums ripen. Torrential rains accompany the typhoons.

Rainfall patterns vary with topography, but most of the country receives more than 40 inches (1,020 millimetres) of precipitation annually. The smallest amount of precipitation occurs on eastern Hokkaido, where only 37 inches fall annually at Obihiro, whereas the mountainous interior of the Kii-hantō (Kii Peninsula) of central Honshu receives more than 160 inches annually. Varying amounts of snow fall on Japan. From November to April snow blankets Hokkaido, northern and interior Honshu, and the northwest coast.

Plant and animal life. Much of the original vegetation has been replaced by agriculture or by the introduction of foreign species to the islands. Semitropical rain forest prevails in the Ryukyu and Bonin islands and contains various kinds of mulberries, camphor, oaks, and ferns (including tree ferns); madder and lianas are found as undergrowth. In the Amami-shotō (Amami Islands) this type of plant life occurs only on lowlands, but it grows at higher altitudes southward. There are a few mangrove swamps along the southern coast of Kyushu.

Vegetational regions

The laurel forest zone of evergreen, broad-leaved trees extends from the southwestern islands northward to the lowlands of northern Honshu. Camphor, pasania, Japanese evergreen oak, camellia, and holly are typical trees, and various kinds of ferns grow as undergrowth. In Kyushu, the evergreen zone reaches to more than 3,300 feet, but its altitudinal limit decreases northeastward across Honshu. In general, camphor dominates in the littoral lowlands; pasania, in sunny and well-drained sites; and Japanese evergreen oak, in the foggy and cloudy inlands. In the southwestern Hondo region (Honshu, Shikoku, and Kyushu) are ficus and fan palm. The coastal dunes are dominated by pine trees. Natural stands of Japanese cedar, some containing trees that are more than 2,000 years old, occur above 2,300 feet on Yaku-shima, south of Kyushu.

Deciduous broad-leaved forests develop in the higher and northward portions of the laurel forest zone. In Kyushu, this type of forest occurs above 3,300 feet, but it gradually descends northward to meet the shoreline in northern Honshu. Its upper limit reaches 6,000 feet in Shikoku and 5,000 feet in central Honshu. The representative trees are beech, katsura tree, maple, oak, and birch; while various kinds of bamboo grasses grow as undergrowth. All of these trees, but especially the maples, are admired for their beautiful fall colours. The trees have been occasionally replaced by larch, false cypress, false arborvitae, Japanese cedar, Japanese red pine, Japanese black pine, and other coniferous species. The deciduous zone extends into western Hokkaido, where beeches terminate at the southwestern peninsula, and further northeastward is replaced by basswood and maple. Some stands of conifers are mixed with the representative forests of this zone.

Coniferous trees are numerous in the north and eastern periphery of Hokkaido up to 2,300 feet above sea level. Sakhalin spruce, Sakhalin fir, blue fir, and Yezo spruce are mixed with such deciduous trees as birch, oak, and maple and dense undergrowth of mosses and lichens. Coniferous trees are mixed with deciduous vegetation in southwestern Hokkaido and occur in the higher portion of central Honshu and Shikoku.

High-altitude small shrubs, the creeping pine, and alpine plants grow in the high mountain knots of central Honshu above 8,000 feet. This zone gradually descends northward to Hakkōda-san, in northern Honshu, at 4,600 feet and to Daisetsu-zan, in central Hokkaido, at about 3,600 feet.

The cherry tree, celebrated for its spring blossoms, is planted all about the country. Many varieties have been cultivated, and natural stands are also found in the mountains.

Animal life

Despite the country's large human population, the land mammals of Japan are relatively numerous in the remote, heavily forested mountain regions. These animals include bears, wild boars, raccoon dogs (*tanuki*), foxes, deer, antelope, hares, and weasels; some species are distinct from those of the neighbouring Asian continent. Wild monkeys (the Japanese macaque) inhabit many places; those found at the northern tip of Honshu represent the northern limit of monkey habitation in the world.

Japanese waters are inhabited by whales, dolphins, porpoises, and fish such as salmon, sardines, sea bream, mackerel, tuna, trout, herring, grey mullets, smelts, and cod. Crustaceans include crabs, shrimp, prawns, clams, and oysters and are important as food; clams and oysters are raised commercially. The rivers and lakes abound in trout, salmon, and crayfish. In addition, several varieties of seaweed are collected for food.

Reptiles include sea turtles, freshwater tortoises, sea snakes, and lizards. There are two species of poisonous snakes, but most of the snakes, including the five-foot-long Japanese rat snake, are harmless. Toads, frogs, and newts are common; and the Japanese giant salamander of Kyushu and Honshu attains a length of four feet.

Water birds include gulls, auks, grebes, albatrosses, shearwaters, herons, ducks, geese, swans, and cranes. The cormorant is sometimes trained to catch fish. There are about 150 species of songbirds, as well as eagles, hawks, falcons, pheasant, ptarmigan, quail, owls, and woodpeckers. Storks

Cherry trees in blossom on a small street in Kyōto.
Lonnie Duka—TSW-CLICK/Chicago

and the Japanese crested ibis (*toki*), once abundant, have nearly become extinct.

Administrative regions. The concept of regions in Japan is inseparable from the historic development of administrative units. Care was always taken to include various physical features in the larger administrative units so as to create a well-balanced geographic whole. Many of the ancient terms for administrative units have survived in the form of place-names.

The origin of Japan's regions

The Taika reforms of 646 established the *ri* (roughly corresponding to the later village community) as the basic social and economic unit and the *gun* (district) as the smallest political unit to be governed by the central government. The *gun* were grouped to form more than 60 *kuni* (provinces), the largest political units, which were ruled by governors appointed by the central government. Each *kuni* was composed of maritime plains, interior basins, and mountains to constitute a more or less independent geographic entity. Several adjacent *kuni* that were linked by a trunk road or a convenient sea route were grouped into a *dō;* the term signified both the route and the region. The core region of the country was called the Kinai, or the land adjacent to the shifting Imperial capitals.

During the Nara (710–784) and Heian (794–1185) periods, the region of Honshu to the east of the three great mountain barriers of Arachi, Fuwa, and Suzuka north, east, and southeast of Biwa-ko was called Kantō (*kan,* "barrier"; *tō,* "east"); and that to the west Kansai (*kan,* "barrier"; *sai,* "west"). As the empire's frontier shifted to the northeast, Kantō came to indicate the region to the east of Hakone barrier, and Kansai gradually came to include limited areas near the capital of Kyōto as far as Ōsaka and present Kōbe. Northern areas that had not come under direct control of the central government were called *Yezochi,* "the land of non-subjugated people."

A third regional system was applied after the 10th century, in which *kuni* were amalgamated according to their distance from Kyōto. The larger units were *kingoku,* or proximate *kuni; chūgoku,* or intermediate *kuni;* and *engoku,* or remote *kuni.* Mutsu and Dewa in northeastern Honshu and islands such as Sado, Oki, Tsushima, and Iki were termed *henkyō,* or peripheral, lands.

In 1871 the feudal system was dissolved and the *ken,* or prefectural, system was established. At first the more than 300 prefectures were mostly the former fiefs of feudal lords, who were appointed as governors. Through amalgamation and partition there were frequent changes in the *ken* pattern, until by 1888 the present configuration of 43 *ken* (including Okinawa), three *fu* (urban prefectures) of Tokyo, Ōsaka, and Kōbe, and one *dō* (Hokkaido) was established; in 1943 Tokyo was given the status of *to,* or metropolis.

Modern regional units

Early in the 20th century it was recognized that larger geographical divisions were needed. By 1905 a system of eight *chihō* (regions) had been set up that divided the country from northeast to southwest. The *chihō* are Hokkaido; Tōhoku (northern Honshu); Kantō (eastern Honshu); Chūbu (central Honshu); Kinki (west central Honshu); Chūgoku (western Honshu); Shikoku; and Kyushu (including the Ryukyus). Another system used by some governmental agencies is a modification of the *chihō* system. Chūbu-chihō, for example, is subdivided into Hokuriku, Tōsan, and Tōkai. This system is devised so as to group prefectures of similar geographic character into one *chihō* and is more effective for illustrating regional contrasts and comparing statistics. In addition, planners have come to refer to the string of industrialized and urbanized areas along the Pacific seaboard between Kantō-chihō and northern Kyushu as the Pacific Belt Zone (Taihei-yō Beruto Chitai). This zone includes most of the Japanese cities with populations of more than 1,000,000, as well as more than half of the country's total population.

Settlement patterns. Since the late 19th century, economic and social changes have affected even the most remote rural villages, but many things Japanese have survived. In the villages, many features that are in common with those of other Asian villages are well preserved. Autonomous and cooperative systems of agricultural practices and rituals, as well as mutual assistance among the villagers, have been handed down to the present. An autonomous rural unit, generally known as a *mura,* consists of some 30 to 50 or more households. Now called an *aza,* this unit should not be confused with the administrative terms *mura* or *son* in use after 1888.

Rural settlement patterns

Most of the rural settlements are of age-old origin, and their histories are lost in time. Historically traceable settlements owe their origins largely to land reclamation after the 16th century. They are commonly called *shinden,* or "new paddy fields," but in terms of social structure they do not radically differ from the older settlements.

Considerable local difference is evident in the settlement pattern. Some villages are agglomerated, as are those of Kinki-chihō; some are dispersed, as in northeastern Shikoku; some are elongated, such as those on the rows of sand dunes in the Niigata-heiya (Niigata Plain) and on the natural levees of deltas; while others are scattered on the steeper mountain slopes. These differences are only superficial; without exception, the inhabitants are bound together to form a firm village community.

No village is regarded as purely rural. Those that are near industrialized urban centres include commuters and industrial workers. The more remote settlements send out seasonal labourers during the winter months. The villages of Hokkaido are based on commercial agriculture, and each household has direct contact with a nearby town.

Fishing villages were absent in Tōhoku-chihō until the beginning of the 17th century, when northward movement began. They were originally dependent upon nearby rice-producing villages; although some dried, salted, or smoked fish found more distant markets. The fishing villages are most numerous in the southwest, where an exchange economy has long been in practice. Mountain villages that depend solely on mountain products other than rice are exceedingly rare. Many of them were founded after the 17th century, when lumber, charcoal, and other such products found markets in the growing towns on the plains. There were also some villages in the mountainous interior of western Tōhoku that relied purely upon hunting, but these have all but disappeared.

Urban centres

Urban settlement is generally of recent origin. Except for the former capital cities of Nara, Kyōto, and Kamakura, no sizable town of any significance appeared before the 16th century. Most of the provincial capitals, or *kokufu,* of ancient Japan were only administrative centres that contained official residences and were not developed towns. After the latter part of the 16th century, influential temples and feudal lords began to build towns by gathering merchants and craftsmen close to their headquarters. The power of the feudal lords stabilized when they built *jōka-machi* (castle towns), which were located so as to command and control the main transportation routes and surrounding areas; the majority of Japan's important cities, including Tokyo, have developed from them.

Next in importance were the port towns, such as Hakata and Sakai, which have experienced more vicissitudes than the castle towns. In addition, some of the religious towns have grown to a considerable size, as in the case of Ise, Kompira, and Zenkōji. Under the regime of the Tokugawa shogunate (1603–1867), peaceful conditions fostered nationwide pilgrimages on a scale unknown in the preceding period, and temple and shrine towns such as Kyōto and Nara flourished.

Widespread urban growth began in the late 19th century with the development of the international ports of Kōbe, Yokohama, Niigata, Hakodate, and Nagasaki and the naval bases of Yokosuka, Kure, and Sasebo. Industrialization ushered in the rapid growth of Japanese cities, and some of the industrial towns (*e.g.,* Yawata, Niihama, Kawasaki, and Amagasaki) were founded in response to economic development. Most of the former castle towns, and especially those along the Pacific belt, have been expanded directly or indirectly by industrialization. In Hokkaido and in southern Kyushu raw materials and power resources have attracted a limited number of industrial plants, which alone are responsible for the existence of cities such as Tomakomai, Muroran, Nobeoka, and Minamata.

Japanese cities are bewildering mixtures of old and new, East and West. Oriental congestion exists side by side with the most modernized business centres and industrial establishments; and the fragmented, patchwork pattern of land ownership is a formidable obstacle in ever-expanding cities of skyscrapers, subways, and underground plazas. Other serious problems are the shortage of better housing, the increasing use of the automobile, overcrowded public transportation systems, the shortage of open space, environmental pollution, and the constant menace of earthquakes and floods. (Ak.W./Y.M.)

THE PEOPLE

Ethnic and linguistic groups. The Japanese people are classified as a branch of the Mongoloid race and are closely akin to the peoples of eastern Asia; they constitute the overwhelming majority of the population. The indigenous Ainu largely have been assimilated into the general population, although small groups survive in Hokkaido. There are two minority groups in the country who generally have not been assimilated into Japanese society and who are subject to varying degrees of discrimination. The larger of these, the *burakumin,* are the descendants of a formerly outcast class of ethnic Japanese. Koreans constitute the second group. Most are descended from Koreans who migrated to Japan in the first half of the 20th century, when Korea was a Japanese colony; despite having been born and raised in Japan, they are classified as resident aliens. The Okinawans form another group, who though they are Japanese citizens are often relegated to a second-class status.

Japanese is the national language, and Ainu is almost extinct. The Japanese language is generally included in the Altaic linguistic group and is especially akin to Korean, although the vocabularies differ. Some linguists also contend that Japanese contains elements of Southeast Asian languages. The introduction of the Chinese writing system and of Chinese literature in about the 4th century AD enriched the Japanese vocabulary. Until that time Japanese had no written form, and at first Chinese characters (called *kanji* in Japanese) were used to write Japanese; but by the 9th century two syllabaries, known collectively as *kana* (*katakana* and *hiragana*), were developed from them. Since then, a combination of *kanji* and *kana* has been used for written Japanese. Although some 3,000 to 5,000 *kanji* are in general use, after World War II the number of characters necessary for a basic vocabulary was reduced to about 2,000, and the writing of these characters was simplified. Several thousand Western loanwords, principally from English, have also been adopted.

The distribution of Japanese nearly coincides with the territory of Japan. Standard Japanese, based on the dialect spoken in Tokyo, was established in the late 19th century through the creation of a national educational system and through more widespread communication. There are many local dialects, which are often mutually unintelligible, but standard Japanese is understood nationwide.

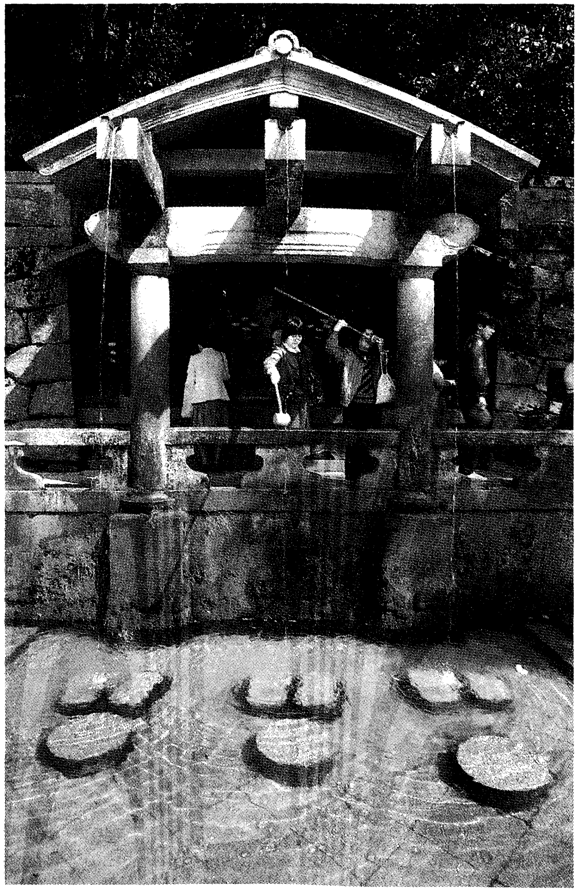
The dialects of the Japanese language

Japanese is broadly divided linguistically into the two major dialects of Hondo and Nantō. The Hondo dialect is used throughout Japan and may be divided into three major subdialects of the east, west, and Kyushu. The Eastern subdialects were established in the 7th and 8th centuries and became known as the *azuma* ("Eastern") language. After the 17th century there was a vigorous influx of the Kamigata (Kinai) dialect, which was the foundation of standard Japanese. Among the Western dialects, Kinki dialect was long the standard language of Japan, although the present Kamigata dialect of the Kyōto–Ōsaka region is of recent origin.

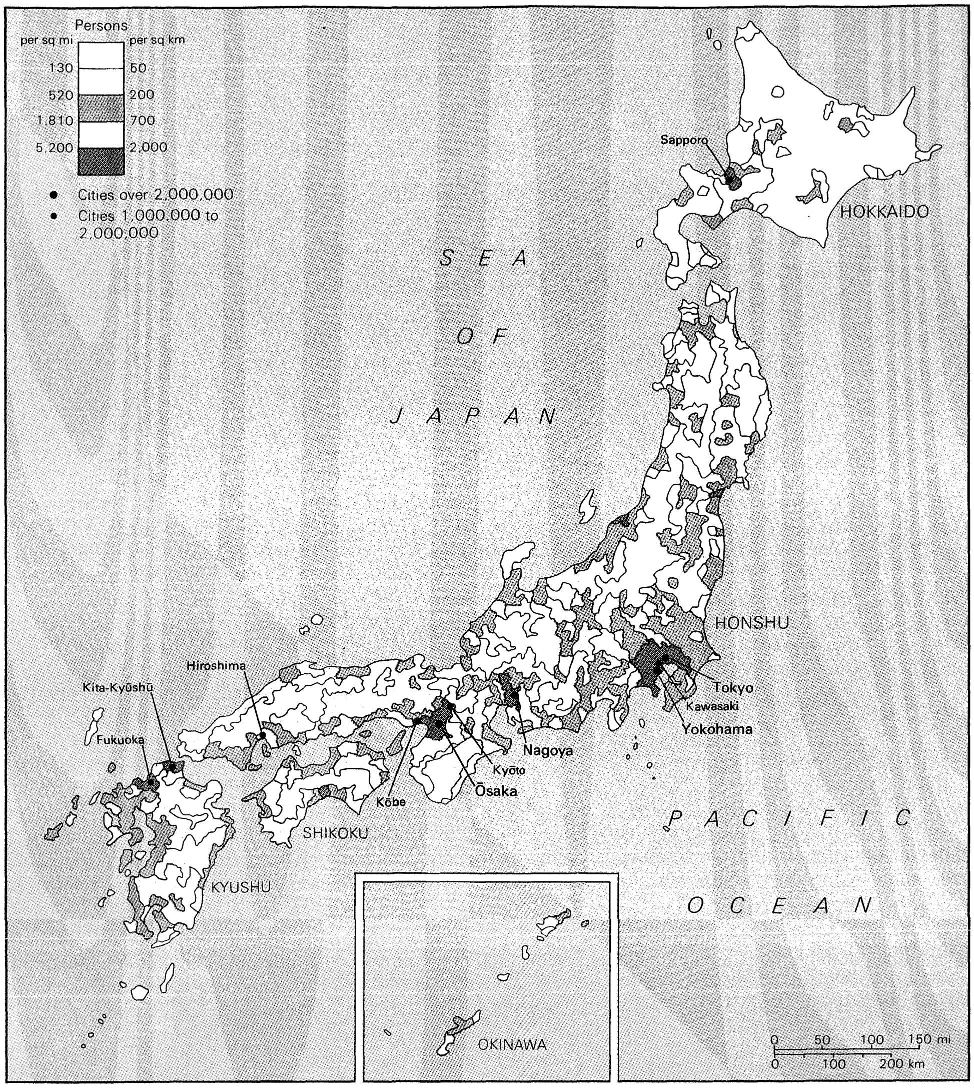
The Kyushu dialects have been placed outside the mainstream of linguistic change of the Western dialects and retain some of the 16th-century forms of the latter. They extend as far south as Tanega and Yaku islands. The Nantō dialects are used by Okinawa islanders from Amami-Ōshima in Kagoshima Prefecture as far west as Yonaguni-jima. Long placed outside the mainstream of linguistic change, they strongly retain their ancient forms.

Religions. In Japan, the indigenous religion, Shintō, various sects of Buddhism, and Christianity exist together with a number of "new religions" (*shinkō shukyō*) that have emerged since the 19th century; not one of the religions is dominant, and each is affected by the others.

D.E. Cox—TSW-CLICK/Chicago



Purification shrine in the Kiyomizu temple, Kyōto.



Population density of Japan.

For example, one person or family may believe in several Shintō gods and at the same time belong to a Buddhist sect. Intense religious feelings are generally lacking except among the adherents of some of the new religions.

Shintō and Buddhism

Shintō is a polytheistic religion. People, commonly major historical figures, as well as natural objects have been enshrined as gods. Some of the Hindu gods and Chinese spirits were also introduced and Japanized. Each rural settlement has at least one shrine of its own; and there are several shrines of national significance, the most important of which is the Grand Shrine of Ise in Mie Prefecture. After the Meiji Restoration, Shintō became a state-supported religion, but this institution was abolished after World War II.

Buddhism was officially introduced into the Imperial court from Korea in the mid-6th century AD. Direct contact with central China was maintained, and several sects were introduced. In the 8th century Buddhism was adopted as the national religion, and national and provincial temples, nunneries, and monasteries were built throughout the country. In the early 9th century, the Tendai and Shingon sects were introduced by Japanese monks who had studied in China. These sects have continued to exert profound influence in some parts of Japan. Zen Buddhism, introduced at the beginning of the Kamakura period, has maintained a large following. Most of the major Buddhist sects of modern Japan, however, have descended from those that were modified in the 13th century by monks such as Shinran, who established an offshoot of Pure Land (Jōdo) Buddhism called the True Pure Land sect (Jōdo Shinshū), and Nichiren, who founded Nichiren Buddhism.

Christianity was introduced into Japan in the 16th century by Roman Catholic missionaries and was well received both as a religion and as a symbol of European culture. After the establishment of the Tokugawa shogunate, Christians were persecuted, and Christianity was totally banned in 1637. Inaccessible and isolated islands and the peninsula of western Kyushu continued to harbour "hiding Christian" villages until the ban was lifted by the Meiji government in 1873. Christianity was reintroduced by Western missionaries, who established many Russian Orthodox, Roman Catholic, and Protestant congregations.

The great majority of what are now called the "new religions" were founded after the mid-19th century. Most have their roots in Shintō, but they have also been influenced by Buddhism, Neo-Confucianism, and Christianity. One of the largest, the Sōka Gakkai ("Value Creation Society"), is based on a sect of Nichiren Buddhism and has powerful political organization. Another new Nichiren sect to attract a large following is the Risshō Kōsei-kai. New Shintō cults include Tenri-kyō and Konkō-kyō.

Demographic trends. The increase in Japan's population since the 19th century has kept pace with economic development, and the standard of living has risen steadily. Japan's overall population density is exceeded by only a few countries, although population distribution is highly variable. Some 80 percent of the country consists of mountainous areas, which has caused the congested concentration of population within the limited plains and lowlands. The increased population has been absorbed into the ever-expanding urban areas, while the population of rural districts has declined considerably.

Population growth trends

Japan has experienced spectacular growth since 1868, when the population numbered about 33,000,000. This increase is directly related to slow but steady urban growth; the development of Hokkaido, Tōhoku, and southern Kyushu; and the introduction of commercial agriculture. In 1897, when industrialization first began, the population numbered more than 42,000,000. From 1898 to 1918 growing industrial cities and mining towns absorbed a large population, as did Hokkaido and the sericultural (silkworm-raising) rural districts.

In 1920, when the first precise census was conducted, the population was nearly 57,000,000. Between 1919 and 1945 Tokyo, Ōsaka, Nagoya, and northern Kyushu developed as the nation's four major industrial districts. At the same time, some of the smaller cities lost their ability to sustain a growing population, and some of them declined. By 1940 the population had grown to more than double that of 1868. During World War II, there was a marked migration to the rural areas to avoid aerial bombing, and some cities such as Ōsaka were reduced to one-third their previous size. After 1945, the repatriated population of nearly 9,000,000 and the temporarily explosive increase in the birth rate caused abnormally high growth.

The rapid rehabilitation of industry after 1950 has resulted in the continuing concentration of population in the Pacific coastal areas. The expansion of the Keihin area is not confined to Tokyo, Yokohama, and their adjacent suburbs but extends to a much wider circle. The same is true of the Keihanshin (Kyōto–Ōsaka–Kōbe) and Chūkyō (Nagoya) areas. Rural areas outside the direct influence of urbanization have been subjected to a marked decline. Adult males migrate to the Pacific coast, and many of those who remain at home periodically leave as temporary labourers, creating a constant outflow of population from the mountainous areas and isolated islands. In many places, emigration is so marked that the remaining population cannot maintain a balanced community, and whole settlements are abandoned. A striking demographic feature since World War II is the decline of birth and death rates in response to improved birth control measures and health conditions. Thus, Japan's rate of population increase is one of the world's lowest, and its life expectancy is among the highest. (Ak.W./Y.M./Ed.)

THE ECONOMY

Japan's regional and world standing

Japan is remarkable for its extraordinarily rapid rate of economic growth in the 20th century, especially after World War II. This growth has been based on unprecedented expansion of industrial production and on an aggressive export trade policy. As a result, Japan has become the second largest free-market economic power, ranking only behind the United States. It is one of the world's largest producers of motor vehicles, steel, and high-technology manufactured goods. Although Japan's standard of living did not increase as rapidly as did the overall economy in the early postwar decades—due in large part to the high percentage of capital reinvestment in those years—it gradually has caught up and become comparable with that found in other developed countries.

Administration of the economy. *Private enterprise, the role of the government, and taxation.* Japan's system of economic management is probably without parallel in the world. Although the extent of direct state participation in economic activities is limited, the government's control and influence over business is stronger and more pervasive than in most other free-enterprise countries. This control is not exercised through legislation or administrative action but through constant—and to an outsider almost obsessive—consultation with business and through the authorities' deep indirect involvement in banking. Consultation is mainly by means of joint committees and groups that keep under review, monitor performance of, and set targets for nearly every branch and sector of the economy. In addition, there are several agencies and government departments that concern themselves with such aspects of the economy as exports, imports, investment, and prices, as well as with overall economic growth. These are staffed by experts, who are not only in constant touch with business but are also close to the minister concerned; they form an integral part of a system that is quick to collate and interpret the latest economic indicators and to respond to changes in the situation. The most important of these agencies is the Economic Planning Agency. Like the Bureau of Statistics, it forms part of the Prime Minister's Office and, apart from monitoring the daily running of the economy, it is also responsible for long-term planning.

The system works well, without any major crises in government–business relations, because of the unusual self-discipline of Japanese businessmen in their dealings with the authorities and the government's deep understanding of the role, needs, and problems of business. The need for large-scale government participation in economic activities is thereby obviated, and, unlike many governments in the free-enterprise world, the state appears to be reluctant to extend its direct role. In 1985, for example, the government relinquished to the private sector its monopolies over the tobacco and salt industries and domestic telephone and telegraph services; and in 1987 the publicly owned Japanese National Railways was privatized, its operations divided into several constituent companies that operate together under the name Japan Railways (JR) Group. The government retains an interest in international telecommunications services and radio and television broadcasting. It plays no part in gas production or—except for providing electricity in economically underdeveloped areas—in electricity generation.

The government's role in banking

The government's economic influence is supplemented by its substantial role in banking. The state owns a number of financial institutions, such as the Japan Development Bank, the Export-Import Bank, the Small Business Finance Corporation, and the Housing Loan Corporation, whose principal objectives are to finance private enterprise in areas that are considered particularly desirable. The Ministry of Finance and the Bank of Japan have considerable influence over business investment decisions because of the close interdependence of business, the commercial banks, and the central bank.

Tax revenues account for a great majority of the government's total income. Japanese taxes can be divided into three main categories: taxes on individuals, taxes on business, and miscellaneous levies. Individuals are subject to progressive income, prefectural, and municipal inhabitant taxes. Income tax is divided into withholding tax, which is deducted at the source of income, and assessment tax, which is payable annually. The local taxes consist of a per capita levy and an income levy. An enterprise tax is sometimes levied on individuals who carry on specified business activities. Businesses are subject to local inhabitant taxes, which consist of a per capita and a corporation levy, enterprise tax, and corporation tax. Other taxes include gift and inheritance taxes, excise taxes levied on a number of consumer goods, and liquor, gasoline, travel, entertainment, securities-exchange, and automobile-acquisition taxes.

Trade unions and employers' associations. Japanese trade unions have had a relatively short history. Although there were several labour organizations before World War II, trade unions became important only after the U.S. occupation forces introduced legislation that gave workers the right to organize, to bargain with employers, and to strike. Because Japanese trade unions are generally organized on a plant or enterprise basis, their number is relatively large, and in many cases there are different organizations for different plants of the same company. The great majority of the enterprise unions are affiliated to federations that are loosely organized on craft lines, such as the Federation of Iron and Steel Workers Unions and Federation of Textile Workers Unions, and to one of the national labour organizations. They retain much of their independence, however, in dealing with employers. While the craft and national federations formulate general policy, discuss and advise on strategy, and coordinate wage offensives, serious negotiations are usually conducted by individual unions and the employees. One result of Japan's industrial, as opposed to craft, unionism is that demarcation disputes and interunion rivalry for members are relatively rare. Furthermore, if judged in terms of working days lost, Japanese labour relations have been noticeably more amicable than those in other developed countries, such as the United Kingdom, Italy, and the United States.

Labour organizations

The national labour organizations are the left-wing and highly political General Council of Trade Unions of Japan (Sōhyō), the more moderate and less political Japan Confederation of Labour (Dōmei), the National Federation of Industrial Labour Organizations (Shinsambetsu), and the Federation of Independent Unions (Chūritsu Rōren). Sōhyō is the largest of the four, and Dōmei is its principal rival; Chūritsu Rōren often associates itself with Sōhyō, especially during the annual "spring labour offensive."

Apart from the political demands of the left-wing unions, labour organizations have been mainly concerned with such questions as wages, prices, and working conditions. Problems of the evolution of a comprehensive industrial policy, of greater centralization, and of the union of the rival national organizations became important during the late 20th century mainly because of the trend toward growing concentration in industry and greater cooperation among the various employers' organizations. There is also a growing feeling that in this age of rapid technological progress and change, the ideology-ridden policies of Sōhyō are no longer adequate. Although all of the interested parties pay lip service to unification, the pressure for greater cooperation, pragmatism, and professionalism comes mainly from the independent unions in private industries. Moves to unify these private-industry unions have been undertaken in the hope that unification would create a large politically moderate union capable of restraining the influence of the highly ideological government and public sector unions led by Sōhyō.

Resources. *Minerals.* Although Japan's mineral deposits are fairly diverse, with a few exceptions the reserves are small and production is inadequate to meet more than a small part of domestic requirements. The quality of the minerals mined is often poor, and since deposits are widely scattered the extractive industry is characterized by a large number of small and relatively inefficient mines that do not lend themselves to the application of modern, large-scale mining methods. Coal, iron ore, zinc, lead, copper, chromite, and manganese are among the most important minerals, and although a large number of others are mined on a minor scale, there is an almost complete lack of nickel, cobalt, bauxite, nitrates, rock salt, potash, phosphates, and oil.

Coal is the country's most important mineral. Japanese coal is of relatively poor quality and is difficult to mine, and production costs are therefore high. Production is concentrated in Hokkaido and Kyushu. Despite the Kyushu field's proximity to the sea, it has lost much of its importance because of the gradual exhaustion of the richer seams and the poor quality of the coal mined. The seams of the Hokkaido field are thicker than those of Kyushu, and the use of mechanized, large-scale mining methods has resulted in a relatively high output per miner.

Oil deposits are meagre, domestic oil production accounting for a negligible fraction of Japan's oil consumption. The oil-bearing belt extends from northern Honshu on the Sea of Japan to the Ishikari-Yūfutsu lowlands in Hokkaido; virtually the whole of the country's output comes from Niigata and Akita prefectures. Natural gas is produced in Honshu; two major fields are in the south Kantō region and in Niigata Prefecture.

Although Japan ranks as one of the world's major steelmakers, domestic resources and production of iron ore are small; Japan imports almost all of its iron ore. Japanese ore is of poor quality and is obtained mostly from small mines in Hokkaido and northern Honshu.

Copper, once Japan's most important metallic ore, is produced in mines situated in Hokkaido and the prefectures of Akita and Iwate. Domestic production is far from enough to meet demand, and the majority of copper ore is imported. Lead and zinc are often found in conjunction with copper; zinc production is concentrated in Gifu Prefecture. Other metallic ores mined include silver, gold, tungsten, tin, antimony, molybdenum, and titanium. Japan also has large sulfur and limestone deposits.

Agricultural and forest resources. Because of the country's mountainous terrain, the supply of agricultural land is limited. The soil, largely infertile and immature, requires careful husbandry and fertilization. Timber resources are extensive, consisting of broad-leaved and coniferous forests, but a sizable proportion of the forestland is located in inaccessible mountain areas. Most of the forest area is privately owned, and much of it is distributed among a large number of relatively small holders. The rest is publicly owned; it is in these areas that large-scale reafforestation has taken place.

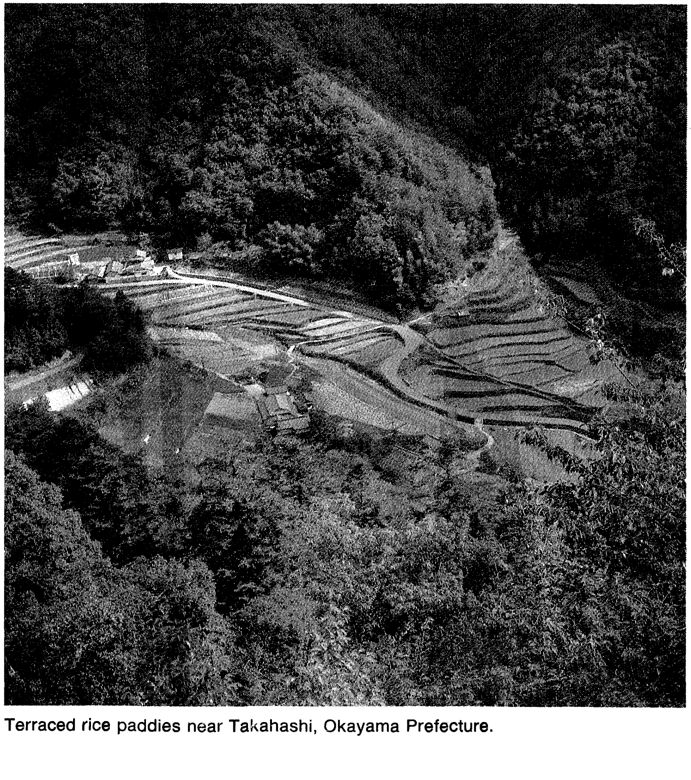
Timber resources

Hydroelectric resources. As a result of its climate, Japan has considerable water resources. It also has an extensive network of rivers that can be used for irrigation, although flooding is a serious problem in many parts of the country. As a result of the mountainous terrain, the ample hydroelectric potential is distributed in an uneven fashion. Hydroelectric development is largely concentrated in central Honshu along the Shinano, Tenryū, Tone, and Kiso rivers; in Tōhoku; and in some parts of Kyushu. This pattern of distribution ensures that Japan's hydroelectric capabilities are well located in relation to the important industrial areas. Although there is significant undeveloped potential, the best sites are already utilized, and further additions to capacity are increasingly expensive. Most hydroelectric power plants cannot operate at full capacity for more than a few months of the year because of seasonal variations of rainfall and the difficulty of constructing adequate storage facilities.

Agriculture and fishing. Agricultural production, including forestry and fishing, has grown less rapidly than national output or has actually declined, and it accounts for only a small proportion of the national income. Despite rapid increases in yields, agricultural output per person is considerably less than in other sectors of the economy. The agricultural sector employs a relatively large percentage of the working population in comparison to its contribution to the national economy, although increasing numbers of farm workers are leaving agriculture for other industries. Japanese agriculture is characterized by a large number of small and often inefficient farms, and many farmers have to rely on outside occupations for a substantial part of their income. Larger farms are generally found in Hokkaido, where units of 25 to 50 acres (10 to 20 hectares) are not uncommon. The country's principal crop is rice. Other important farm products include wheat, barley, potatoes, fruits, vegetables, and tea.

The main objectives of the government's agricultural policy have been to encourage self-sufficiency in the more important commodities, to enlarge the size of the average

© Suzanne J. Engelmann—Shostal Assoc.



Terraced rice paddies near Takahashi, Okayama Prefecture.

The government's agricultural policy

holding, which is small by advanced agricultural standards, and to close the gap between rural and urban incomes. The central feature of this policy has been an artificially high producer rice price. This has succeeded in raising farm incomes and has led to increases in the production of rice; a growing surplus of rice, however, has prompted the government to encourage farmers to raise livestock and to grow vegetables, wheat, and other crops instead of rice. Livestock-raising has become one of the most important farming activities, although most feeds must be imported.

Japan has one of the largest catches of fish of any nation in the world. In spite of its dominant international position, the Japanese fishing industry faces a number of serious problems, partly the result of structural weaknesses within the industry and partly a result of the restrictions placed upon it by nations that have claimed a 200-mile economic zone in their coastal waters. Despite the efforts of the government to create larger and more efficient units, small and medium-sized enterprises and individual fishermen still account for the bulk of the total income. Imports of fishery products exceed exports.

Industry. *Mining and quarrying.* Mining is a relatively unimportant branch of the economy. Most of the value of mining production is accounted for by coal, with copper, limestone, oil and natural gas, lead and zinc, sulfur, silver, and gold also significant contributors. The coal industry has suffered from the competition of cheaper foreign coal and from the rapid shift to oil after World War II. Despite the closure of a large number of uneconomic pits and a growth in productivity, foreign competition has had an unfavourable effect on the financial results of most coal-mining companies, which repeatedly has forced the government to aid the industry.

Manufacturing. The most notable feature of Japan's economic growth since World War II has been the rapid development of manufacturing. Progress has been made in terms of quantitative growth, quality, variety, and efficiency. Japan has become a greatly feared competitor whose products are in increasing demand. It is one of the world's principal shipbuilders and automakers and is a major producer of crude steel, synthetic rubber, aluminum, sulfuric acid, plastics, cement, pulp and paper, refined copper, and cotton yarn. It has some of the world's largest and most advanced industrial plants.

The most spectacular growth has been in the production of motor vehicles, iron and steel, machinery (including robots), petrochemicals, precision equipment (notably cameras), and advanced electronic products (including computers, telecommunications equipment, and consumer goods). Some of the older industries, however, have advanced relatively slowly. The lumber and wood industry, textiles, and foodstuffs have failed to match the expansion in manufacturing as a whole.

Reasons for industrial growth

A principal reason for Japan's industrial performance has been the high level and rapid growth of capital investment. A boom in equipment investment provided the iron-and-steel and machine-building industries with a rapidly growing home market, allowed for a spectacular increase in productive capacity and in the scale of operations, and led to a rapid replacement of old machinery. This, in turn, resulted in a considerable improvement in productivity throughout the economy and enabled industry to grow, despite the acute shortage of skilled labour that developed. Despite rising wages, many sectors of Japanese manufacturing have a formidable advantage over their rivals, a fact that is well illustrated by the country's growing exports. Also important to Japan's strong industrial position in the world are the high rate of labour productivity, relative to other major industrial countries, and the extensive use of technological innovations.

During the late 20th century, industry has also been characterized by a growing tendency toward tie-ups, mergers, and takeovers among the larger manufacturing and industrial concerns. These actions have been made possible by the gradual relaxation and the increasingly flexible interpretation of the country's antimonopoly laws enacted after World War II. The authorities have accepted the argument that greater concentration at the top is essential in order to improve efficiency, to make better use of the existing resources, and to increase or maintain international competitiveness. This argument has been given additional force by the need to strengthen Japanese enterprises in the face of growing direct foreign investment, made possible under the government's capital decontrol program. The merger of the Yawata and Fuji Steel companies into the Nippon Steel Corporation, for example, created one of the world's largest producers of steel.

There are also a large number of small and relatively inefficient manufacturing enterprises that use a substantial amount of the scarce labour. By and large, small firms tend to be more prominent in the less dynamic industries.

Despite Japan's rapidly increasing consumption of energy, per capita consumption remains low when compared to that of other industrialized countries. The largest single source of energy is oil; almost the entire demand is satisfied through imports, an important share of which comes from fields developed by Japanese companies. In addition, the difficulty and expense of mining coal has steadily reduced its importance. Because of the difficulties involved in hydroelectric development, the majority of the total electric power is generated by thermal plants. Oil is important, but there has been an increase in the importance of coal-fired electricity plants. There are also a number of nuclear and geothermal plants. The growth in gas production has been greatest for natural gas and liquefied natural gas, which account for the largest share of total production.

Finance. Japan's complex financial system is different from that of other developed countries in a number of important aspects. The Bank of Japan, established in 1882, is the sole bank of issue; it also plays an important role in determining and enforcing the government's economic and financial policies. The bulk of the domestic banking business is transacted through commercial banks, of which the city banks (such as Dai-Ichi Kangyo, Fuji, Mitsubishi, and Sanwa) are the most important. There are also a number of long-term credit banks, some government financial institutions—including the Japan Development Bank and the Export-Import Bank—and many mutual savings and loan banks and credit associations.

Relationship between banks and industry

One of the more unusual features of the Japanese financial system is the high degree of interdependence between the central bank, the commercial banks, and industry. Traditionally, industry has relied on banks for a large part of its borrowing requirements, and, although the importance of its own capital has increased, private and government financial institutions still account for a substantial part of the total. Since the commercial banks are responsible for most credit extended to industry, their influence on their client companies is considerable. Their active lending policy also means that their liquidity ratios tend to be low by Western standards and that they are forced to rely on call money (money that is readily available to banks as loans) and on large-scale borrowing from the Bank of Japan. The central bank is thereby in a strong position to influence bank operations and to bring about a quick adjustment in the volume of credit through credit ceilings, moral pressure, and other methods. Other sources of finance that are less susceptible to central bank influence include mutual savings and loan banks, credit associations, life insurance companies, and other nonbank financial institutions.

The bond market is relatively undeveloped because the government's low, long-term interest rate policy has made bonds relatively unattractive as compared with the comparatively high level of short-term rates. Individuals and institutional investors tend to buy discount debentures only. Bond buying, therefore, is confined chiefly to banks and other financial institutions, which are expected to purchase government and government-guaranteed bonds according to an unofficial allocation quota. The secondary bond market has been in operation since the mid-1960s, and, although over-the-counter transactions have risen rapidly, a significant proportion of the total business is accounted for by trading in financial debentures. It is generally accepted that improvement of the efficiency of the bond market is important, but significant progress cannot be made without rectifying the imbalance between short- and long-term interest rates.

There are several stock exchanges in Japan; the two most important, Tokyo and Ōsaka, account for almost all of the total business. Foreign investors have taken an interest in Japanese stocks, and the Tokyo exchange has become one of the world's largest.

Trade. An outstanding feature of Japan's economic development has been the rapid advance in its overseas sales, even though the share of exports in the country's gross national product has not changed significantly. From the point of view of individual industries and as a generator of growth, however, exports are much more important than their contribution to the national income would suggest.

Reasons for this spectacular export performance are the growing variety of Japan's industrial output, the shift to products with a relatively high value added, more advanced sales-promotion techniques, and the country's export competitiveness. The rise in labour productivity enabled industry to absorb much of the rapid rise in wage costs, with the result that for many years Japan's export prices tended to be lower than those of its principal competitors. Increasingly, however, Japanese exports have faced strong competition from such developing industrial nations as South Korea and Taiwan.

A significant change in the composition of exports occurred in the late 20th century. The share of total exports of textiles and food products decreased considerably, while exports of machinery and transport equipment grew dramatically, accounting for the largest component of exports. Other important exports include metal and metal manufactures and chemicals. The United States is Japan's largest single customer; East and Southeast Asia, western Europe, and the Middle East are other important customers.

Growth of imports

Imports have also grown steadily. Because of Japan's meagre natural resources, the bulk of its imports are raw materials, foodstuffs, and fuels. The major components of manufactured goods are machinery and allied products and chemicals. Japan's largest suppliers include East and Southeast Asia, the United States, the Middle East, western Europe, and Australia. (E.I.U./Y.M.)

Transportation. Until the latter part of the 19th century, the majority of Japanese people traveled on foot. There were no vehicular means of transportation except for small wagons, carts, or palanquins (*kago*) carried by men or animals. The first railway was built between Tokyo and Yokohama in 1872, and others soon followed, though the rugged terrain required the construction of many tunnels and bridges. Iron ships were built at about the same time, and modern ports were constructed. Road construction, however, tended to lag behind the development of other means of transport, resulting in the present congestion of most urban areas.

Japan's great cities attract large numbers of passengers and vast quantities of goods. Tokyo, especially, is an incomparably large focus for transportation; it is followed by the Ōsaka metropolitan area, including the three cities of Ōsaka, Kōbe, and Kyōto. The third largest focus of transportation is Nagoya. All of these large urban agglomerations are served by large and internationally known ports. Other cities, such as Kita-Kyūshū, Fukuoka, Sapporo, Sendai, and Hiroshima, also function as hubs of the transportation network.

Transport networks

The largest volume of intercity or interregional transport, in both passengers and goods, moves between the two largest metropolitan regions—by rail, road, coastal waterway, and air. Kyushu is connected with Honshu by the world's first undersea railway tunnel (built in 1941), by an undersea double-decked road tunnel (built in 1958), and by a huge suspension bridge linking the two (opened in 1973). With the opening in 1988 of a railway tunnel between Hokkaido and Honshu and of a multiple-span railway–road bridge between Honshu and Shikoku, all four of Japan's main islands were linked by surface transport.

The Japanese network of telecommunications and of postal services is among the best in the world. Its hundreds of islands, as well as its remotest villages deep in the mountains, are effectively linked by these services. During the 1980s Japan became a world leader in the use of advanced telecommunications, including fibre-optics networks and electronic-mail systems.

Road networks. The development of the road network is retarded in comparison with Japan's general economic progress and in view of its large number of cars. Some expressways have been built between major cities and to some scenic areas. In addition, the metropolitan regions of Tokyo and Ōsaka have a limited-access highway network within their respective built-up areas. The construction of a nationwide network of expressways, however, is expected to require many years. The fact that many roads, both in built-up and in rural areas, are narrow and winding causes additional planning problems. Until the 1920s most Japanese roads were used almost exclusively by pedestrians; even horse-drawn vehicles rarely used them. This fact, as well as the limited area of land in proportion to population, affected the pattern of road development considerably. Even in the late 20th century a great number of narrow roads were still being constructed, although subject to a minimum width. In rural areas footpaths are still in use, despite the existence of many newly built or widened motorable roads. Extremely dense networks of footpaths in many mountain areas are used for recreation.

Japanese city street patterns are manifold. Cities such as Kyōto and Nara still preserve the gridiron street pattern of the ancient Chinese city plan, though with modifications in built-up inner parts of the cities. In many rural areas as well, the ancient pattern of land division and the resultant road pattern take the rectangular gridiron form, which is similar to the U.S. township layout in concept and pattern, although it evolved much earlier and is smaller in size. Feudal towns, especially fortified (castle) towns, have somewhat similar street patterns, though with many modifications for defense purposes.

The number of four-wheeled motor vehicles and three-

Orion Press, Tokyo

Multiple-span bridge over the Inland Sea, linking Kojima, Honshu, with Sakaide, Shikoku.

Increases in vehicular traffic

wheeled trucks has increased at a phenomenal rate. Japan has an extremely high density of motor vehicles per unit area in the plains and in other inhabited areas. Trucks represent a much higher proportion of vehicular traffic than in other major motorized countries, and not until the late 20th century did passenger cars outnumber trucks. The quantities of freight transported by trucks have surpassed those carried by rail. Many families now have two or more automobiles, and commuting to work by car has become an increasingly common practice, resulting in road congestion in the big cities and in industrial areas. Although railways still play the major role in carrying commuters, there appears to be no practical solution to the problem of how to reduce the number of cars on the roads. The increases in poisonous exhaust gases and in the noise of the traffic have become serious problems. Steps taken to alleviate them include the development of pollution-control devices for automobiles and the installation of noise barriers on highways in densely populated areas. In addition, the desire to preserve natural or historical sites or to keep areas free from traffic noise often has proved an obstacle to road construction.

Railways. Despite the competition from road transport, railways play an extremely important role in transporting passengers. Railroads continue to give way, however, to competition from road and air transport.

The first Japanese rail line was financed by the British and built by British engineers. There was strong opposition to its construction, because many feared the expansion of foreign economic and political influence. This opposition was silenced somewhat after the line was completed. Other early railroad construction faced strong local opposition, but Japanese engineers began to build railroads at a rapid rate. The first streetcar line was constructed in Kyōto in 1891, using the electricity from the nation's first power station. In subsequent years Japan, unlike most other Asian countries, developed quite extensive intraurban and suburban railroad systems; the period between the two world wars, in particular, was characterized by the construction of many railroad lines to the suburbs to serve the needs of growing numbers of middle-income people. In 1927 the first subway was built in Tokyo's downtown district. Construction of new railroads continued until the outbreak of World War II, but the Japanese defeat and its aftermath prevented further construction for some time. From about 1955 onward, however, railroad construction was resumed. Subway construction, in particular, progressed, and systems have been built in Tokyo, Ōsaka, Nagoya, Sapporo, Yokohama, Fukuoka, Kyōto, and Kōbe.

The New Tōkaidō Line of the Shinkansen

In 1964 the New Tōkaidō Line of the Shinkansen—then a part of the government-owned Japanese National Railways and now one of the companies of the JR Group—began operations. Named for the Tōkaidō, the ancient highway between Kyōto and Tokyo, the line provides high-speed passenger service on an electrified, double-track route between Tokyo and Ōsaka. It is part of an eventual nationwide network of high-speed trains linking all major cities. Following completion of the New Tōkaidō Line, Shinkansen service was extended westward to Okayama in 1972 and then to Hakata (Fukuoka) on Kyushu in 1975. Two lines radiating outward from Tokyo—north to Niigata and northeast to Morioka—were opened in 1982.

The most serious traffic problem is caused by congestion on railroad transport within the large cities. Most commuter trains are very crowded during rush hours, with some trains carrying twice—and (rarely) three times—the number of passengers for which they were designed. Services have been expanded to cope with the growing demand. In Tokyo, a monorail (completed in 1964) operates over a distance of about eight miles between downtown Tokyo and the airport at Haneda. Monorails are numerous in Japan and are used primarily for commuting and recreational purposes. Many cable cars also operate in the mountains. The length of daily commuting time has become a serious problem; in some instances it consumes two hours each way.

Port facilities. Japan is one of the world's principal seagoing nations and has one of the world's largest merchant fleets. It has engaged in seafaring since early times. In about 1600 the port of Sakai, just south of present-day Ōsaka, prospered from its trade with China and Southeast Asia. Soon afterward the feudal regime greatly restricted foreign trade by imposing a policy of isolation from the rest of the world—a policy that was to last for about 250 years. As a consequence, few ports engaged in foreign trade, the notable exceptions being Nagasaki and Kagoshima in southern Kyushu and the Tsushima archipelago between Japan and Korea. After Japan reopened its doors to foreign trade in 1859, it was some time before large modern trading ports were developed; Yokohama and Kōbe became and have remained the leading trade ports of Japan, the former being the outport of Tokyo and the latter the outport for Ōsaka and Kyōto. Many other modern ports subsequently came into existence, including Nagoya, Kawasaki, Chiba, Kita-Kyūshū, Mizushima, and Sakai.

Air transport. Both domestic and international air transportation play important roles in Japan. Before World War II, air transportation was considerably restricted, but, since the foundation in 1953 of Japan Air Lines (JAL), international flights have proved profitable. Despite competition by railways, especially the Shinkansen, the volume of domestic air transport has grown considerably. Tokyo is the nation's largest single focus of air transport, followed by Ōsaka. Other major airports are in Nagoya, Sapporo, and Fukuoka. All other metropolitan areas in Japan are also connected by air routes. Generally speaking, southwestern Japan is covered by a denser network of air transport than other regions, primarily because of the presence of many islands.

GOVERNMENT AND SOCIAL CONDITIONS

Government. *Constitution and structure of government.* Japan's constitution was promulgated in 1946 and came into force in 1947, superseding the Meiji Constitution of 1889. It differs from the earlier document in the following points: first, the emperor, rather than being the embodiment of all sovereign authority (as he was previously), is the symbol of the state and of the unity of the people, while sovereign power rests with the people; second, Japan renounces war as a sovereign right; and third, fundamental human rights are guaranteed as eternal and inviolable. Furthermore, the government is now based on a constitution that aims at maintaining Japan as a peaceful and democratic country in perpetuity.

The role of the emperor

The emperor has no powers related to government. His major role as emperor consists in such formalities as appointing the prime minister—who is first designated by the Diet (Kokkai)—and appointing the chief justice of the Supreme Court (Saikō Saibansho), convoking sessions of the Diet, promulgating laws and treaties, and awarding state honours—all with the advice and approval of the Cabinet (Naikaku).

Legislative powers are vested in the Diet, which is popularly elected and consists of two houses. The House of Representatives, or lower house (Shūgiin), ultimately takes precedence over the House of Councillors, or upper house (Sangiin). Membership in the House of Representatives is based on proportional representation from prefectural districts, while that in the House of Councillors is divided between proportional representation and at-large representation. The House of Representatives controls the budget and approves treaties with foreign powers. Executive power is vested in the Cabinet, which is organized by the prime minister. If the House of Representatives passes a resolution of no-confidence or refuses to pass a vote of confidence in the government, the Cabinet must resign, unless the House of Representatives is dissolved within 10 days. There are governmental ministries and agencies in addition to the Prime Minister's Office. All offices of the central government are located in and around the Kasumigaseki district in central Tokyo. An independent constitutional body called the Board of Audit is responsible for the annual auditing of the accounts of the state.

Local government. Japan is divided into 47 prefectures, 43 of which are *ken* (prefectures proper), one of which (Tokyo) is a *to* (metropolitan prefecture), one (Hokkaido) is a *dō* (district), and two (Ōsaka and Kyōto) are *fu* (urban prefectures). Prefectures, which are administered by governors and assemblies, vary considerably both in area and in population. The largest prefecture is Hokkaido, with an area of 32,246 square miles, while the smallest is Ōsaka, with 720 square miles. The most populous prefecture is Tokyo, and the least populous is Tottori. A prefecture is further subdivided into minor civil divisions; these include the city (*shi*), town (*machi* or *chō*), and village (*mura* or *son*). All these local government units have their own mayors, or chiefs, and assemblies. Before World War II, there were also counties (*gun*), consisting of towns and villages but excluding cities within a prefecture. This county system only survives in the form of statistical units.

An intermediate level of governmental services is formed between the central and prefecture levels; the branch offices of the government are located in certain cities, which—as regional centres—generally administer several prefectures together. Designated cities (*shitei toshi*), which must have populations of at least 500,000 each, are divided into wards (*ku*). These cities include Yokohama, Ōsaka, Nagoya, Kyōto, Kōbe, Kita-Kyūshū, Sapporo, Kawasaki, Fukuoka, and Hiroshima. A ward has a chief and an assembly, the former being nominated by the mayor and the latter elected by the residents. Tokyo has 23 special wards (*tokubetsu ku*), the chiefs of which are elected by the residents. These special wards, created after the metropolitan prefecture was established in 1943, demarcate the city of Tokyo from the other cities and towns that make up the metropolitan prefecture; the city proper, however, no longer exists as an administrative unit.

Terms of office; constituencies

The political process. Members of both the House of Representatives and the House of Councillors are chosen by general elections. Members of the House of Representatives serve for a four-year term, but this may be terminated earlier if the house is dissolved. Members of the House of Councillors are elected for a six-year term, with half of the members being elected every three years. The electoral procedure for the House of Councillors differs from that for the House of Representatives in that about two-fifths of the total are elected from a national constituency, in which each voter casts a vote for a national candidate; the remaining members are elected from the prefectural constituencies. The number of seats for each constituency was determined largely by the population density in each area in 1947, with some modifications resulting from the population increase in urban constituencies. Heads of local governmental units, such as prefectures, cities, special wards, towns, and villages, are elected by local residents. Universal adult suffrage is available to all men and women who are 20 years or older.

Freedom to organize political parties is guaranteed by the constitution; any organization that supports a candidate for political office is required by law to be registered as a political party. Since the enactment of the 1947 constitution, many political parties have been organized, have merged, or have been dissolved. There are now more than 10,000 parties, most of them of local or regional significance. The Liberal-Democratic Party (LDP; Jiyū-Minshutō) represents somewhat conservative elements. Such conservative parties have been the dominant force in government since the mid-20th century.

The Japan Socialist Party (JSP; Nihon Shakaitō), long a major opposition party, draws much of its support from labour unions and inhabitants of the large cities. It has a neutralist policy, urging the establishment of a peaceful security system covering Japan and East Asia by means of a treaty between Japan, the United States, the Soviet Union, and China.

The Clean Government Party (Kōmeitō), also a major opposition party, draws its main support from the Sōka Gakkai. Although the party was formed in 1964 under the influence of the Sōka Gakkai, it has renounced any formal ties with it. The Clean Government Party's major policy aim is the establishment of a welfare system and the promotion of "human Socialism." The Democratic Socialist Party (DSP; Minshatō) was formed in 1960 by a right-wing splinter group of the JSP. It advocates an independent foreign policy and aims at creating a Socialist society through democratic processes, and it draws most of its support from the same sources as the JSP itself. The Japan Communist Party (JCP; Nihon Kyōsantō) is a small but important party that exerts a strong influence on political and intellectual developments in society. The New Liberal Club (NLC; Shin Jiyū Kurabu), founded in 1976 by several of the more progressive members of the LDP, plays a small but important role in Diet politics.

The role of the citizen in politics is often discussed both in journalism and in daily conversation, especially at election time. Many citizens wish to participate in solving such problems as traffic congestion, waste disposal, air and water pollution, the shortage of parks and playgrounds, and noise control. In many cities there are numbers of commercial and residential streets that have been closed to automobile traffic.

Justice. The judiciary is completely independent of the executive and legislative branches of the government. The judiciary system consists of the Supreme Court, eight high courts, a district court in each prefecture, with the exception of Hokkaido, which has four, and many summary (informal) courts. Family courts also abound.

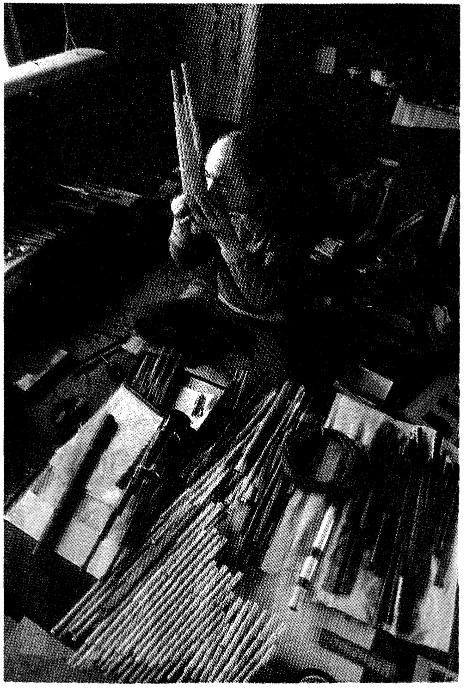
The Supreme Court consists of one chief justice and 14 other justices. The chief justice is appointed by the emperor upon designation by the Cabinet, while the other justices are appointed by the Cabinet. The appointment of the justices of the Supreme Court is subject to review in a national referendum, first at the time of the general election following their appointment and then at the general election every 10 years thereafter. An impeachment system also exists; the court of impeachment consists of members of the House of Representatives and of the House of Councillors. The Supreme Court determines questions of the constitutionality of any law, order, regulation, or official act. Lower court judges are appointed by the Cabinet from a list of persons nominated by the Supreme Court. The appointment term is for 10 years, and reappointment is allowed. All judges of lower courts must retire at the age of 70, according to law.

The Self-Defense Forces

Armed forces. Under its present constitution, Japan cannot maintain armed forces for purposes of aggression; in consequence, national security is maintained by the Self-Defense Forces (Jieitai), as well as by the collective security system in which the United States participates. The Self-Defense Forces play a role that is entirely limited to defense and internal security.

Between 1945 and 1950, Japan had no armed forces

Jim Brandenburg—TSW-CLICK/Chicago



Craftsman testing a *shō*, a type of mouth organ made from wood and bamboo that is used in *bugaku* (court music).

except for police; after the outbreak of the Korean War, however, the government, at the suggestion of the Allied occupation forces, established a National Police Reserve, which later became the Self-Defense Forces, and increased the strength of the Maritime Safety Agency. The Self-Defense Forces consist of ground, maritime, and air forces, under the civilian-controlled National Defense Council.

The constitutionality of the Self-Defense Forces has often been disputed. The Supreme Court ruled in 1959 that it did not violate the constitution, because its purpose was defensive. United States military bases operate in many parts of Japan under the Treaty of Mutual Cooperation and Security, concluded between Japan and the United States in 1960 and reaffirmed in 1970. The treaty may be terminated one year after either signatory indicates such an intention.

Japan's police services are under the supervision and control of the National Police Agency. Police services operate relatively smoothly; many problems that plague other countries are absent because of Japan's insularity and its nearly uniform ethnic composition. The relatively low ratio of extremely violent crimes to total crimes stands in contrast to that of most advanced countries.

Education. A great many educational institutions existed even in the feudal period preceding the Meiji Restoration of 1868, a number of which had been subjected to Chinese cultural influences since ancient times. Numerous private temple schools (*terakoya*), mostly in towns, functioned as elementary schools; reading, writing, and arithmetic were taught by monks, unemployed warriors, or others. Provincial lords (daimyo) also established special schools for children of the warrior class. Yet another type of school instructed primarily the children of wealthier merchants and farmers.

The modern educational system was introduced immediately after the Meiji Restoration. The government set up elementary and secondary schools throughout Japan in 1872, and in 1886 a system providing three to four years of education was inaugurated. The introduction of modern education did not encounter many problems, primarily because it was possible to utilize the educational system already functioning. Free compulsory education was introduced in 1900, and in 1908 it was extended to a period of six years. Since 1947, compulsory education has been for a nine-year period, beginning at the age of six.

The educational system of Japan is organized as follows: kindergarten lasts from one to three years but is not compulsory. Compulsory elementary school lasts six years, compulsory middle school three years, and high school (not compulsory) another three years. Higher education institutions consist of junior colleges, lasting for two to three years, and ordinary colleges and universities, lasting for four years. A master's degree can be obtained in two years after earning a bachelor's degree, and a doctor's degree in three years after earning a master's degree. In addition, there are five-year technological colleges that combine high school and junior college education. Public elementary and middle schools are free.

Higher education

Japan is one of the few countries in the world that provides a complete and thorough education for almost all the people. Although neither kindergartens nor high schools are compulsory, attendance at both has become virtually universal; higher education has also become popular. There are more than 1,000 institutions of higher education in the country. A large number of preparatory schools (*juku*) have been established to help students prepare for the difficult and highly competitive university entrance examinations.

The Tokyo metropolitan area, including Yokohama and many other satellite cities, has a high concentration of students, which has led to congestion as well as to an active intellectual life. The Tsukuba Science City, located about 40 miles northeast of Tokyo, consists of government research facilities and two universities.

Educational administration is decentralized, with the Ministry of Education playing a coordinating role. Responsibilities for the budget, curriculum, teacher appointments, and the supervision of elementary and middle schools are in the hands of local educational boards.

Health and welfare. Higher living standards, including better nutrition and better living conditions, as well as progress in medical care, have contributed much to an increase in the life span. Numerous hospitals, clinics, and health centres throughout the nation, as well as health education in schools and among the public at large, have virtually eliminated such diseases as typhus, diphtheria, and scarlet fever. Tuberculosis and dysentery are much less prevalent than they once were. Cholera, leprosy, and rabies have long been practically nonexistent. Increases in the so-called diseases of civilization have, however, become a serious problem. Stroke, high blood pressure, heart ailments, mental disorders, and similar diseases have become principal causes of death, as have traffic accidents. Cancer has also become a major cause of death. Japanese medical practice is usually of the Western type, but classical Chinese techniques are also used.

Changing dietary conditions

The Japanese people are obtaining increasingly better food. Although calorie consumption is generally lower than that of Europeans or Americans, overnutrition causing excess weight has nevertheless become a serious problem. The traditional Japanese food has been replaced partly by Western types of food and partly by Chinese food—to such a degree that the average Japanese no longer regards Western or Chinese food as alien.

During the feudal period, there was a social division of commoners into four classes (warrior, farmer, craftsman, and merchant), with a peer class above and an outcast class below. With the exception of the *burakumin,* the former outcast class, this social-class system has almost disappeared. In 1959, for example, Crown Prince Akihito married a commoner. Insofar as a social-class system does persist it does not have the ethnic basis that can exist in multiracial societies since the Japanese regard themselves as belonging to the same ethnic group. The few exceptions include resident aliens (non-Japanese citizens)—particularly Koreans who came to Japan as labourers before and during World War II and their descendants—and Japanese citizens of Ainu origin, who are scattered over the island of Hokkaido. Before World War II there was a tendency to distinguish the people of Okinawa from other Japanese because many of them exhibited minor differences in physiognomy and cultural life; this tendency has diminished but not disappeared. Okinawan culture, including its dialect and religion, is now recognized as sharing many traits with Japanese culture and, in some respects, represents a prototype. The ethnic unity of the Japanese is reflected in the fact that, of all the peoples of the world, the Japanese have been among the least inclined to intermarry with foreigners. A vernacular word for foreigner, *gaijin* (literally, "outside person"), is often used in daily conversation; the term implies that *gaijin* are fundamentally different from Japanese and, as such, cannot understand Japanese culture.

Vast discrepancies between the conditions of the rich and poor have been reduced since World War II largely as a result of the agricultural reforms of 1946 to 1950 and of the application of a graduated income tax. Although differences exist in income and property, the great majority of the Japanese regard themselves as in the middle-income group. Most of those in the upper middle-income group own their own homes, usually houses with several rooms surrounded by a garden; those in the lower middle-income group usually live in a two- to five-room house or (in urban areas) in an apartment house. The social attitudes of the Japanese and the absence of strict zoning in urban areas have contributed to the mixed land uses characteristic of Japan's cities; thus, functionally different establishments, such as shops, factories, or houses, are found adjoining one another, so that mixed rather than exclusive social patterns result.

Social-welfare services have expanded considerably since the end of World War II. Social welfare programs include social insurance, services for the aged and the physically and mentally handicapped, and care for disadvantaged children. Social insurance itself consists of health insurance, pension insurance, unemployment insurance, and workmen's accident compensation insurance. After 1961 the health-insurance system covered all Japanese people.

The scale of payments varies and in some cases no payments are required. The government has established hundreds of health centres throughout Japan, aiming primarily at improving environmental sanitation and at preventing communicable diseases in their early stages. Most of the hospitals are operated by unions, associations, or individuals, and the remainder by local governments and the national government. Aged people receive many services, including medical examinations, home-help services, recreational services, and institutional care, as well as varying amounts of financial aid. Local governments are obliged to provide welfare services for the physically handicapped and mentally retarded. Various children's welfare programs also exist—for example, medical-care services are free to expectant mothers and to young children from low-income families. Many voluntary and private associations also provide supplementary services.

Wartime destruction of houses

Housing. The housing shortage has posed major political and social problems. It is due to the following: the destruction of 70 percent of the houses in 70 percent of the major cities in Japan during World War II; constantly rising land prices, especially in and around major cities; the general use of lumber as a building material, requiring earlier replacement than brick or stone; the frequent occurrence of earthquakes, typhoons, and heavy rains bringing floods; the government's inclination to encourage economic growth rather than house construction; a rapidly rising standard of living, creating a demand for larger and better houses; and an increase in the number of families, resulting from the breaking up of extended families into smaller units.

The penchant for living in single-family homes may be attributed to the Japanese love of nature, to the influence of the garden-city movement, and to the fashion for imitating the tree-shaded residences of the feudal warrior class. The government and some private construction companies have encouraged construction of multistory apartment houses.

To cope with the housing shortage, a semigovernmental agency, the Housing Loan Corporation, was established in 1950 to finance house construction at low interest rates. In 1955, another semigovernmental agency was organized: the Japan Housing Corporation (since 1981 called the Housing and Urban Development Corporation), which has contributed significantly to housing construction. In addition, local governments have built a number of units, mostly of the apartment-house type and primarily for low-income families, and many large corporations maintain low-cost housing for their employees. A major proportion of the houses built are subsidized by governmental or semigovernmental funds. Generally, the size of housing units is increasing as per capita national income rises.

CULTURAL LIFE

The cultural milieu. Japan's long history has produced a cultural milieu that differs significantly from that of other countries. In general, this milieu is characterized by an inseparable mixture of traditional Japanese culture with introduced Chinese and Western cultural forms.

Ancient Chinese influences

Prehistoric Japanese culture was subjected to ancient Chinese cultural influences that were introduced some 1,500 years ago. One consequence was the imposition of the gridiron system of land division, which long endured; it is still possible to trace the ancient place-names and field division lines of this system. Chinese writing and many other Chinese developments were also introduced. The Buddhist religion, which originated in India and underwent modification in Central Asia, China, and Korea, also exerted a profound influence on the Japanese cultural life, but in the course of history, the process of Japanization continued. The Japanization of the introduced Chinese culture was greatly accelerated during the 250-year period of isolation that ended in 1868. The climate of Japan, for example, which was much more humid than that of China, led to such cultural adaptations as the use of lumber for building in place of the mud and brick used in China. Similarly, the Chinese characters had only a limited use because they did not fit the Japanese language.

After the Meiji Restoration of 1868, Japan began to modernize and to industrialize on the European and U.S. pattern. In the period since then, the United States generally has exerted a more conspicuous influence on Japanese cultural and social life than has Europe. Western cultural traits have been introduced on a large scale through the schools and the mass-communication media. Western scientific and technical terms have been widely diffused in translation and have even been re-exported to China and Korea. U.S., English, French, German, or Soviet influences on Japanese culture are in evidence in literature, the visual arts, music, education, science, recreation, and ideology.

Effects of Western influences

Modernization has often been accompanied by cultural changes. Rationalism and Socialism based on Christianity, as well as Marxism, have become inseparably related to everyday Japanese life. Western or Westernized music seems to be preferred to traditional Japanese music at most social levels. Although Japanese Christians form a fractional percentage of the population, Christmas is enjoyed, if not celebrated, quite widely, almost as a folk event. The use of Western dress among the Japanese, in place of the kimono, is widespread, although the kimono tends to be used by women at celebrations and by a considerable number of male adults and older women for home wear. House construction has also been changed considerably by the introduction of Western architectural forms and functions. In shape, in colour, and in building materials, many contemporary Japanese houses are significantly different from the traditional ones; they now have more modernistic shapes, use more colours, and are more often made of concrete and stucco.

Everyday use of modern transport and of modern communication media has brought Japanese urban life close to that of the West. Japanese forms of recreation are similar to those in other developed countries, although there are some notable differences. Outdoor recreational activities include hiking, mountaineering, skiing, skating, golf, swimming, boating, fishing, baseball, tennis, and football (soccer). Indoor recreations include Shōgi (a kind of Chess), Go (a strategy game also similar to Chess), Mah-Jongg, Japanese and Western card games, basketball, volleyball, table tennis, bowling, wrestling, gymnastics, and such martial arts as judo, kendo, karate, and aikido. Sumo wrestling is also practiced, or watched, both indoors and out.

Arts, folk traditions, and popular culture. The Japanese cultural tradition includes many forms of the fine arts and folk arts. Local variations are found throughout Japan's mountainous archipelago, where most river basins, valleys, or islands have their own specific folklores.

The highly refined traditional arts of Japan include flower arranging (*ikebana*), the tea ceremony (*cha-no-yu*), painting, calligraphy, dance, music, theatrical plays (including such forms of drama as Kabuki, a highly stylized form of drama characterized by singing and dancing; *bunraku,* the puppet theatre; Nō, the classic form of dance-drama), and *gagaku* (court music), gardening, and architecture. Delicacy and exquisiteness of form, together with simplicity, characterize traditional Japanese artistic taste. The Japanese tend to view the traditional Chinese arts generally as being too grandiose or showy. The newly introduced Western arts are also felt to suffer from the same flaws, though in a different fashion.

With the advance of modernization, many folk traditions and forms of folklore are rapidly disappearing. The widespread use of standard Japanese has accelerated this trend, since local cultures are directly related to dialects. Folk songs, for example, are generally no longer commonly sung except in some remote areas in northern and southwestern Japan. Folk music and dance are related to local life and are often significantly concerned with the local religion (whether animistic, Shintō, or Buddhist), agriculture, or human relations (including the theme of love). Some, however, still enjoy a great popularity, which has been increased through the mass media. On informal social occasions, even in the large cities, folk and popular songs are often sung. Such traditional arts as *ikebana, cha-no-yu,* and calligraphy are studied and practiced by a great many Japanese; *ikebana* and *cha-no-yu,* in particular, are popular among young unmarried women, since these are regarded as cultural or aesthetic attributes for future housewives. Traditional Japanese painting, dance, and music have, however, lost much of their traditional popularity, though the poetic forms of haiku and waka continue to flourish.

Changes in social customs

In social life, the arranged marriage (*miai-kekkon*) is being replaced by the love match, though a significant proportion of marriages are still arranged or initiated by parents or other older persons or, sometimes, by friends. Modern (usually Western) popular culture has gained a strong foothold in Japan. Jazz, rock, and the blues are enjoyed by the younger generation, along with half-Westernized or half-Japanized folk and popular songs. Many more or less Japanese songs are sung to the accompaniment of Western musical instruments; at the same time, many more or less Western subjects are treated in Japanese-style drama or song.

Japan has 12 national holidays. New Year's Day is traditionally regarded as the most important of these holidays, with millions of people engaging in a kind of pilgrimage that leads to shrines and temples starting at midnight of December 31. For three days thereafter, people visit shrines and temples, their families, and the homes of friends. In addition to the 12 national holidays, there are also such nationwide festivities as the Doll Festival, or Girls' Day (March 3), which is comparable to Boys' Day (May 5), now celebrated as Children's Day (a national holiday). May Day (May 1) is celebrated by many workers. Many temples and shrines celebrate their own specific festivals, attracting large numbers of people. City, town, and village authorities, as well as local communal bodies, often organize local festivals.

Cultural institutions. In addition to its cultural institutions such as libraries, museums, art galleries, theatres, parks, gardens, and schools of various kinds, Japanese department stores also play a role in the dissemination of culture by offering free or low-cost exhibitions.

The National Diet Library in Tokyo (which also includes branch libraries) is the single largest library in Japan. Higher educational institutions, including universities, colleges, junior colleges, and technical colleges, have hundreds of libraries. Secondary and elementary schools are also equipped with libraries as a matter of course. In addition, there are city, town, and village libraries, some equipped with mobile facilities. The overwhelming majority of library books are in the Japanese language.

Museums

There are museums of all kinds; these include general, science, historical, art, and outdoor museums, as well as zoos, botanical gardens, and aquariums. Museums of all kinds have been increasing in number, as well as in the quantity of their exhibits and in their attendance.

Local governments provide youth educational services, offering classes on various topics. Adult education is also conducted by local governments, as well as by private institutions, offering classes in general education, vocational training, technology, homemaking, home economics, arts, physical education, and recreation. Many institutions also help to promote nature studies and recreation through public and private youth hostels, national lodging houses, national vacation villages, national parks, quasi-national parks, and a great many prefectural natural parks.

Special and miscellaneous schools also function as agencies of cultural dissemination. These schools, which are recognized by the local authorities, offer courses in such subjects as dressmaking, handicrafts, cooking, abacus calculation, foreign languages, driving, and nursing.

Press and broadcasting. The print and broadcast media have long been influential in Japan, although their activities were somewhat circumscribed by the government until the end of World War II.

The press. Japan ranks as one of the major book publishing countries in the world, and Tokyo is the centre of the Japanese publishing industry. Several thousand magazines are also published, with more than half of these being weeklies.

The role of newspapers is of great importance. Major newspapers print both morning and evening daily editions, and daily circulation is relatively high. Several newspapers have nationwide circulation, and some local papers also have large circulations. Japan's largest dailies rank among the highest in the world in circulation, and all of the large papers are generally considered to maintain high editorial standards.

Radio and television. Radio and television are used in Japan far more extensively than in any other Asian country and, indeed, than in most other countries in the world. Radio broadcasting began in 1925 with the establishment of Nippon (Nihon) Hōsō Kyōkai, or NHK (the Japan Broadcasting Corporation)—a public corporation financed by license fees that according to law must be paid by television-set owners. NHK broadcasts many quality programs, both on radio and on television; no commercial advertisements are permitted.

The first television broadcast was made by NHK in 1952. Television stations now broadcast to all parts of Japan, including all of the isolated islands. NHK has been broadcasting overseas programs such as "Radio Japan" since 1953 and now broadcasts in more than 20 languages. Private commercial broadcasting began in 1951 and has gained widespread popularity. In addition, satellite and cable television reception has become common.

Commercial advertising has become an immense industry in Japan. Television and newspapers are the most important media of advertising; magazine and radio advertising are less significant.

For statistical data on the land and people of Japan, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (Y.M.)

History

ANCIENT AND MEDIEVAL JAPAN TO C. 1550

Prehistoric Japan. *Pre-Ceramic culture.* It is not known when man first settled on the Japanese archipelago. It was long believed that there was no Paleolithic occupation in Japan, but since World War II a number of Paleolithic tools have been uncovered. These include both core tools, made by chipping away the surface of a stone, and flake tools, made by working with a stone flake broken off from a larger piece of stone. It seems likely that the people who used these implements moved to Japan from the Asian continent. At one stage, land connections via what are now the straits of Korea and Tsushima made possible immigration from the Korean peninsula, while another connection, via what are now the Sōya and Tsugaru straits, allowed people to come in from northern Asia.

The Paleolithic Age in Japan is variously dated from 30,000 to 10,000 years ago. Nothing certain is known of the culture of the period, though it seems likely that people lived by hunting and gathering, used fire, and made their homes either in pit-type dwellings or in caves. No bone or horn artifacts of the kind associated with this period in other areas of the world have yet been found in Japan. There was no knowledge whatsoever of pottery; hence, the period is referred to in Japan as the Pre-Ceramic or Pre-Pottery era.

The Japanese Neolithic cultures

Jōmon culture (5th or 4th millennium to c. 250 BC). The Pre-Ceramic period is followed by two Neolithic cultures, the Jōmon and the Yayoi. The former takes its name from the *jōmon* ("cord marks") pottery found throughout the archipelago. A convincing theory dates the period during which Jōmon pottery was used from about 10,000 years ago until the 2nd or 3rd century BC. Of the features common to Neolithic cultures all over the world—progress from chipped tools to polished tools, the manufacture of pottery, the beginnings of agriculture and pasturage, the development of weaving, and the erection of monuments using massive stones—the first two are prominent features of the Jōmon period, but the remaining three do not appear until the succeeding Yayoi period. The manufacture of pottery, however, was highly developed, and the work of Jōmon culture has a diversity and complexity of form and an exuberance of artistic decoration. It is customary to take changes in the type of pottery used as a basis for subdividing the age and to distinguish very early, early, middle, late, and very late periods. It must be remembered, however, that since Jōmon culture spread over the whole of the archipelago, it also developed regional differences, and this combination of both chronological and regional variations gives the evolution of Jōmon pottery a high degree of complexity.

The pottery of the very early period includes many deep, urnlike vessels with tapered bases. In the early period, the vessels of eastern Japan become roughly cylindrical in shape, with flat bases, and the walls contain an admixture of vegetable fibre. In the middle period there were rapid strides in pottery techniques; the pots produced during this time in the central mountain areas are generally considered to be the finest of the whole Jōmon era. The surface of these generally cylindrical vessels is covered with complex patterns of raised lines, and powerfully decorative projections rise from the rim to form handles. From the middle period onward there is increasing variety in the types of vessels, and a clear distinction developed between high-quality ware using elaborate techniques and simpler pots made for purely practical use. The amount of the latter increases steadily, preparing the way for the transition to Yayoi pottery.

Jōmon dwelling sites have been found in various parts of the country. They can be classified into two types: one, the pit-type dwelling, consisted of a shallow pit with a floor of trodden earth and a roof; the other was made by laying a circular or oval floor of clay or stones on the surface of the ground and covering it with a roof. Remains of such dwellings have been found in groups ranging from five or six to several dozen, apparently representing the size of human settlements at the time. Most of these settlements form a horseshoe shape, with a space in the centre that seems to have been used for communal purposes. Nothing certain is known, however, concerning social or political organization at this period. It can be deduced that each household was made up of several family members and that the settlement made up of such households was led by a headman or magician.

The people of the Jōmon period lived mainly by hunting and fishing and by gathering edible nuts and roots. The appearance of large settlements from the middle period onward has been interpreted by some scholars as implying the cultivation of certain types of crop—a hypothesis supported by the fact that the chipped stone axes of this period are not sharp but seem to have been used for digging soil. Weaving was still unknown, and clothes were probably made of skins or bark. Jewelry included bracelets made of seashells, earrings of stone or clay, and necklaces and hair ornaments of stone or bone and horn. From the latter part of the period, the custom also spread over the whole country of extracting or pointing certain teeth, probably performed as a rite marking the attainment of adulthood.

No especially elaborate rites of burial were evolved, and the dead were buried in a small pit dug near the dwelling. Sometimes the body was buried with its knees drawn up or with a stone clasped to its chest, a procedure that probably had some religious or magical significance. A large number of clay figurines have been found, many representing female forms that were probably magical objects associated with primitive fertility cults.

The Ainu

For years certain scholars have claimed that the people responsible for the Jōmon culture were not of Japanese stock but were ancestors of the Ainu, an aboriginal Caucasian people now found in northern Japan. Scientific investigation of the bones of Jōmon people carried out since the beginning of the 20th century has disproved this theory. The Jōmon people were a particular people who might be called proto-Japanese, and they were spread all over the country. Despite certain variations in character arising from differences in period or place, they seem to have constituted a single stock with more or less consistent characteristics. The present Japanese people were produced by an admixture of certain strains from the Asian continent and from the South Pacific, together with adaptations made in accordance with environmental changes. Linguistic evidence suggests that a people speaking a language belonging to the primitive Ural-Altaic family moved eastward across Siberia and entered Japan via Sakhalin and Hokkaido. Nothing can yet be proved concerning their relationship with the people of the Pre-Ceramic period, but it cannot be asserted that they were entirely unrelated.

Yayoi culture (c. 250 BC to c. AD 250). The new Yayoi culture that arose in Kyushu (the southernmost of the four main Japanese islands), while the Jōmon culture was still undergoing development elsewhere, spread gradually eastward, overwhelming the Jōmon culture as it went, until it reached the northern districts of Honshu, the largest island of Japan. The name Yayoi derives from the name of the district in Tokyo where, in 1884, the unearthing of pottery of this type first drew the attention of scholars. Yayoi pottery was fired at higher temperatures than Jōmon pottery and was turned on wheels. It is distinguished partly by this marked advance in technique and partly by an absence of the proliferating decoration that characterized Jōmon pottery. It developed, in short, as pottery for practical use. It is accompanied by metal objects and is associated with the wet cultivation of rice. Culturally, it represents a notable advance over the Jōmon period and is believed to have lasted for some five or six centuries, from the 3rd or 2nd century BC to the 2nd or 3rd century AD.

In China, the 3rd and 2nd centuries BC corresponded with the period of the unified empire under the Ch'in (221–206 BC) and Han (206 BC–AD 220) dynasties, which had already entered the Iron Age. In 108 BC, the emperor Wu Ti occupied the Korean peninsula and established Lo-lang and three other colonies. They provided a base for a strong influx of Chinese culture into Korea, which, in turn, spread to Japan. The fact that Yayoi culture had iron implements from the outset, and bronze implements somewhat later, probably indicates borrowings from Han culture. Iron objects rust easily, and comparatively few have been found, but they seem to have been widespread at the time. These include axes, knives, sickles and hoes, arrowheads, and swords. The bronze objects are also varied, including halberds, swords, spears, *taku* (small bell-shaped devotional objects from China), and mirrors. The halberds, swords, and spears seem not to have been used in Japan for the practical purposes for which they were evolved in China but to have been prized as precious objects.

Wet cultivation of rice

The wet cultivation of rice, possibly borrowed from southern China, was one of the most important features of Yayoi culture. The earliest Yayoi pottery and sites, discovered in northern Kyushu, have yielded marks of rice husks as well as carbonized grains of rice; this suggests that rice growing was carried on in Japan from the earliest days of the culture. Traces of paddy fields, their divisions marked with wooden piles, have been found close to sites of settlements in various districts, along with irrigation channels equipped with dams and underdrains, showing that techniques of making and maintaining paddy fields were quite advanced.

Generally speaking, the settlements of this period were built on low-lying alluvial land to facilitate the irrigation of the paddies, but at one stage they were built in the hills or on high ground instead. It is not clear whether this was dictated by the needs of defense or whether dry cultivation was being practiced. Much as in the Jōmon period, there were two types of dwelling, the pit type and the type built on the surface; but in addition to these, raised-floor structures appeared and were used for storing grain.

With the acquisition of a knowledge of textiles, clothing made great strides compared with the Jōmon period. The cloth was woven on primitive looms using vegetable fibres.

The dead were buried in either large clay urns or heavy stone coffins. Both were common in northern Kyushu and neighbouring areas, and similar urns and coffins are also found in Korea, where they probably originated. The graves were usually marked by mounds of earth or circles of stones, but a special type employed a dolmen (a large slab of stone supported over the grave by a number of smaller stones). Since the erection of dolmens was widely practiced in Manchuria and Korea, these, too, are believed to be a sign of an influx of continental culture. Normally, graves occur in clusters, but occasionally one is found apart, surrounded by a ditch and with swords, beads, and mirrors buried along with the dead. Such special graves suggest that society was already divided into classes.

It is natural to suppose that these new cultural elements represent a migration to Japan from Korea or China. Yet it is also certain that the migration was not of an order to change the character of the men who had inhabited the islands from Jōmon times. Although Yayoi culture undoubtedly represents an admixture of new sanguineous elements, it seems likely that the chief strain of proto-Japanese found all over the country during the Jōmon period was not disrupted but was carried over into later ages. This point of view is supported by the accounts of the "men of Wo," found in the Chinese history *Wei chih.*

Chinese chronicles. Japan first appears in Chinese chronicles under the name of Wo (in Japanese, Wa). The Han histories relate that "in the seas off Lo-lang lie the men of Wo, who are divided into more than 100 states, and who bring tribute at fixed intervals." Lo-lang was one of two Han colonies established in the Korean peninsula in 108 BC, and it is beyond doubt that the country of Wo was Japan. A Later Han (23–220) history records that in AD 57 the "state of Nu in Wo" sent emissaries to the Later Han court and that the Emperor gave them a gold seal. The "state of Nu," located around what is now Hakata Bay, in Kyushu, was one of more than 100 states into which Wo was divided. This account was confirmed by a gold seal, apparently the identical seal awarded by the Chinese emperor, unearthed on the island of Shikano, at the mouth of Hakata Bay, in 1748. Later, in the latter half of the 2nd century, there was civil war in the state of Wo; a woman called Pimiko (Himiko in Japanese) used her religious authority to pacify the land, and there came into being a union of more than 30 states, which opened communications with the Wei dynasty (AD 220–264) in China. Wei, too, sent emissaries to Wo, and friendly relations between the two sides continued during the first half of the 3rd century. The *Wei chih* contains a detailed account of the route from Lo-lang to the court of the Wo queen in "Yamatai." Scholars are divided as to whether Yamatai was located in northern Kyushu or in the Kinai district. If it was in northern Kyushu, then the union of states was a purely local government, unrelated to the Yamato court of later times, but if it was in the Kinai district, then it would be natural to see it as the ancestor of that court. This would suggest, in turn, that Japan had already achieved a considerable degree of political unification. But it seems most likely Yamatai was a local centre of power in Kyushu and that unification did not take place until a century later.

The "state of Nu in Wo"

According to the *Wei chih,* the people of Wo had already reached a fairly high degree of civilization. Society had clear-cut divisions of rank, and the people paid taxes. There were impressive raised-floor buildings. The various provinces held fairs where goods were bartered. Since there were exchanges of letters with Wo, it seems, too, that there were already some who could read and write.

The ancient period (c. AD 250–710). *The Yamato court and the unification of the nation.* The question of how the unification of Japan was first achieved and of how the Yamato court, with the *tennō* ("emperor of heaven") at its centre, came into being has inspired many hypotheses, none of which has so far proved entirely convincing. Thanks to Chinese and Korean records, however, it is possible to get at least an approximate idea of the date when unification occurred. The relations that Yamatai had begun with Wei were continued with Chin (AD 265–317), the dynasty that replaced Wei; but following the dispatch of a mission in 266, all records of exchanges cease, and it is not until 147 years later, in 413 during the Eastern Chin dynasty (AD 317–419) in China, that the name of Wo again appears in Chinese documents. It is most likely that the blank period resulted from conditions within Japan that made exchanges with other countries impossible. The collapse of Yamatai and the birth pangs of the united nation that took its place probably occurred during this period.

Date of unification

It is possible to push the date of unification of the nation back a few decades earlier than 413: a memorial erected in 414 commemorating the achievements of King Kwanggaet'o (Japanese Hotae) of Koguryŏ (a Korean state, 37 BC–AD 668), describing the fighting between Wo and Koguryŏ that took place on the Korean peninsula from the end of the 4th century into the beginning of the 5th century, makes special mention of a great army sent to the peninsula in 391 by Wo that succeeded in subjugating the kingdoms of Paekche, Kaya, and Silla. Such military success presupposes a long period of preparation and the prior establishment of a Wo foothold on the peninsula. The 8th-century *Nihon shoki* ("Chronicles of Japan"), one of Japan's two oldest histories, mentions the dispatch of troops by Japan in 369. Displays of strength of this kind would hardly have been possible unless Japan were already unified, and the date of the unification of the country may therefore be set at the middle of the 4th century at the latest.

The rise and decline of the Yamato court. At the time of unification, Japan already seems to have been an extremely powerful nation, as attested by the fact that it took on Koguryŏ, which dominated Korea, and established a base for its own power in southern Korea. Paekche, in the west of southern Korea, was a friendly state that paid tribute to Japan, while Kaya (Japanese, Mimana), at the southern extremity of the peninsula, was under direct Japanese jurisdiction. Tributes from these states lined the coffers of the Yamato court and encouraged a marked rise in standards of living. Weavers, smiths, and irrigation experts migrated to Japan from these areas, and the Chinese ideographic script came into Japan at this time, together with Confucian works written in that script.

The Yamato court reached its peak in the early 5th century and thereafter went into a rapid decline. The main reason was that the states of the Korean peninsula, as a result of shifts in international relationships, broke away from Japan, so that the latter was no longer able to rely on tribute from them. Japan therefore conceived the idea of borrowing the authority of the Chinese court in achieving the subjugation of the Korean kingdoms. Beginning in 421, it sent envoys to the Liu-Sung dynasty (420–479) to ask that the Japanese emperor be granted the title of generalissimo, with military control over the states of Korea. Such envoys were sent on a number of occasions during the Liu-Sung dynasty and continued until 502, during the Liang dynasty (502–557). These attempts proved futile and are themselves evidence of the decline of Japan's military power.

The Korean rebellion

Japan's difficulties abroad were paralleled by an impasse in domestic affairs. The Yamato court was headed by a hereditary emperor, while its members were drawn from the group of powerful *muraji* (clan leader) families, which had been vassals of the emperor from the start, and another group of powerful families, the *omi* (chieftain), which had sworn allegiance during the process of national unification. The highest officers of government were the *ō-muraji* and the *ō-omi*, the heads and representatives of those two groups. In time, however, some members of these families began to cool in their allegiance to the emperor or even to plot with the states of Korea. In addition, there were ceaseless struggles involving succession to the throne within the Imperial family itself. As a result, Mimana, Japan's domain in Korea, was captured in 562 by the kingdom of Silla, depriving Japan of a powerful foothold on the peninsula. By the end of the 6th century, Japan had reached a low point in both foreign and domestic affairs.

During the declining years of the Yamato court, however, there was one event of the utmost cultural importance: the introduction of Buddhism from the Korean kingdom of Paekche. The date of its introduction is traditionally set at either 538 or 552, but it seems likely that Buddhist beliefs had begun spreading among ordinary Japanese at a much earlier date. Buddhism at first was an object of wonder and admiration, a rare item of foreign culture symbolized by its beautiful statuary, its imposing religious paraphernalia, and its majestic temples. The Buddhism that first spread among the Japanese was almost certainly a simple reliance on the magical powers of the religion in seeking various benefits in the present world. True understanding of its doctrines did not come until the time of Shōtoku Taishi (Prince Shōtoku).

Introduction of Buddhism

The period from the latter half of the 3rd century until the beginning of the 7th century is known to archaeologists as the age of the Tumulus, or Kofun, culture, for burial mounds were then erected over a wide area, and their shapes and contents (the objects buried with the dead) give a good idea of the material aspects of the everyday life of the time. These include the well-known *haniwa* tomb sculptures.

The idealized government of Shōtoku Taishi. The Yamato court that fell into such desperate straits toward the end of the 6th century was to be resuscitated by efforts made within the Imperial family itself, efforts that in the course of a century reformed the government of the country and set it moving toward formation of a centralized state more suited to the new age. The movement was touched off by the theories of ideal government expounded by Prince Shōtoku, who, as regent for his aunt, the empress Suiko, took charge of the nation in these difficult times. Prince Shōtoku took the Buddhist spirit of peace and salvation for all beings as the ideal underlying his government. He made no move, even, to charge the known murderer of the previous emperor but worked to convince him gradually, through the ideas of Buddhism, of the wrong he had done.

The Prince's most striking achievement in the field of domestic government was his establishment of a system of 12 court ranks in 603 and the "Seventeen Article Constitution" in 604. The former, which made clear the relative stations of persons working at the court by giving them caps of different colours, aimed to encourage efficient use of persons of ability and give the court a proper organization and etiquette of its own. The constitution consists of 17 simple articles setting forth the ideals of the state and rules for human conduct. It distinguishes the ruler, his ministers, and the people as the three human elements making up the state and clearly lays down the duties and rights of each; it thus set the pattern of a centralized state presided over by a single ruler, and it provided a kind of basic law of the nation.

The constitution of 604

Shōtoku's chief achievement in foreign relations was the opening of relations with Sui dynasty (581–618) China. The exchanges between Japan and China in the 5th century had placed Japan in the position of a tributary state. Prince Shōtoku opened relations with Sui on an equal basis, and envoys were exchanged by the two countries. He also sent Japanese students to China to learn directly from Chinese culture, which had hitherto reached Japan via the states of Korea. Shōtoku was a profound student of Buddhism who gave lectures on the scriptures and himself wrote commentaries. His commentary on the Lotus Sūtra, four volumes of which survive in the original draft written by the Prince himself, may be called the oldest written work of known authorship in Japan.

As Buddhism gained ground, imposing temples were built in the Chinese style. The astonishment aroused by these great buildings—often with more than one story and with massive tiled roofs—that were built where there had been only low, thatched houses may well be imagined. A new civilization descended on Japan almost overnight. Of the temples built at the time, all that has survived of most of them are the foundation stones, but the Hōryū-ji, founded in 607 at Ikaruga in present Nara Prefecture, still preserves its ancient wooden structures; indeed, it is the oldest wooden building in the world.

The Taika reforms. The death of Prince Shōtoku, in 622, prevented his ideals of government from bearing full fruit. The Soga family, regaining its former powers, exterminated Prince Shōtoku's son Yamashiro Ōe in 643 and all his family. At the same time, however, the students whom Shōtoku had sent to China were returning to Japan with accounts of the power and efficiency of the T'ang dynasty (618–907), which had overthrown the Sui dynasty and unified China. These accounts impressed on educated men the need to reform the government, strengthen the power of the state, and take every step to prepare against possible pressure from outside.

In 645 Prince Nakano Ōe and Nakatomi Kamatari engineered a coup d'état within the palace, killing the Soga family and wiping out all forces opposed to the Imperial family. They then set about establishing a system of centralized government with the emperor as absolute monarch at its head. An edict issued in 646 abolished private ownership of land and men by the wealthy families. The land thus taken over by the state was to be parcelled out among all who had attained a certain age, with the right to cultivate, in exchange for which the tenants were to pay a fixed tax. Provisions were also made for a governmental system embracing a capital city and local administration and for defense and communications facilities. A system was also established whereby a kind of "complaints box" was installed at court to give people a chance to appeal directly to the emperor. The main outlines of the reforms were drawn up in about five years. They are given the name Taika reforms after the *nengō* ("year name")—the first such in Japanese history—that was given to the era at that time. In the countries of East Asia, era names are a symbol of an independent nation, a sign that the sovereign's authority is effective.

Abolition of private land-ownership

Not long after the Taika reforms, Japan became involved in a dispute that led it to send troops to Korea. Paekche, whose royal castle had been captured in 660 by the combined forces of T'ang (China) and Silla (another Korean kingdom), called on Japan for help. Japan, which had traditionally been friendly with Paekche, sent a large army; it was crushed, however, in 663, by a combined T'ang and Silla army at the mouth of the Pak River. Japan withdrew entirely and gave up any further intervention on the peninsula. The Japanese ruler of the time, the empress Saimei, went to northern Kyushu and directed operations personally, even though she was already 67 at the time. The Empress was succeeded by Prince Nakano Ōe, who, as the emperor Tenji, directed his attention to domestic affairs. He built fortifications in Kyushu to prepare for an expected T'ang and Silla invasion and amended the system established by the Taika reforms so as to make it more suitable to the practical needs of the state. His younger brother, the emperor Temmu, similarly devoted his energies to domestic government; he had the Taika reforms set forth in written codes, which comprised *ritsu-ryō* political structure.

The ritsu-ryō system. The *ritsu-ryō* can be divided between *ritsu,* the criminal code, and *ryō,* the administrative and civil codes. A similar system had long been in force in China, and the Japanese *ritsu-ryō* was an imitation of the *lü-ling* of T'ang China and incorporated some of its articles just as they stood. Where different local conditions called for amendment, however, amendments were made without hesitation; it is a good early example of the skill of the Japanese in importing foreign culture.

The Japanese emperor, for example, was in some respects an absolute monarch who ruled over the whole country as the head of a bureaucracy in the same manner as the emperor of China. Yet at the same time he was also the traditional high priest who maintained peace for the land and people by paying tribute to the gods and sounding out their will. Thus the central government was headed by twin agencies—the Council of State (Dajōkan), which combined within its functions the various practical aspects of administration, and the Office of Deities (Jingikan), which was in charge of the worship of the gods. Prospective bureaucrats were required to study at a central college and to pass prescribed examinations; during their term of office their achievement was subjected to scrutiny once a year, and their rank and position were adjusted in accordance with the results. This was based on the highly developed bureaucratic system of China, yet the *ritsu-ryō* system was not too bound by its provisions to provide special favours for men of high rank and good family. This, too, was a compromise between the new principles of the *ritsu-ryō* system and the old spirit of respect for birth. The provinces were divided into three types of administrative division: the *kuni,* or *koku* (province), the *kōri,* or *gun* (county), and the *sato,* or *ri* (village), to be administered by officials known as *kokushi, gunji,* and *richō,* respectively. The posts of *kokushi* were filled by members of the central bureaucracy in turn, but the posts of *gunji* and *richō* were filled by members of prominent local families.

The Imperial government

The people were divided into two main classes, freemen and slaves. The slaves were the possession of the government, the aristocracy, and the shrines and temples; as such they were obliged to provide unlimited labour, but their total number accounted for less than one-tenth of the population. The majority of the free population were farmers. At the age of six, each male child was apportioned paddy fields that remained his to cultivate for life. A tax was levied on the produce of the paddies, and a head tax was levied on adult males. The paddy field tax was low (about 3 percent of the crop), but the head tax, payable in handicrafts such as silk and hemp, imposed a heavy burden. Moreover, the transport of the goods from the provinces to the capital was the responsibility of the taxed, which involved an enormous labour for those living in distant parts. Adult males were also obliged to give military service and to provide labour for public works at the command of the local *kokushi,* amounting to not more than 60 days per year. Since the government's finances depended on such tribute from the common people, whenever the latter found the burden too much and fled from their registered homes to avoid paying taxes the government felt the pinch immediately.

The lowest ranking freemen were the groups of smiths, tanners, and others engaged in manufacturing. They were mostly the descendants of immigrants who inherited their trades and paid their taxes in the form of manufactured goods or by working for fixed periods in the government workshops.

Shrine and temple aristocracy

All land was in principle the property of the state. Most of the land was distributed equally among the people, but, apart from this, land of a certain annual yield was given to bureaucrats and other high-ranking persons as stipends and to Shintō shrines and Buddhist temples as sources of revenue. Land other than paddy fields was left to the individual to use as he pleased. There was a need to open up new paddy fields as a means of providing for a growing population, but the *ritsu-ryō* system made inadequate provision for this process. In time, the government began to encourage the opening up of new land, and in 743 the system was changed to give permanent private possession of such land to the person who had first put it under cultivation. As a result, the aristocrats and the shrines and temples frantically set about putting land under cultivation in order to increase their own privately owned territories. The principle of public ownership of land provided for in the *ritsu-ryō* system began to crumble, and as it did so the whole system of government grew increasingly shaky.

Nara period (710–784). *Beginning of the Imperial state.* In 710 the Imperial capital was shifted a short distance from Asuka to Nara. For the next 75 years, with minor gaps, Nara was the seat of government, and the old custom of changing the capital with each successive emperor was finally discarded. During this period, the centralized government provided for under the *ritsu-ryō* structure worked well, but a still conspicuous feature is the brilliant flowering of culture, especially Buddhist culture. The leaders in its promotion were the emperor Shōmu and his consort, Kōmyō. Immediately on his accession, Shōmu, who from childhood had been given a thorough schooling as future emperor, showed an eager concern to promote the stable livelihood of the people. Convinced that the Buddhist faith was a means to ensuring both the happiness of the individual and peace for the country as a whole, he introduced strong doses of Buddhism into his government.

Official encouragement of Buddhism

One of the measures he took was the founding of the temples known as *kokubun-ji.* Each province was to build a monastery known as *kokubun-ji* and a nunnery known as *kokubun-niji,* each with a seven-story pagoda and each housing a statue of the Śakyamuni Buddha. Each monastery was to have 20 monks, each nunnery 10 nuns, whose constant task would be to recite the scriptures and offer up prayers for the welfare of the nation. Just as the temporal world had its *kokushi* (governors) in each province to attend to its administrative and juridical matters, so the spiritual world would have officially appointed monks and nuns, distributed evenly among the provinces, to attend to the spiritual needs of the people.

The second measure taken by Shōmu was the construction of the Tōdai-ji as *kokubun-ji* of the capital and the installation within it of a huge bronze figure of the Vairocana Buddha as supreme guardian deity of the nation. The casting of the Daibutsu (Great Buddha) was a tremendously difficult task. The Emperor, however, called on the people at large to contribute to the project, in however humble a way, and thereby partake of the grace of the Buddha. The great image that was produced as a result, though damaged in later ages, still stands in the Tōdai-ji and is famous the world over as the Great Buddha of Nara.

The marriage of Buddhism and politics that was Shōmu's ideal was to cause trouble in the following era. The temples gradually amassed vast wealth, and the monks acquired high political positions and began to interfere in secular affairs. A movement to counter such abuses arose among the aristocracy, the leaders of the movement being the Fujiwara family, descendants of Nakatomi Kamatari, who had played such an important role in the Taika reforms. Kamatari and his son Fuhito (later given the surname Fujiwara) had supervised compilation of the codes that comprised the *ritsu-ryō* system and had become prominent figures at court as a new type of bureaucrat-noble. The subsequent progress of the family's fortunes was not always smooth. In particular, the emphasis on Buddhism in government had obliged them to lie low. At the end of the 8th century, however, when it seemed to them that the evils of Buddhistic government were threatening the future of the nation, they set on the throne a new emperor, Kōnin, who had no leanings toward Buddhism. Kōnin's son, the emperor Kammu, who was of a similar mind, shifted the capital to Heian-kyō (present Kyōto) to sever connections with the temples of Nara and reestablished government in accordance with the *ritsu-ryō* system.

Culture in the Nara period. The cultural flowering centring on Buddhism was an outcome of lively exchanges with other nations. Four times within 70 years the government sent official missions to the court of T'ang, and each time they were accompanied by a large number of students who went to study in China. By this time T'ang had formed a great empire that controlled not only the central plains of China but parts of Mongolia and Siberia to the north and of Central Asia to the west.

The *Man'yō-shū* anthology

Japanese culture, borrowing from the T'ang, whose capital, Ch'ang-an, was a great international city, thus showed in the Nara period marked international flavour. The ceremony of consecration of the Great Buddha of the Tōdai-ji, for example, was conducted by a Brahmin high priest born in India, while the music was played by musicians from all over the Far East. But despite this strongly international flavour, respect was also shown for traditional Japanese ways and outlooks. An outstanding example of this respect is the collection of Japanese verse known as *Man'yō-shū* (*c.* 8th century AD), an anthology of 4,500 poems both ancient and contemporary. The poets range over all classes of society, from the emperor and members of the Imperial family through the aristocracy and the priesthood to farmers, soldiers, and prostitutes, and the scenery celebrated in the verse represents districts all over the country. The poems deal directly and powerfully with basic human themes, such as love between men and women or between parents and children, and are deeply imbued with the traditional spirit of Japan, scarcely influenced at all by Buddhist or Confucian ideas. The anthology had immense influence on all subsequent Japanese culture.

The compilation of Japan's two most ancient histories, the *Koji-ki* and *Nihon shoki,* also took place at the beginning of the 8th century. Both works are extremely important, for they draw on oral or written traditions handed down from much earlier times.

The Heian period (794–1185). *Changes in ritsu-ryō government.* In 794, as noted above, the emperor Kammu shifted his capital to Heian-kyō, cut the ties between government and Buddhism, and revived government in accordance with the *ritsu-ryō.* Commanding that the provisions of the *ritsu-ryō* system be enforced, he also amended those articles that were no longer relevant to the age. Since it was difficult in practice to carry out the allocation of rice fields once every six years, this was amended to once in 12 years. A tighter watch was imposed on corruption among local officials. The original system of raising conscript troops was abolished, and troops were thenceforth selected from among the sons of local government officers and persons of rank. An alien tribe known as the Emishi in the northern districts of Honshu was brought under government control. Those Emishi who submitted to government forces were resettled throughout the empire and quickly assimilated to the existing population.

Buddhism was forbidden to interfere in affairs of state, but as a religion it was encouraged to fulfill its proper functions. Two brilliant monks, Saichō and Kūkai, were sent to China to study. Each of them, on his return to Japan, established a new sect of Japanese Buddhism: the Tendai sect, founded by Saichō, and the Shingon sect, founded by Kūkai. In the Nara period, Buddhism had been no more than a transplantation of the Buddhism of T'ang China, but the two new sects, though basically derived from China, were reworked in a characteristically Japanese fashion. As headquarters of their new sects, Saichō and Kūkai founded, respectively, the Enryaku-ji (also known as the Hieizan-ji) on Mt. Hiei (Hiei-zan) and the Kongōbu-ji (*ji*, "temple") on Mt. Kōya (Kōya-san). The two sects were thenceforth to form the twin mainstreams of Japanese Buddhism.

After Kammu, successive emperors carried on his policies, and society enjoyed some 150 years of peace. The formal aspects of government, at least, were carefully observed, and the supplementing of the legal codes, the compilation of histories, and the minting of coins all took place frequently in accordance with precedent. The social reality, however, became increasingly chaotic, so that form and actuality were soon travelling along quite different courses. The very foundations of *ritsu-ryō* government had begun to crumble because of the difficulty of finding enough rice fields to distribute to the people and the decline in government revenue resulting from the impoverishment of the masses.

Decline of *ritsu-ryō* government

A good example of the split between form and reality is the fact that while, on the surface, appointments to official posts at court were made just as they had always been, real power shifted to other posts that were newly created as the occasion demanded. Typical of such new posts were those of *kurōdo,* a kind of secretary and archivist to the emperor, and the *kebiishi,* who had total control over the police and the judicature. The two supreme examples of such posts were those of *sesshō* (regent) and *kampaku* (chief councillor). The original role of the *sesshō* was to attend to affairs of state during the minority of the emperor, while the *kampaku*'s role was to attend to state matters for the emperor even after he had come of age. Neither post had been foreseen by the *ritsu-ryō* system, which was rooted in the principle of direct rule by the emperor.

The rule of Imperial succession

In the middle of the 9th century, however, when the emperor Seiwa ascended to the throne at the age of nine, his maternal grandfather, Fujiwara Yoshifusa, became *sesshō.* Yoshifusa's son Mototsune became *sesshō* during the minority of the emperor Yōzei, then in the reign of the emperor Uda he created the post of *kampaku.* It thus became the established custom that a member of the Fujiwara family should serve as *sesshō* and *kampaku.* In order to become *sesshō* or *kampaku,* it was necessary that the person concerned should marry his daughter into the Imperial family, then establish the resulting offspring as emperor. In other words, an important qualification was that one should be the emperor's maternal grandfather or father-in-law. As a result of this complex system, there were constant struggles at court involving the expulsion of members of other families by the Fujiwara family or wrangling among the Fujiwara themselves.

One of the most celebrated affairs involving the expulsion of a member of another family by the Fujiwara was the removal of Sugawara Michizane from his post as minister and his exile to Kyushu. Born into a family of scholars, Michizane was himself an outstanding scholar whose ability in writing Chinese verse and prose was said to rival that of the Chinese themselves. Recognizing his talent, the emperor Uda singled him out for an attempt to break the authority of the Fujiwara family. As part of his plan, Uda appointed Michizane and Fujiwara Tokihira to a succession of government posts. In 899 Uda's successor, the emperor Daigo, simultaneously appointed Tokihira as his minister of the left and Michizane as minister of the right. The Fujiwara objected strenuously. In 901 Tokihira falsely reported to Daigo, who was sympathetic to the Fujiwara, that Michizane was plotting treason. The matter was taken up officially and Michizane was demoted to a ministerial post in Kyushu, effectively sending him and his family into exile.

The culture of the 9th century was a continuation of that of the 8th, insofar as its foundations were predominantly Chinese. The writing of Chinese prose and verse was popular among scholars, and great respect for Chinese customs was shown in the daily lives of the aristocracy. Many Buddhist monks went to China to bring back as yet unknown scriptures and iconographic pictures. Buddhist sculpture and paintings produced in Japan were done in the T'ang style. At the end of the 9th century, however, Japan cut off formal relations with T'ang China, perhaps because of the expense involved in sending regular envoys and perhaps because of the political unrest accompanying the breakup of the T'ang empire. The practical result was the stimulation of a more purely Japanese cultural tradition. Japanese touches were gradually added to the basically T'ang styles, and a new culture slowly came into being; but it was not until the 10th century and later that this tendency became a strong current.

Aristocratic government at its peak. From the 10th century and through the 11th, successive generations of the Fujiwara family continued to control the nation's government by monopolizing the posts of *sesshō* and *kampaku,* and the wealth that poured into their coffers enabled them to lead lives of the greatest brilliance. The high-water mark was reached in the time of Fujiwara Michinaga. Four of his daughters became consorts of four successive emperors, and three of their sons became, respectively, the emperors Go-Ichijō, Go-Suzaku, and Go-Reizei. Government during this period was based mostly on precedent, and the court had become no more than a centre for ceremonies. Court ministers were content to perform prescribed rites on prescribed days and were utterly unfitted to deal with any sudden social crisis that might confront them.

The growth of private estates

The *ritsu-ryō* system of public ownership of land and men survived in name alone; land passed into private hands, and men became private citizens. Typical of the new privately owned lands were the *shōen* ("manors"), which developed on the basis of rice fields under cultivation since the adoption of the *ritsu-ryō* system. Since the government encouraged the opening up of new land during the Nara period, the temples and aristocrats with resources at their disposal hastened to develop new areas, and vast private lands accrued to them. The owners of the new lands used one pretext or another to obtain special exemption from taxes, so that the *shōen* gradually became nontaxpaying estates. The increase in such *shōen* thus came to pose a serious threat to the government, which accordingly issued frequent edicts intended to check the formation of new estates. This merely served, however, to establish more firmly the position of those already existing and failed to halt the tendency for such land to increase. Since the owners of the *shōen* were the same aristocracy and high officials that made up the government, it was extremely difficult to change the situation.

Although the aristocracy and temples around the capital enjoyed exemption from taxes on their *shōen,* the same privileges were not available to powerful families in the provinces. These, accordingly, presented their *shōen* to members of the Imperial family or the aristocracy, concluding agreements with them that the latter should become owners in name while the former themselves retained rights as actual administrators of the property. Thanks to such agreements the estates of the aristocracy went on increasing steadily and their incomes swelled proportionately. The *shōen* of the Fujiwara family in particular reached such vast proportions that it was said that among them they owned the whole country.

While the aristocracy was leading a life of luxury on the proceeds from its estates, the first stirrings of a new power in the land—the warrior, or samurai, class—were taking place in the provinces. Younger members of the Imperial family and lower ranking aristocrats who were dissatisfied with the Fujiwara monopoly of high posts would take up posts as local officials in the provinces, where they settled permanently, acquired lands of their own, and established their own power. In order to protect their territories, they began to press the local inhabitants into service and to give them arms, thus building up armed forces of their own. As a consequence, when men of true martial ability and sufficient autonomy emerged, the slightest incident involving any one of them was liable to provoke armed conflict. The risings of Taira Masakado (died 940) in the Kantō district and of Fujiwara Sumitomo (died 941) in western Japan had an enormous effect in lowering the government's prestige and encouraging the desolation of the provinces.

The development of the *kana* syllabaries

During the 10th century a truly Japanese culture developed, one of the most important contributing factors being the emergence of indigenous scripts, the *kana* syllabaries. Until then, Japan had no writing of its own; Chinese ideographs were used partly for their sense and partly for their pronunciation in order to represent the Japanese language, which was entirely different from Chinese. The educated men and women of the day, however, gradually evolved a system of writing that used a purely phonetic, syllabic script formed by simplifying a certain number of the Chinese characters; another script was created by abbreviating Chinese characters. These scripts, called *hiragana* and *katakana*, respectively, made it possible to write the national language with complete freedom, and their invention was an epoch-making event in the history of the expression of ideas in Japan. Thanks to the *kana,* a great amount of verse and prose in Japanese was to be produced.

Particularly noteworthy in this respect were the daughters of the Fujiwara, who, under the aristocratic government of the day, became the consorts of successive emperors and surrounded themselves with talented women who vied with each other in learning and the ability to produce fine writing. The *hiragana* script provided such women with an opportunity to create works of literature. Among such works, the *Genji monogatari* (*The Tale of Genji*), a novel by Murasaki Shikibu; and the *Makura no sōshi* (*The Pillow Book of Sei Shōnagon*), a collection of vivid scenes and incidents of court life by Sei Shōnagon, who was a lady-in-waiting to the empress Sadako, are masterpieces that hold a place in world, and not merely Japanese, literature. The waka, the native Japanese verse form, was an indispensable part of the daily lives of the aristocracy, and proficiency in verse-making was counted an essential accomplishment for an intellectual. Such circumstances led to the compilation in 905 of the *Kokinshū* (or *Kokin wakashū*), the first of a series of anthologies of verse made at Imperial command.

The same trend toward the development of purely Japanese qualities became strongly marked in Buddhism also. Both the Tendai and Shingon sects produced a succession of gifted monks and continued, as sects, to flourish. But, being closely connected with the court and aristocracy, they tended to pursue worldly wealth and riches at the expense of purely religious goals, and it was left to the Jōdo (Pure Land) sect of Buddhism to preach a religion that sought to arouse a desire for salvation in ordinary men. Pure Land Buddhism, which became a distinct sect in the 12th and 13th centuries, expounded the glories of the paradise of Amida (Amitābha, or Buddha of Infinite Light)—the world after death—and urged all to renounce the defilements of the present world for the sake of rebirth in that paradise; it seemed to offer an ideal hope of salvation in the midst of the collapse of the old order. It was a very approachable religion in that it eschewed difficult theories and ascetic practices, teaching that in order to achieve rebirth it was only necessary to invoke the name of Amida and dwell on the marks of his divinity. This same teaching also inspired artists to produce an astonishing number of representations of Amida in both sculpture and painting. The mildness of his countenance and the softly curving folds of his robe contrasted strongly with the grotesque Buddhist sculpture in the preceding age and represented a much more truly Japanese taste.

The signs of the growing independence of Japanese culture, apparent in every field, were an indication that by now, two centuries after the first busy ingestion of continental culture, the process of absorption was nearing completion.

Government by cloistered emperors (insei) and the rise of the samurai. The powerful authority wielded by members of the Fujiwara family as *sesshō* and *kampaku* was maintained by their blood relationship on the maternal side to successive emperors; once such a relationship disappeared, their power was bound to weaken. This, in fact, is what happened; the emperor Go-Sanjō ascended the throne even though he was not born of a daughter of the Fujiwara, while Michinaga's sons Yorimichi and Norimichi both gave their daughters to be Imperial consorts without obtaining the desired birth of an Imperial prince. As a result of these and other circumstances, real political power passed from the *sesshō* and *kampaku* to the "cloistered emperors" during the latter half of the 11th century; in other words, it passed to emperors who had already abdicated and taken Buddhist vows—thus nominally renouncing the world—yet who wielded a very real power behind the scenes. This system, known as *insei* ("cloistered government"), was perhaps a more natural arrangement insofar as it represented a shift from government by matrilinear relatives of the emperor to government by patrilinear relatives. In continuing to treat the actually reigning emperor as a pure figurehead, however, it was no whit better than the old *sesshō-kampaku* system. The cloistered-emperor system continued for a long period, although the emperors Shirakawa, Toba, and Go-Shirakawa—who retired for periods of 43, 27, and 34 years, respectively—were the only ones to wield absolute, behind-the-scenes power.

The power of retired emperors

The *insei* system was inspired by no particular ideal and conformed to no particular rules. The one common feature of each reign was that the emperor became a Buddhist priest and governed in a way that theoretically respected the teachings of Buddhism. In practice, this "Buddhism" was more preoccupied with construction of ostentatious temples than with true belief. Other signs of the same trend were the frequent journeys of retired emperors to worship at distant temples and their edicts strictly prohibiting the killing of living creatures; it did not concern them that such edicts deprived a large number of their subjects of their occupations.

The nominal respect for Buddhism spurred on the secularization of the religion. Properly speaking, the world of Buddhism should have been one in which factionalism could have no part, a world in which nothing counted but wisdom, virtue, and experience. But at this time large numbers of aristocrats were taking holy vows and going to live in the temples, which thus became centres of factionalism and intrigue. Most of the higher positions in the religious world were occupied by members of the Imperial family and former aristocrats. This effectively closed advancement to commoners, and the lower ranking monks in the temples harboured a grudge against their superiors on this account. Whenever some particularly serious grievance arose, they would march in a body on the capital and try to force acceptance of their demands by a direct appeal to the court. Some idea of the nuisance they constituted can be gained from the fact that even the retired emperor Shirakawa—the most powerful of the cloistered emperors—ranked them with the waters of the Kamo River and the dice in games of chance as one of three superhuman forces that he was powerless to control. Nor did the monks hesitate to resort to armed force; it was an age in which a priesthood ostensibly committed to compassion and respect for life in all its forms could openly bear arms and engage in slaughter.

The rise of the warrior class

Another feature of the age was the rise of the warrior class. With the development of government by the cloistered emperors, the more powerful of the samurai, who, as noted above, first established their power in the provinces, gradually gathered in or near the capital, where they acted as military police. Associating with the Fujiwara court nobles and aristocracy, they gradually established a foothold at court. Outstanding among these samurai were the Minamoto (or Genji) family, descendants of the emperor Seiwa, and the Taira (or Heike) family, descendants of the emperor Kammu. The Taira had at first settled in the Kantō district, where they extended their influence over a wide area; but they had suffered a setback in the uprising of Taira Masakado and had finally lost their hold on the Kantō district as the result of another, later rising by Masakado's descendant Tadatsune. The Minamoto clan, favourites of the Fujiwara, had been a prominent family in the capital from the start, but their fame as a warrior clan was greatly heightened in the mid-11th century when after 12 years of hard fighting they quelled a rising by the Abe family in the Tōhoku district. Minamoto Yoshiie, who played an important part in the fighting, became the nation's most celebrated warrior, and many powerful clans made voluntary vows of allegiance to him and presented him with land in return for his protection. Yoshiie, however, had no son to match him in military prowess, and the Taira took advantage of this relative decline to advance their own fortunes again. Taking advantage of the *insei* system as a means to their own political advancement, they curried favour with the retired emperors. Taira Masamori and his son Tadamori served as governors in provinces in western Japan, building up their own power in the area, and aided the retired emperors' programs of temple building by erecting a large number of new temples and pagodas. Tadamori also tried his hand at trade with Sung dynasty China as a means of amassing wealth. In such ways, the social position of the Taira rose steadily, so that Tadamori's son Kiyomori could take his place alongside the aristocracy.

Discord between retired emperors and reigning emperors combined with internal differences within the Fujiwara family to split the Imperial family and nobility into two parties, which enlisted the Minamoto family and the Taira family, respectively, on their own sides. The two sides eventually clashed openly in Kyōto in what is known as the Hōgen Disturbance (July 1156). The conflict was on a small scale, the outcome determined by a single night's fighting, yet it was highly significant as showing that the power of the samurai was sufficient to sway the nation's government. In the Heiji Disturbance (1159) that followed, the Minamoto were thoroughly defeated, and Taira Kiyomori emerged as the chief power in the land. Although he was a samurai by birth, Kiyomori shared certain aristocratic tendencies of the Fujiwara, and the 20-odd years of Taira rule that followed had a special character of their own. He himself became grand minister of state (*dajō-daijin*) at the court, and more than 50 other official posts were filled by members of his family. His daughter became the consort of the emperor Takakura, and the prince born of the union ascended to the throne in his infancy—a return to government by matrilinear relatives of the emperor. Kiyomori's rule also had its more drastic, soldierlike aspects; thus in a single move he swept 42 court officials from their posts and into exile, and he razed to the ground troublesome temples such as the Tōdai-ji and Kōfuku-ji. His repairing of the Inland Sea route and his encouragement of trade with Sung China were measures that would never have occurred to a Fujiwara government.

The Heiji Disturbance

While the Taira family thrived in the capital, the descendants of the original Minamoto were quietly building up their strength in the provinces. Finally Yoritomo, a descendant in the direct line of the Minamoto family, who grew up in exile at Izu, rallied the Minamoto and sent his younger brothers Yoshitsune and Noriyori to attack Kyōto. The final rout of the fleeing Taira forces on the sea off the island of Shikoku put a more or less decisive end to the swing of fortune between Minamoto and Taira.

It also marked an important turning point in Japanese history, since Yoritomo's establishment of a military government, or shogunate, in Kamakura may be seen as the beginning of rule by a samurai class backed up by a feudal system and the end of the ancient monarchical system of court and aristocracy. The shogunate, or *bakufu* (literally, "tent government," the name for the field headquarters of a campaigning warrior), was to hold effective political control in Japan until the restoration of Imperial power in 1868. (T.Sa.)

The Kamakura bakufu (1192–1333). *The establishment of military government.* The establishment of the shogunate by Minamoto Yoritomo at the end of the 12th century marks the beginning of a new era, one in which independent government by the warrior class successfully opposed the political authority of the civil aristocracy. Yoritomo established his headquarters in Kamakura and entrusted the suppression of the powerful Taira family to his younger brothers Noriyori and Yoshitsune. Meanwhile, he gathered a following in eastern Japan as a foundation for a new military government. As a first step he set up in 1180 the Samurai-dokoro (Board of Retainers), a disciplinary board to control his military vassals. General administration was handled by a secretariat, which was opened four years later and known as the Kumonjo (later renamed the Mandokoro). In addition, a judicial board, the Monchūjo, was set up to handle lawsuits and appeals. Under these institutions, the organization of the *bakufu* gradually took shape.

Bakufu institutions

In 1185, after the destruction of the Taira family, Yoritomo appointed military governors (*shugo*) in all the provinces and military stewards (*jitō*) in both public and private landed estates. It was the job of the *shugo* to recruit metropolitan guards and keep strict control over subversives and criminals. The *jitō* collected taxes, supervised the management of landed estates, and maintained public order.

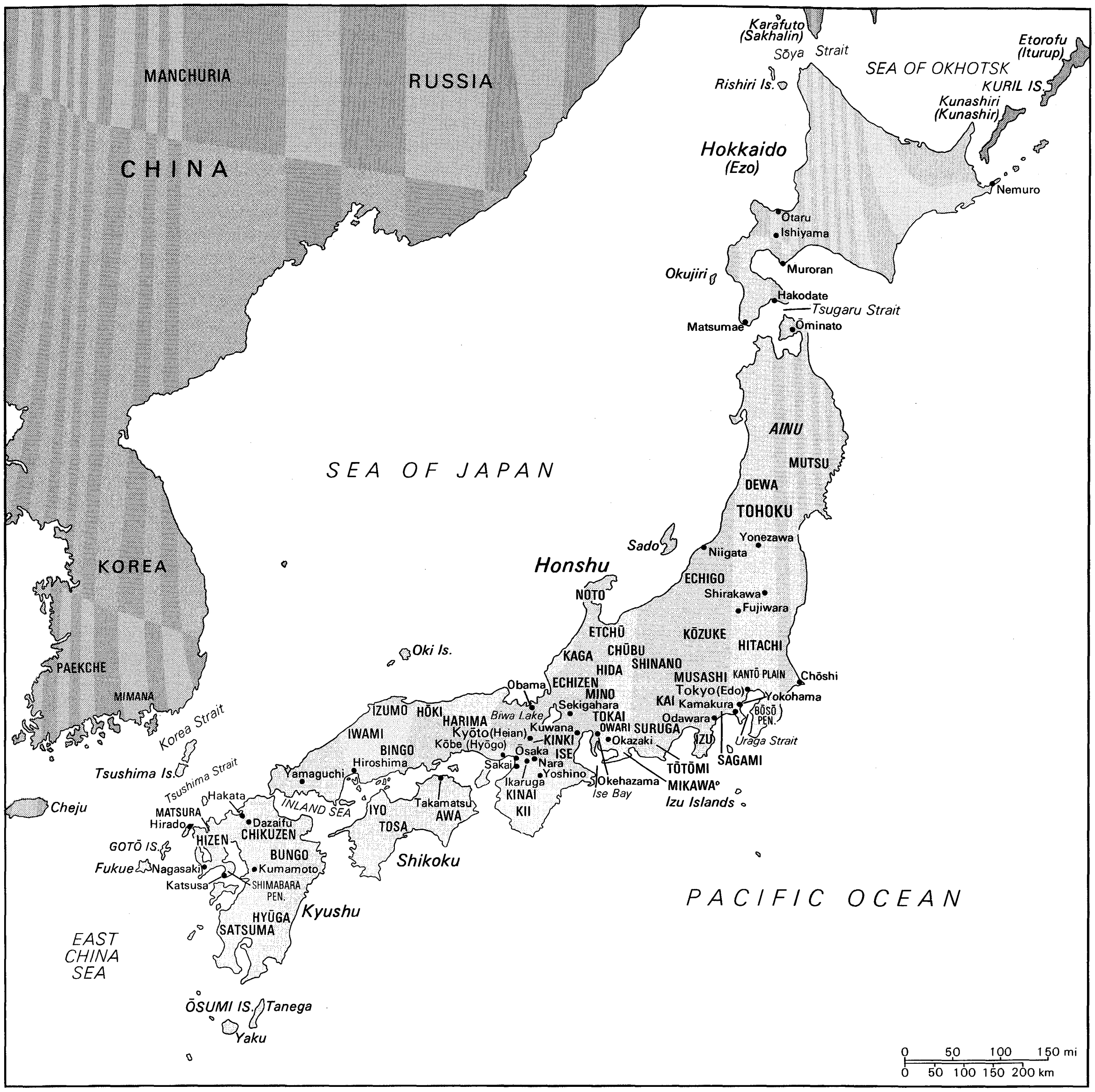
In 1189 Yoritomo finally destroyed the great Fujiwara family of Mutsu Province, which had sheltered his rebellious brother Yoshitsune. Three years later Yoritomo went to Kyōto and was appointed shogun (an abbreviation of *seii taishōgun*; "barbarian-quelling generalissimo"), the highest honour that could be accorded a warrior. At first the chief base of the *bakufu* lay in the landed estates seized from the Taira family and in the limited administrative revenues from public estates in provinces granted to Yoritomo by the Imperial court. But later the *bakufu* was able to expand its influence over those public estates that were still controlled by the civil provincial governors, as well as the private estates of the civil aristocracy and the temples and shrines.

The regency government. After the death of Yoritomo in 1199, real power in the *bakufu* passed into the hands of the Hōjō family, from which Yoritomo's wife, Masako, had come. In 1203 Hōjō Tokimasa, Masako's father, assumed the position of regent (*shikken*) for the shogun, an office that was held until 1333 by nine successive members of the Hōjō family. Taking advantage of disputes among Yoritomo's generals, the Hōjō overthrew their rivals, and after three generations the direct line of descent from Yoritomo had become extinct. Though wielding actual power, the Hōjō family was of low social rank, and its leaders did not aspire to become shoguns themselves. Kujō Yoritsune, a distant relative of Yoritomo, was appointed shogun, while Tokimasa's son Hōjō Yoshitoki (*shikken* 1205–24) handled most government business. Thereafter, the appointment and dismissal of the shogun followed the wishes of the Hōjō family.

This increasing political power of the military led to a conflict with the aristocracy. Hence, the emperor Go-Toba, seeing in the demise of the Minamoto family a good opportunity to restore his political power, in 1221 issued a mandate to the country for the overthrow of Yoshitoki, but few warriors responded to his call. A *bakufu* army occupied Kyōto, and Go-Toba was arrested and banished to the island of Oki in an incident known as the Jōkyū Disturbance, after the era name Jōkyū (1219–22). The *bakufu* now set up its headquarters in Kyōto to supervise the Imperial court and to control the legal and administrative business of the western provinces. The estates of the civil aristocrats and warriors who had joined Go-Toba were confiscated and distributed as rewards among the shogun's vassals. The political power of the *bakufu* now extended over the whole country.

The Jōkyū Disturbance

Meanwhile, the regent Hōjō Yasutoki, to strengthen the base of his political power, reorganized the council of leading retainers into an advisory council known as the Hyōjō-shū. In 1232 the council drew up a legal code known as the Jōei Shikimoku (Jōei Formulary). Its 51



Important Japanese historical sites.

articles set down in writing for the first time the legal precedents of the *bakufu.* Its purpose was simpler than that of the *ritsu-ryō,* the old legal and political system of the Nara and Heian civil aristocracy. In essence it was a body of pragmatic law laid down for the proper conduct of the warrior way of life. Distinctive features of the formulary included a strong emphasis upon the lord-vassal relationship and paternal power and a recognition of female inheritance of land; its sway was gradually extended over the whole country. In 1249 the regent Hōjō Tokiyori also set up a supreme court, the Hikitsuke-shū, to secure greater impartiality and promptness in legal decisions.

The Mongol invasions. The establishment of the regency government coincided with the rise of the Mongols under Genghis Khan in Central Asia. In the space of barely half a century they had established an empire extending from the Korean peninsula in the east and as far west as Russia and Poland. In 1260 Genghis Khan's successor, Kublai, became Great Khan in China and fixed his capital at present-day Peking. In 1271 Kublai adopted the dynastic title of Yüan; shortly thereafter the Mongols began preparations for an invasion of Japan. In the autumn of 1274 a Mongol and Korean army of some 40,000 men set out from present-day South Korea. On landing in Kyushu it occupied the Matsura district of Hizen Province (part of present-day Saga Prefecture) and advanced to Chikuzen. The *bakufu* appointed Shōni Sukeyoshi as military commander, and the Kyushu military vassals were mobilized for defense. A Mongol army landed in Hakata Bay, forcing the Japanese defenders to retreat to Dazaifu; but a typhoon suddenly arose, destroying over 200 ships of the invaders, and the survivors returned to southern Korea.

The *bakufu* took measures against a renewed invasion. Coastal defenses were strengthened, and a stone wall was constructed extending for several miles around Hakata Bay to thwart the powerful Mongol cavalry. Apportioned among the Kyushu military vassals, these public works took five years to complete and required considerable expenditure. Meanwhile, the Mongols made plans for a second expedition. In 1281 two separate armies were arrayed: an eastern army consisting of about 40,000 Mongol, North Chinese, and Korean troops set out from South Korea; and a second army of about 100,000 South China troops under the command of the Mongol general Hung Ch'a-ch'iu. The two armies met at Hirado and in a combined assault breached the defenses at Hakata Bay. But again a fierce typhoon destroyed nearly all of the invading fleet, forcing Hung Ch'a-ch'iu to retreat precipitately. The remnants of the invading army were captured by the Japanese; it is said that of 140,000 invaders, fewer than one in five made good their escape. The defeat of the Mongol invasions was of crucial importance in Japanese history. The military expenditure undermined the economic stability of the Kamakura government and led to the insolvency of many of the military vassals. It led to another prolonged period of isolation from China that was to last until the 14th century. Moreover the victory gave a great impetus to a feeling of national pride, and the Kamikaze (Divine Wind) that destroyed the invading hosts gave the Japanese the belief that they were a divinely protected people.

Samurai groups and farming villages. The Japanese feudal system began to take shape under the Kamakura *bakufu.* At its inception in the Kamakura (1192–1333) and Muromachi (1338–1573) periods, Japanese feudalism presents a different appearance from that of later eras. In the earlier periods feudal warrior-landlords lived in farming villages and carried on agriculture themselves, while the central civil aristocracy and the temples and shrines held huge public and private estates in various provinces and wielded power comparable to that of the *bakufu.* These estates were, in reality, managed by influential resident landlords who had become warriors. They were often the original developers of their districts, becoming officials of the provincial government, agents of the private estates, and military stewards appointed by the *bakufu.* As leaders of a large number of villagers, they laboured to develop the rice fields and irrigation works in the areas under their jurisdiction, and they and other influential landlords constructed spacious homes for themselves in the villages and hamlets where they lived.

Among these were some who were military vassals of the shogun and others who were connected to the aristocracy or the temples and shrines. The military vassals owed their loyalty to the shogun, for whom they performed public services such as guard duty in Kyōto, where the emperor lived with his court. In return, the shogun not only guaranteed these men the traditional landholdings but rewarded them with new lands. This connection between lord and vassal, on which grants of landownership or management were based, gave Japanese society a distinctly feudal character.

The samurai served on the battlefield and in times of peace engaged in hunting and training in the military arts, nourishing a rugged and practical character. The *kyūba no michi* ("the way of the bow and horse"), the samurai ideal of chivalry, grew out of this daily training and the experience of actual warfare. Pride of family name was especially valued, and loyal service to one's overlord became the fundamental morality. This was the origin of the more highly developed code of Bushidō, or Way of the Warrior, of later ages.

The status of women in the samurai families was comparatively high; they were allowed to inherit a portion of the estates, a practice that gradually came to be restricted.

Economic changes in the Kamakura period

After the middle of the Kamakura period, the farming villages in which the warriors resided underwent changes as agricultural practices advanced, and other aspects of society were changing as well. Artisans were usually attached to the proprietors of the private estates and progressively became more specialized along with the growth of consumer demand. Centres for metal casting and metalworking, paper manufacture, and other skills appeared in the localities. The exchange of agricultural products, manufactured goods, and other products thrived; local markets, held on three fixed days a month, became common. Copper coins from Sung (960–1279) China circulated in these markets, while itinerant merchants increased their activity. Bills of exchange were also used for payments to distant localities. In the large ports, specialized wholesale merchants appeared who were called *toimaru.* They served as contractors who stored, transported, and sold goods. Further, it was common for many merchants and artisans to form guilds, known as *za,* organized under the temples, shrines, or civil aristocrats. Such guilds, which resembled European trade guilds in the Middle Ages, gained special monopoly privileges and exemptions from customs duties.

Kamakura culture: the new Buddhism and its influence. During the Kamakura period the newly arisen samurai class came to dominate the ancient civil aristocracy, which continued to maintain the classical culture. Vigorous overseas trade fostered the transmission of Zen Buddhism (in Chinese, Ch'an) and Neo-Confucianism from Sung China. Chinese influences were seen in monochrome painting style (*suiboku-ga*), forms of architecture, certain skills in pottery manufacture, and the custom of tea drinking—all of which assisted the formation of the samurai culture and exerted an enormous influence on everyday life.

Introduction of Zen sects

In matters of religion, the great social changes that took place between the end of the Heian period and the early Kamakura period caused the people to demand a simple standard of faith, in place of the complicated teachings and ceremonies of the ancient Buddhism. The warriors of the farming villages, in particular, demanded a religion that would suit their personal experience. Several new Buddhist sects sprang up which eschewed difficult ascetic practices and recondite scholarship. Among these may be included the Jōdo, or Pure Land, sect and its offshoot, the Shin (True) school. The Zen school sought to open the way to insight by self-effort; hence, it met with a ready response, satisfying the demands of the samurai. At the same time, scholarship and the arts were still deeply linked with esoteric Buddhism, which was a vigorous influence even in Shintō circles (see BUDDHISM).

In scholarly and literary circles, the civil aristocrats of Kyōto confined themselves to the annotation and interpretation of the ancient classics and to the study of precedents and ceremonies. But at the beginning of the Kamakura period, a brilliant circle of waka (poems of 31 syllables) poets gathered around the retired emperor Go-Toba, and an Imperial selection of poems entitled the *Shin kokin wakashū* was compiled. The waka of this period is characterized by the term *yūgen,* which may be described as a mood both peaceful and profound.

Just before the Jōkyū Disturbance the monk Jien (posthumous name Jichin) completed his *Gukanshō* ("Jottings of a Fool"). This is the first work of historical philosophy in Japan, and it provides a comprehensive picture of the rise and fall of political powers from the Buddhist viewpoint. Meanwhile, with the rise of the samurai style of life, warriors with a love of scholarship and a delight in waka poetry appeared. One was Hōjō Sanetoki, who collected Japanese and Chinese books and founded a famous library, the Kanazawa Bunko, in the Shōmyō-ji (at what is now Kanazawa). Military epics became popular. The most famous that has come down to us is the anonymously written *Heike monogatari* (*The Tale of the Heike*). They were recited throughout the country by Buddhist troubadours called *biwa hōshi.* After the middle Kamakura period, as Buddhist pessimism grew fainter, various kinds of instruction manuals and family injunctions were composed, while collections of essays such as Yoshida Kenkō's *Tsurezure-gusa* (*Essays in Idleness*) also made their appearance. The new nationalistic fervour found expression in Kokan Shiren's *Genkō shakusho* (1332), a 30-volume history of Buddhism in Japan.

Impact of military on the arts

In the visual arts the carving of wooden images of famous monks flourished, and, after the middle of the Kamakura period, Chinese styles of the Sung dynasty also began to enter Kamakura wood carving. In painting, in addition to Buddhist themes, picture scrolls (*emaki-mono*) became popular, taking as their themes the history of temples and shrines, the biographies of founders of the sects, and military epics.

Decline of Kamakura society. In the later Kamakura period occurred the breakdown of family solidarity among the samurai (see below).

During the troubled state of society at the end of the period, feudal landlords, out of the necessity of defending their own lands, seem to have devoted themselves chiefly to the military arts and to have entrusted the running of agriculture to their household dependents. Moreover, lands were no longer divided among younger sons but were now kept entirely in the hands of the eldest son. Power thus became concentrated in the head of the household, to whom other family members were subordinated in a lord–vassal relationship.

At the same time, major economic changes began to undermine the position of the *bakufu* vassals. Yet, despite the social crises among the landholders, trade was flourishing. Coins came into increasing circulation, and city styles of living began to be imitated in the country. But the landowners were often unable to meet their expenditures from the income of their limited holdings. Therefore, they borrowed money at high rates of interest from rich moneylenders, and many were forced to surrender their holdings when unable to repay their loans. Thus the gap between rich and poor became marked among the shogun's vassals. In particular, the local military governors, who had the right to raise troops, progressively gained control over the resident landlords, establishing a lord–vassal relationship with them. Moreover, deputies sent out by the heads of families to oversee their distant landholdings often broke with the main family and became vassals of the military governors. This powerful new class of local magnates was called daimyo (domain lords) and soon began to challenge the authority of the Hōjō regents in the *bakufu.*

The Ashikaga, Sasaki, Shōni, and Shimazu families were among the most powerful of the new class. The *bakufu* began to totter, shaken by the disputes between the Hōjō family and the rival military governors. When the Andō family raised a revolt in Mutsu Province at the end of the Kamakura period, the government found it difficult to suppress, partly because of the remoteness of the site of the uprising. In addition, regional unions of small landlords developed in the Kinai (the five home provinces centred around Kyōto).

Taking advantage of the accumulating weaknesses of the *bakufu,* a movement arose among the civil aristocracy to regain political power from the military. The occasion was provided by the question of the Imperial succession. In the mid-13th century there emerged two competing lines for the succession—the senior line centred on the Jimyō-in in Kyōto and the junior line centred on the Daikaku-ji on the western edge of the city. In the last half of the century, each side sought to win the support of the *bakufu.* In 1317 the *bakufu* proposed that the two lines serve by turns as emperor, but the dispute did not cease. Finally, in 1318 Prince Takaharu of the junior line acceded to the throne as the emperor Go-Daigo.

Dispute over the succession

The Muromachi (or Ashikaga) shogunate (1338–1573). *The Kemmu Restoration and the dual dynasties.* On the accession of Go-Daigo, the retired emperor Go-Uda broke the long-established custom and dissolved the "cloistered" Imperial government. As a result, the entire authority of the Imperial government was concentrated in the hands of a single emperor, Go-Daigo. A party of young reforming court nobles gathered around the Emperor, who revived the Kirokusho (Records Office) and strove to renovate the government. But to realize his ideal of a true Imperial restoration, it was necessary for Go-Daigo to rid himself of the interference of the *bakufu.* His plans for its overthrow were discovered, however, and he was arrested and exiled to Oki Island. But in the Kinai area, local leaders, supported by militant Buddhist monks, raised an army to overthrow the *bakufu.* The Imperial forces were led by Prince Morinaga and Kusunoki Masashige, but the decisive victory was brought about by the powerful Kantō families of Ashikaga Takauji and Nitta Yoshisada, discontented vassals of the Hōjō family. In 1333 Takauji successfully attacked the Hōjō headquarters and forced the *bakufu* leaders to commit suicide. Yoshisada meanwhile conquered the *bakufu* in Kamakura. Thus, after 140 years' rule, the *bakufu* government was brought to an end.

Overthrow of the Kamakura *bakufu*

The return of Go-Daigo to Kyōto in 1333 is known as the Kemmu Restoration (Kemmu no Chūkō). The Emperor immediately set about to restore direct Imperial rule. He abolished the powerful office of *kampaku* and set up a central bureaucracy. He established the Kirokusho to settle lawsuits in the provinces; the Zassho Ketsudansho (Court of Miscellaneous Claims) to handle minor suits; and a *musha-dokoro* (guard station) to keep order among the warriors in Kyōto. He placed Prince Morinaga in charge of his military forces and set up members of the Imperial family as provincial leaders in the north and east.

Those local warriors, however, who had joined the Imperial forces in the overthrow of the *bakufu* were disappointed in the division of the spoils. A rebel army formed under the leadership of Ashikaga Takauji, and in 1336 it drove the Emperor from Kyōto. Takauji set up an emperor from the senior Imperial line (the Jimyō-in), while Go-Daigo and his followers set up a rival court in the Yoshino Mountains near Nara. For the next 60 years political power was divided between the Southern Court in Yoshino and the Northern Court in Kyōto. It remained for Takauji's grandson Yoshimitsu to establish peace (1392) between the two courts; thereafter, the Imperial succession remained with the descendants of the Northern Court. Throughout the long dispute, however, local warriors attached themselves to the military governors, who increasingly asserted their independence from the declining central authority.

Division of Northern and Southern courts

The establishment of the Muromachi bakufu. After the withdrawal of Go-Daigo to Yoshino, Ashikaga Takauji set up a *bakufu* at Nijō Takakura in Kyōto. But in 1378 Takauji's grandson, the shogun Yoshimitsu, moved the *bakufu* to the Muromachi district in Kyōto, where it remained and took final shape. Yoshimitsu, assisted by the successive *kanrei* (deputy shogun) Hosokawa Yoriyuki and Shiba Yoshimasa, gradually overcame the great military governors. He destroyed the Yamana family in 1391, and in 1399, with his power further enhanced by his success in uniting the Northern and Southern courts, he attacked and destroyed the great military governor Ōuchi Yoshihiro, thus gaining control of the Inland Sea. Yoshimitsu was now raised to the highest office of grand minister of state, or *dajō-daijin.* He constructed the famed Golden Pavilion (Kinkaku-ji; see below *The establishment of military culture*) at his country seat in Kitayama, taking great pride in its luxurious display; he also carried on trade and diplomacy with Ming dynasty China under the title of king of Japan.

The Muromachi *bakufu* inherited almost unchanged the structure of its Kamakura predecessor (see above), setting up a Mandokoro, Monchūjo, and Samurai-dokoro. But after the appointment of Hosokawa Yoriyuki as *kanrei* (deputy shogun), this post became the most important in the *bakufu* government. The official business of the Mandokoro was to control the finances of the *bakufu;* and later the Ise family, who were hereditary retainers of the Ashikaga, came to inherit this office. The Samurai-dokoro, besides handling legal judgments, was entrusted with the control of the capital. Leading officials called *shoshi* who held the additional post of military governor of Yamashiro Province (Kyōto Prefecture) were next in importance to the *kanrei.* In local administration, a special administrator was set up in Kamakura to control the 10 provinces of the Kantō area. This office came to be held by heads of the Ashikaga Motouji family. The 11 provinces of Kyushu were controlled by an office known as the Kyushu *tandai.*

Structure of the Muromachi *bakufu*

The finances of the Muromachi *bakufu* could not be met simply from its receipts from the lands under its direct control. So, according to their needs, the military governors and stewards of each province were ordered to levy a money tax, on either every unit of land or every household, but this was not fully effective. As a result, taxes were levied from such dealers as pawnbrokers, and sake brewers, who were among the wealthiest merchants of the time. Financial deficiencies were also supplemented by engaging in trade with China. But the foundations of this *bakufu* began to be shaken by the increasing power of the military governors and by the frequent uprisings of local samurai and farmers.

In the Kamakura period, the authority of the military governors was limited to security matters. In the latter half of the Northern and Southern courts period, their executive power over the areas under their control was increased. As disturbances increased, they gained wide powers of military command. Sometimes the private estates were made depots for military supplies on the pretext of protecting them from the depredations of the samurai, and half their yearly taxes were given to the warriors. This was called the equal tax division, or *hanzei.* Many military governors succeeded to their domains by inheritance, and in cases such as that of the Yamana family a single military governor sometimes held a number of provinces. Thus arose a new class of official known as the *shugo* daimyo.

From the outset, the controlling power of the Ashikaga *bakufu* was weak, and, especially after the death of Yoshimitsu, the tendency for powerful military governors to defect became marked. Hence, as time passed the office of shogun became increasingly impotent.

The rise of commerce and manufacturing

In the villages around Kyōto, the status of farmers rose markedly as agriculture became more highly developed, and commerce and small-scale manufacturing prospered. Also, confederations of the middle and small landlords, or *myōshu,* proceeded apace and often resulted in uprisings. Such confederations appeared where farming by the greater *myōshu* had dissolved and middle and small *myōshu* had established themselves on a wide scale. These smaller landlords endeavoured to defend themselves against the ravages of local warfare, and they formed unions to manage the forests in common and to maintain irrigation works. In such confederations, a leader called the elder, or *otona,* would be selected to carry on village government. Assemblies were held regularly among its members at the village shrine or temple, and regulations were drawn up for the maintenance of the community life.

As self-government became strong in the communities, the resistance of farmers became fierce. After the northern and southern courts dispute, armed uprisings arose among the farming villages, demanding reductions in yearly taxes against the old proprietors and a moratorium on debts owed to the moneylenders. A large-scale uprising of this kind took place in 1428 in the last years of Yoshimitsu's reign. In 1429 an uprising broke out in Harima Province (present Hyōgō Prefecture) aimed at the expulsion of the warriors from the province. In 1441 farmers living around Kyōto attacked the pawnbrokers and demanded a moratorium on debts from the *bakufu.* Thereafter, uprisings occurred on a greater or lesser scale almost yearly—testimony to the fading power of the *bakufu.*

The *kangō* trade

Trade between China and Japan. Trade with Ming dynasty China began after the suppression of Japanese piracy. Ashikaga Takauji had sent ships of the Tenryū-ji to trade with the Yüan dynasty. But trade then ceased on account of the internal disturbances, and pirates from the maritime districts of west Japan raided the Korean peninsula and the continental mainland. When Korea came under the control of the Yi dynasty and when the Ming dynasty emerged in China, it requested the *bakufu* to open formal trade relations with the aim of suppressing piracy. Yoshimitsu, both in response to the desires of the merchants and in order to supplement the finances of the *bakufu,* began formal trade relations with Ming China and Korea, repatriating a large number of Chinese who had been taken captive by the pirates. In response, the Ming also began to trade, but under the form of tribute from Yoshimitsu, "King of Japan," to the Emperor of China. In this trade, in order to distinguish between pirate ships and trading ships, seals received from the Ming called *kangōfu* were used. Hence the use of the term *kangō* trade.

The profits of this trade were important to the *bakufu,* but later the control of this trade came into the hands of the Hosokawa and Ōuchi families, under whose protection trading merchants became active in Hakata, Hyōgo, and Sakai. After the Ōnin War (see below), the Ōuchi controlled this trade, but on their destruction the *kangō* trade ceased and piracy again became rife. Trade with Yi dynasty Korea was carried on through the agency of the Sō family of Tsushima, and various domain lords and the merchants of Hakata were actively involved in it, importing cotton and other goods. Japanese traders resided in Pusan and elsewhere in Korea. Also included in the trade with China and Korea were goods imported by Japanese merchants from the Ryukyu Islands, lying between Japan and Taiwan, and dye materials, pepper, and other special products from the South Seas.

Edicts for cancellation of debts

The Ōnin War (1467–77). In the reign of the shogun Ashikaga Yoshimasa a general civil war broke out in the area around Kyōto, caused by economic distress and by a dispute over the Imperial succession. Indeed, severe famines engendered rebellion nearly every autumn, and it is said that during his term as shogun Yoshimasa issued 13 edicts for the cancellation of debts known as *tokuseirei,* or "acts of grace." Lacking children of his own, Yoshimasa at first proposed that his younger brother should succeed him. But when he later fathered a child a quarrel arose over the succession for control of the family. The two chief administrators, Shiba and Hatakeyama, and the great military governors also took sides in the power dispute, with Kosokawa Katsumoto and Yamana Sōzen (also known as Yamana Mochitoyo) at the head. In 1467, the first year of the Ōnin era, fighting broke out between the eastern army of the Hosokawa party and the western army of the Yamana party. The eastern army had the advantage of the support of both the Emperor and the shogun, but the western army, assisted by the Ōuchi family, recovered its power, and fighting continued for 11 years, centred on Kyōto. Destruction around Kyōto was severe, and many large temples and residences were burned. After 11 years of warfare, the fighting spread to the provinces. As a result, the farming villages held conferences and quite frequently mounted armed uprisings in self-defense. The leaders of these armed uprisings were local samurai with roots in the farming villages. Such men frequently established themselves as domain lords (daimyo) during the war disturbances. They formed associations and often mounted uprisings that extended over a whole province and challenged the great military governors. In the autumn of 1485, 36 representatives of the local warriors of southern Yamashiro Province met in the Byōdō-in (Byōdō Temple) at Uji and successfully demanded the withdrawal of the two Hatakeyama armies. As a result, the southern Yamashiro area became self-governing for more than eight years.

In these wars, the civil aristocracy and priests lost the income from their private estates. Many of them left the capital, moving to Sakai or Nara or even taking up residence in the castle towns under the protection of local domain lords. This migration assisted in the diffusion of the central culture to the localities. Old traditions were destroyed, but from the ashes a new culture was born.

The shogun Ashikaga Yoshimasa, for example, finally turned his back on a troubled world and set up a country residence in the Higashiyama ward of Kyōto, where he lived in elegance and refinement, paying no attention to matters of government. The political power of the *bakufu* thus became virtually extinct, and real power came into the hands of the chief administrators of the Hosokawa (1490–1558) family. In the 16th century this power then came into the hands of their retainers, the Miyoshi (1558–65) family, until it was finally usurped by their own retainers, the Matsunaga (1565–68) family.

The period of the "Warring Country." *The emergence of new forces.* After the Ōnin War, the power of local leaders became increasingly strong, and there were many instances in which the deputies of military governors usurped the domains of their superiors, retainers overthrew their overlords, and branch families seized power from main families. Because of this tendency for "inferiors to overcome superiors" (*gekokujō*), the previous military governors almost completely disappeared from Kyōto and the surrounding provinces; a new type of domain lord, called *sengoku* ("warring country") daimyo, took their place.

Until the first half of the 16th century, domain lords in the various localities were thus building up strong military bases. During this period, the provinces held by the domain lords were almost completely free of *bakufu* control. The domain lords included the local leaders among their retainers, taking away their independence by enforcing land surveys and directly controlling the farming villages. Domain lords such as the Imagawa, Date, and Ōuchi issued their own laws called *bunkoku-hō.* These provincial laws, while drawing on the samurai laws of the Jōei Formulary and thereafter, also included regulations for farmers, and they applied strict controls over retainers. It was made a principle that inheritance by retainers should be restricted to the main heir alone, and the overlord's permission was necessary for his vassals to inherit property or to marry. In the farming villages the domain lords, in addition to carrying out land surveys, built irrigation dikes and opened new rice fields. In order to concentrate their power they also readjusted the disposition of local fortified strongholds, gathered their retainers into castles, and reorganized roads and post stations to centre on their castle towns.

Growth of cities and trade

Commerce and towns made marked development at this time in Japan's history. Markets also came to be held six times a month and were set up all over the country. Despite the obstructions of the customs barriers, the products of all the districts were available in these markets. In large cities such as Kyōto, commodity exchange markets were set up to handle huge quantities of rice, salt, fish, and other goods; wholesalers, or *toiya,* specialized in dealings with distant areas. The circulation of coined money also became vigorous, but in addition to the various kinds of copper coin imported from China of the Sung, Yüan, and Ming dynasties, privately minted coins also circulated within the country, giving rise to confusion of exchange rates. The *bakufu* and domain lords issued laws to prohibit people from hoarding good coins but without effective results. The guilds now showed a strong monopolistic tendency in order to protect themselves against new-style merchants who emerged while new guilds were set up in the castle towns (*jōka-machi*) under the direct control of the domain lords. Among the cities of the time, next to Kyōto and Nara, Uji and Yamada outside the gates of Ise-jingū (Ise Shrines) flourished. Besides these, towns grew up around the castles of the domain lords, such as Naoetsu of the Uesugi family, Yamaguchi of the Ōuchi family, Ichijōdani of the Asakura family, and Odawara of the later Hōjō. As the castles of the domain lords were moved from mountain fortresses to strongholds in the plains, markets were opened outside the castle walls, and merchants and artisans gathered there to live. Harbour towns (*minato machi*) such as Sakai, Hyōgo, and Onomichi on the Inland Sea, Suruga and Obama on the Japan Sea, and Kuwana and Ōminato on Ise Bay also flourished as exchange centres. Sake brewers, brokers, and wholesale merchants were leading townsmen (*machi shu*), and town elders, called *otona,* were chosen to carry on local government through assemblies. In the trading port of Sakai, for example, an assembly of 36 men drawn from the wholesale guilds carried on the city government. They maintained soldiers and constructed moats and other defenses, and while profiting from the confrontation of the domain lords, they resisted their domination. The Jesuit missionaries (see below) compared Sakai to the free cities of Europe in the Middle Ages and described its flourishing condition in their reports.

The arrival of the Europeans. As the warring domain lords carved out their territories, the central authority ceased to maintain control over overseas trade. Further, Japanese marauders in association with Chinese pirates again became active. At this time, the Spanish and Portuguese made their appearance in the archipelago. In 1543 (according to Japanese sources), the first Portuguese were shipwrecked on the island of Tanega, off southern Kyushu. These were the first Europeans to arrive in Japan, and the art of musket construction they passed on at this time immediately spread to Sakai and other places.

The Jesuit missionaries

In 1549 the Jesuit missionary Francis Xavier arrived in Kagoshima. After missionary work for two years and three months, he left Japan, and thereafter Jesuit missionaries arrived continuously. The missionaries made use of the trade from the Portuguese ships to propagate Christianity, and there were cases in which merchant ships would not enter the ports of domain lords who did not show good will toward missionary activity. Thus, domain lords, seeking the profits of foreign trade and the acquisition of military equipment and supplies, progressively protected Christianity, until finally some domain lords even became Christian converts. Three Kyushu Christian lords—Ōtomo Sōrin, Arima Harunobu, and Ōmura Sumitada—sent an embassy to Rome. Farmers also increasingly became converts under the influence of the social relief work and medical aid that accompanied missionary activity.

The establishment of military culture. While absorbing the traditional culture of the civil aristocracy, the military families that established themselves in Kyōto also introduced the continental culture of the Sung, Yüan, and Ming dynasties, especially the culture associated with Zen Buddhism, thus fashioning a new military family culture. This began with the golden age of the shogun Ashikaga Yoshimitsu at the end of the 14th century. In this period, scholarship and the arts flourished in the five Zen monasteries of Kyōto under the patronage of the shogun. Renga (linked verse) and Nō drama became vigorous. The essence of this culture finds concentrated expression in the Golden Pavilion in Yoshimitsu's country estate at Kitayama. Destroyed by an arsonist in 1950 and rebuilt in 1955, it is now officially the Rokuon-ji (Rokuon Temple) in northwestern Kyōto. Facing a garden of refined elegance, the Golden Pavilion is built in the Japanese *shinden* style (an ecclesiastical style with Zen influence) in its first and second stories, while its upper story is in the *kara* ("Chinese") style of the Zen school. Thus Kitayama culture, while absorbing Zen influences from China, shows many influences of the native aristocratic culture. In the time of the shogun Yoshimasa, the samurai culture, in addition to even deeper Zen taste, shows a refined appreciation of simplicity and quiet profundity. The Silver Pavilion (Ginkaku-ji) and its garden built by Yoshimasa on his country estate at Higashiyama (now part of the Jishō-ji) displays the essence of this polished Higashiyama culture. While adopted by the local domain lords, Higashiyama culture also gave rise to a new culture centred on the townsmen of Kyōto and Sakai, and it is the forerunner of the Azuchi-Momoyama and Edo cultures.

Increasing temporal role of the monasteries

In Buddhism, the ancient great temples like the Enryaku-ji became mere shadows of their former greatness with the diminution of their landed estates. Since the Kamakura period, the Rinzai sect had been linked to the upper military families. The Muromachi shogunal family (the Ashikaga) gave special protection to the group of the priest Musō Soseki of this sect, which flourished in the Gozan monasteries (the five most important Zen monasteries) in Kyōto. The monks of the Gozan became advisers to the *bakufu* in government, diplomacy, and culture; they studied the Neo-Confucian philosophy of Chu Hsi that came from China along with Zen, published books, and wrote poetry and prose in the Chinese style. But the five monasteries became vulgarized because of their excessive links with the political world, and they ceased to prosper as the *bakufu* declined. In contrast, the Myōshin-ji and Daitoku-ji groups arose, and the latter is famous for the monk Ikkyū, who propagated his own teaching. At this time Rennyo (1415–99) of the Shin (True) sect of Jōdo (Pure Land) Buddhism came forth from the Hongan-ji in Kyōto, teaching his principles in simple phrases and spreading the faith by organizing groups called *kō.* He came under persecution from the Enryaku-ji, however, and he fled to Echizen Province, establishing a school of instruction at Yoshizaki. He then moved to Settsu Province, where at Ishiyama (now a part of Ōsaka) the Hongan-ji achieved its golden age. While also persecuted, the Hokke (Lotus) sect progressively gained adherents among the warriors and merchants. At this time, the custom of pilgrimages to the holy places of the Buddhist deity Kannon, to the Shintō shrines at Ise, and to the summit of Mt. Fuji also became popular. Within this trend, a worldly Shintō belief arose, and in the 15th century the scholar Yoshida Kanetomo, while proclaiming Shintō principles, also took the occasion to free Shintō shrines from Buddhist control; he believed that only a deep religious faith could cure the people of their despondency. In the arts the Nō drama developed in the Kamakura period under the influence of agricultural festival dances, and guilds (*za*) were formed to serve at the ceremonies of temples and shrines and at funeral services. From among the four guilds attached to the Kōfuku-ji and the Kasuga Shrine of Yamato Province (present Nara Prefecture), the father and son Kan'ami and Zeami Motokiyo appeared; under the patronage of the shogun Yoshimitsu, they laid the foundations for a flourishing Nō drama, establishing the guidelines for performance and bequeathing many texts. *Kyōgen* (dialogue plays with dance), which developed from the comic elements of an older form of entertainment called *sarugaku,* were performed in the intervals of Nō drama; taking their topics from the everyday life of the common people, *kyōgen* were widely appreciated by them. Traditional Japanese waka verse was also composed, but renga (linked verse) became ever more popular and was enjoyed by the warriors and the common people. After a time, however, renga became overly formal, as the waka did, and lost its freshness; hence, the free-style verse called haikai was born.

Developments in architecture and the visual arts

Along with the prosperity of Zen, the *shoin* style of building closely connected with this school was widely adopted by the military families and civil aristocrats in the construction of their residences, becoming the foundation of present-day domestic architecture. The *shoin* was originally a room in which monks read the Buddhist scriptures. In constructing the *shoin,* an entrance called a *genkan* is built, while within the room straw mats called *tatami* are laid out, paper-covered sliding partitions are used, and an alcove and shelves at different levels are set up. The custom of hanging a monochrome painting in the alcove and placing flowers or an incense bowl before it also arose at this time. In gardens, a delight was first taken in adding the Zen mood of retreat from the world to the *shinden* style, making symbolic use of streams, flowers, and bushes. Later, even more symbolic gardens were constructed using arrangements only of stones, raked sand, and gravel. The carving of images of the Buddha and the Buddhist paintings that had flourished in the Kamakura period now declined, as did the ancient sects themselves, and new ones arose. *Yamato-e* painting also declined, and the picture-scrolls lost their freshness. On the other hand, with the increase of interest in Zen, monochrome painting in the Sung and Yüan style was begun by the monks of the Gozan. In the time of Yoshimasa, the great painter Sesshū broke away from imitation of Chinese models and opened new frontiers in monochrome paintings. The father and son Kanō Masanobu and Kanō Motonobu introduced the gentle models of *Yamato-e* to monochrome painting and became the founders of the succeeding Kanō school. Tea drinking, introduced from Sung China by the Buddhist priest Eisai in the Kamakura period, spread to the warriors and the common people in the northern and southern courts period. In particular, in the time of the shogun Yoshimasa, Jukō came from Nara and began the *wabi-cha* form of tea ceremony by bringing together the *cha-no-yu* of the civil aristocracy and the *cha yoriai* of the common people. This new form spread among the warriors and great merchants and was further purified by the Sakai merchant Jōō. Together with the development of the tea ceremony, new forms were brought about in the construction of tearooms, flower arrangement, pottery, and Japanese cakes. Higashiyama culture became further diffused among the common people, and as the livelihood and education of the merchants and artisans of the cities reached higher levels they enjoyed Nō and *kyōgen* dramas, the tea ceremony, and renga. Fairy stories were also widely enjoyed, being easy to read, and included stories that had been related among the people since ancient times. These became popular not only among the children of the civil aristocracy and warriors but also among those of the townsmen who were educated in temples and shrines. The local domain lords also promoted culture within their domains, enhancing their dignity as lords by building temples and shrines in their castle towns and actively introducing the culture of Kyōto.

Thus, while warfare was rife in the Kamakura and Muromachi periods, these were, nevertheless, eras that gave Japan some of its most distinctive cultural institutions.

(T.T.)

THE EARLY MODERN HISTORY OF JAPAN (1550–1850)

Unification. *The Oda regime.* In the 1550–60 period the *sengoku* daimyo, who had survived the wars of the previous 100 years, moved into an even fiercer stage of mutual conflict. These powerful daimyo were harassed not only by their mutual conflicts but also by the social development of the common people within their domains. The lords sought to resolve these contradictions by acquiring land and people to widen their domains and, finally, by trying to grasp the leadership of the whole country. To accomplish this purpose, they considered it necessary to control Kyōto, the political centre since ancient times. In the midst of these struggles, one such *sengoku* daimyo, Oda Nobunaga of Owari Province in modern Aichi Prefecture, succeeded in entering the capital as the first feudal unifier.

The first feudal unifier

On the emergence of Oda's regime, the feudal disintegration of the previous century began to show a clear tendency toward unification. Oda's bold wars of suppression, which entitled him to be called a military genius, led to a great redrawing of the political map, previously split up among the domain lords of the whole country. In the Kinai district, on which Oda's conquered territory was centred, however, he established control by dividing his new domain among his commanders. Rather than completely abrogating the long-established privileges of the temples, shrines, and local landlords (*kokujin*), he at first recognized these, regarding them as an important adjunct to the strengthening of his military power and using them as followers in the unification wars. The land surveys aimed at strengthening feudal landownership were at this stage carried out not so much to gain control over the complicated landholding and taxation system of the farmers as to define the size of fiefs (*chigyō*) of his retainers in order to confirm the extent of their military services and obligations.

The unification policy of the Oda regime was upheld by the separation of the warriors from the farmers, but it could not be fully achieved because of resistance from old political forces. Unification became more clearly established later, during Toyotomi Hideyoshi's regime.

The Hideyoshi regime. Oda was the son of an Owari domain lord, whereas Hideyoshi was the son of a farmer from the same province. Entering Oda's service, Hideyoshi was greatly esteemed for his brilliant talents, and before Oda's death Hideyoshi had become one of his most powerful commanders. After Oda's death during the Honnō-ji (a monastery in Kyōto) Incident of 1582—when his retainer and commander, Akechi Mitsuhide, rebelled against him—Hideyoshi eliminated many rivals by his superb political judgment and shrewd actions, thus firmly establishing himself as successor. In the footsteps of Oda, Hideyoshi proceeded to unify the whole country at a rapid pace, and, in 1590, all Japan—from Kyushu in the southwest to Tōhoku in the northeast—had come under his control. On the pretext of giving rewards for distinguished service, Hideyoshi gave the Kantō domain, formerly of the Hōjō family, to Tokugawa Ieyasu, causing Ieyasu to move to Edo; this was, in fact, a stratagem to remove the Tokugawa family from the Chūbu district, where its power had been nourished.

The keynote of the Hideyoshi unification policy was its firm establishment in principle of the separation of the warriors and farmers. Thus, the clear-cut contrast between feudal landowners and feudal small farmers (or serfs) became the basic model of the system.

The Taikō land survey

The Taikō land survey played an important part in this process. *Taikō* was a traditional title for former *kampaku* (chancellors), and it was assumed by Hideyoshi in 1591. The Taikō land survey was carried out over the whole country from 1583 to 1598, just before Hideyoshi's death. As a result of this survey, the complicated connections of rights to landownership that had developed since the Kamakura period were set in order. Landowning relations were now based on *kokudaka—i.e.,* on the actual product of the land. Moreover, this actual product now came within the landlord's grasp in every village, and land taxes were now levied on the village as a unit. This is called the *mura-uke,* or village responsibility, system. In addition to this definition of the rights held by the farming population, the *kokudaka* system also applied to the landholdings of domain lords (the *chigyō*) to be distributed among their retainers. In place of previous land taxes (*nengu*) assessed in money as so many hundred or ten thousand *kan* of silver, an assessment of *kokudaka* was made as so many hundred or ten thousand *koku* of rice, a *koku* representing the amount of rice consumed by one person in one year; the amount was also used as a standard on which military services were levied in proportion.

During the Taikō land survey, a land-survey register was drawn up in every village, and farmers so registered were recognized in their rights as cultivators to the extent of the land thus registered; in return they were bound to pay land taxes in rice and forbidden to neglect the cultivation of their fields or to move elsewhere. Farmers were thus reduced to rural serfs, tied to the land, and exploited. The promulgation of an order of social-status control in 1591 prohibited people who neither cultivated the land nor performed military service (*i.e.,* the artisans and merchants) from residing in the villages, showing a further advance from the separation of the warriors and farmers to a feudal social class system of warriors, farmers, artisans, and merchants. The order of "sword hunt" (*katanagari*), which took away arms from the farmers, was also an important prerequisite for this policy. By establishing the *kokudaka* system, the Taikō land survey delivered the final blow to the *shōen* system, a system of manors in medieval times, which had already greatly declined. The feudal *chigyō* system, based on the *kokudaka* system, was established throughout the country, the domain lords all submitted to the despotic control of the Hideyoshi regime, and the alliance-like relationship between Oda and the former *sengoku* daimyo changed to a clear lord-vassal relationship.

The political structure of the Hideyoshi regime was not yet fully equipped, however, to be the unified governing authority controlling the whole country. For example, the *kurairechi,* or lands under its direct control, which were the immediate financial base of the regime, amounted to about 2,000,000 *koku.* But setting aside those of the metropolitan and surrounding provinces, these lands were in many cases divided among the distant, independent domain lords (*tozama* daimyo), and the management of these lands was entrusted to them. Such lands were thus not firmly in the grasp of the regime. By contrast, the lands that later came under the direct control of the Tokugawa shogunate amounted to more than 4,000,000 *koku,* or better than double those of the Hideyoshi regime, and 80 percent of them were managed by officials known as *gundai* and *daikan,* who were direct retainers of the shogunate, with only 20 percent deposited with domain lords. This inner contradiction in the political structure of the Hideyoshi regime gave rise to internal power struggles and finally drove Hideyoshi to such reckless actions as the Korean expeditions (aggressions against Korea in 1592 and in 1597). The Hideyoshi regime collapsed on the failure of that expedition and as the direct result of Hideyoshi's subsequent death. Tokugawa Ieyasu was the strongest candidate to form the next regime.

The Matsudaira family

The bakuhan system. *The establishment of the system.* The ancestors of Tokugawa Ieyasu, the founder of the Edo *bakufu,* were the Matsudaira, a *sengoku* daimyo family from the mountainous region of Mikawa Province (in present Aichi Prefecture). The Matsudaira family had built up their base as domain lords by advancing into the plains of Mikawa. But when they were attacked by the powerful domain lords of the Oda family from the west, they were defeated, and Ieyasu's father, Hirotada, was killed. Ieyasu had earlier been sent to the Imagawa family as a hostage to cement an alliance but had been captured enroute by the Oda family. After his father's death he finally was sent to the Imagawa family and spent 12 years under detention. In 1560 Imagawa Yoshimoto was killed by Oda Nobunaga in the Battle of Okehazama, in which Nobunaga destroyed the Imagawa family and confirmed his course of unification, and Ieyasu was finally released. He returned to Okazaki in Mikawa and brought this province under his control. As Oda's ally, he guarded the rear for the advance on Kyōto, and he thereafter fought his own military campaigns, advancing eastward. By 1582 he was a powerful daimyo, possessing, in addition to his home province of Mikawa, the four provinces of Suruga and Tōtōmi (modern Shizuoka Prefecture), Kai (Yamanashi Prefecture), and southern Shinano (Nagano Prefecture).

When Hideyoshi seized the ruling power, Ieyasu at first opposed him. But he then submitted, and, rising to be the most powerful domain lord among Hideyoshi's vassals, he became chief of the five *tairō,* the highest officers of the Hideyoshi regime. After Hideyoshi's death Ieyasu won the Battle of Sekigahara in 1600, in which all daimyo in the country took part, establishing his national supremacy. In 1603 Ieyasu set up the Edo *bakufu* (more commonly known as the Tokugawa shogunate [1603–1867]) to legalize this position. Control over the domain lords was firmly exercised at this time. On the pretext of allotting rewards after Sekigahara, he dispossessed, reduced, or transferred a large number of opposing daimyo and gave the confiscated lands either to relatives and retainers of the Tokugawa family to establish them as domain lords and to increase their holdings or reserved the lands as the family's domains. In addition, Hideyoshi's son Hideyori was reduced to the position of a domain lord of the Kinki district. Two years after the establishment of the *bakufu,* Ieyasu relinquished the office of shogun to his son Hidetada, retiring to Sumpu (modern city of Shizuoka) to devote himself to strengthening the foundations of the *bakufu.* In 1615 Ieyasu stormed and captured Ōsaka Castle, destroying the Toyotomi family. Immediately afterward, the Buke Shohatto (Laws for the Military Houses) and the Kinchū Narabi ni Kuge Shohatto (Laws for the Imperial and Court Officials) were promulgated as the legal basis for *bakufu* control of the domain lords and the Imperial court. In 1616 Ieyasu died.

Completion of the *bakuhan* system

Under the second and third shoguns, Hidetada and his successor, Iemitsu, the *bakufu* control policy advanced further until the *bakuhan* system—the government system of the Tokugawa shogunate; literally a combination of *bakufu* and *han* (the domain of a daimyo)—reached its completion. By reorganizations in 1633–42 the executive of the *bakufu* government was almost completed, as represented by the offices of senior councillors (*rōjū*), junior councillors (*waka doshiyori*), and three commissioners (*bugyō*) for the temples and shrines of the country, the shogun's capital, and the treasury of the *bakufu.* Confiscations and reductions were continuously made against the domains, and wide-scale transfers also took place, distributing the strategic districts of Kantō, Kinki, and Tōkaidō among the relatives and retainers of the *bakufu,* thus keeping the "outside" (*tozama*) domain lords in check. Along with the rearrangement of the daimyo, the lands under the direct control of the *bakufu* were also increased at key points throughout the country. The most important cities and mines were also placed under direct *bakufu* control and used to control commerce, industry, and trade.

The *bakufu* also revised the Laws for the Military Houses and systematized the *sankin kōtai* (alternative attendance), by which the domain lords were required to pay ceremonial visits to Edo every other year, while their wives and children resided permanently in Edo as hostages. The *sankin kōtai* system was unique to Japanese feudalism and never appeared in European feudalism. In addition, the daimyo were forced to assist in such public works as the construction of castles in the *bakufu* domains, thus being driven into financial difficulties. The *bakufu* domains now amounted to more than 7,000,000 *koku*—about one-fourth of the whole country. Of these lands, more than 4,000,000 *koku* were under its direct control, and 3,000,000 *koku* were distributed among the *hatamoto* and *gokenin,* the liege vassals to the *bakufu.* In addition, because the *bakufu* had a monopoly of foreign trade and alone had the right to issue currency, it had considerably greater financial resources than did the domain lords. In military strength, it was also far more powerful than any individual daimyo.

In step with the structural organization of the *bakufu* as the supreme power, the domain governments of the daimyo also progressively took firm shape. The relationship between the shogun and the daimyo was linked with the lord-vassal relationship, based on the feudal *chigyō* system. The land of the whole country belonged to the shogun, who divided this among the domain lords as a special favour, or *go-on.* In return, the domain lords had a duty to provide military and other services to the shogun. This same connection existed between the domain lords and their retainers, and in order to concentrate and strengthen the ruling power of the domain lords, it was necessary for them to tighten this connection. Applying restrictions to the traditional right of the domain warriors to *chigyō,* or subdomains, retainers were supplied with rice stipends (*kuramai*) in place of *chigyō,* thus increasing their dependence on the domain lord. At the same time, this increased the lands under the direct control of the daimyo, strengthening the economic base of the domain. These were the same methods employed by the *bakufu.* In this way, a hierarchical, feudal *chigyō* system was established by means of the *koku* system, which extended from the shogun and the domain lords to their retainers. This was the fundamental condition that made possible the concentration of ruling power in the hands of superiors, the principal distinguishing feature of the Japanese feudal system.

Having been inherited from the Oda–Hideyoshi regimes, the control of the farmers, which was the main object of feudal rule, was now further strengthened. The Taikō land survey had recognized the rights of the lower-class farmers as the actual cultivators of the land and made them bear the responsibility for the payment of land taxes. While also carried out on this basis, the land surveys of the *bakufu* and the daimyo domains were much more detailed and precise in standards of survey, adopting the policy of extracting the greatest possible tax yield, limited only by the necessity for the farmers to continue to produce. The Tokugawa villages thus differed from those of the preceding ages, which had been controlled by patriarchal landlords, or *myōshu.* The Tokugawa villages were composed of a main core of feudalistic petty farmers, who went under the general title of *hombyakushō* ("farmers proper"). At the same time, the village became an administrative unit of the new feudal regime; and to carry out its functions, a system of village officers was organized in three grades—*nanushi* (or *shōya*), *kumigashira,* and *hyakushōdai.* The inhabitants of the towns and villages of the whole country were required to form *gonin-gumi* ("five-household groups"), or neighbourhood associations, to foster joint responsibility for tax payment, to prevent offenses against the laws of their overlords, to provide one another with mutual assistance, and to keep a general watch on one another. The so-called outer economic controls over farmers were further strengthened: farmers were strictly prohibited from buying and selling land, from abandoning their land, or from changing their occupation; minute restrictions were also applied to keep their clothes, food, and houses as simple as possible. The Keian no Ofuregaki ("Proclamations of the Keian era") was promulgated by the *bakufu* in 1649, a compendium of *bakufu* policies toward rural administration.

Control of the villages

The enforcement of national seclusion. The 1630s also marked an important dividing line in foreign relations with the enforcement of *sakoku,* or national seclusion. The path toward seclusion had been prepared during the formation process of the *bakuhan* system. The seeds of this policy had been sown in trade control and in measures against Christianity by the Oda-Hideyoshi regimes. Driven on by his consciousness of Japan as a "divine country" in his position as feudal overlord, Hideyoshi, though strongly attracted to trade as a source of national wealth and military strength, had issued an order for the exclusion of the missionaries. Ieyasu, even more strongly attracted by profits, made efforts to trade not only with the Portuguese Catholics but also with Protestant Holland and England, protecting trade with the southern regions by granting special licenses, or *shuin-jō* ("red-seal license"), to oceangoing merchant ships. But Ieyasu's encouragement of trade was above all aimed at establishing a *bakufu* trade monopoly. In 1604 a special system for the purchase of silk was established, and the Chinese silk imported to Japan by Portuguese ships was sold at fixed prices to the powerful merchants of Kyōto, Sakai, and Nagasaki, who formed a guild and then distributed this silk to the domestic retail merchants. Moreover, under the name of "emperor's silk" (*go-yō ito*), Ieyasu enjoyed a preferential purchase of a part of the imported silk prior to the said guild's treatment and reaped a huge profit on releasing this to the domestic markets.

As a result of his eagerness for trade, it was natural for Ieyasu to be generous to the propagation of Christianity. But Ieyasu feared that the Christians would link up with the Toyotomi family to resist the *bakufu,* and he took steps to prohibit Christianity before his destruction of the Toyotomi family. Decrees prohibiting Christianity were promulgated in 1612 and 1614, and the persecution of its adherents began immediately thereafter. This persecution became much more severe during the reigns of Hidetada and Iemitsu, until finally it became official policy to stamp out Christianity even at the sacrifice of trade. This policy became manifest with the seclusion orders of the 1630s. Thus, in 1635 Japanese were forbidden to make overseas voyages or to return to Japan from overseas, which was a severe blow to traders.

In 1637, in resistance to heavy taxes and the prohibition of Christianity, an uprising took place in the Shimabara Peninsula of Kyushu, consisting of farmers led by masterless Christian samurai. For five months they put up a fierce fight against the *bakufu* army. Having suffered this bitter experience, known as the Shimabara Rebellion, the *bakufu* thereafter stepped up its strict controls on Christians and attempted to root them out by such means as *fumi-e,* in which one was required to trample on an image of Christ or the Virgin Mary. The system of registration at Buddhist temples, or *teraukeseido,* was begun, and all the inhabitants of the country had to belong as parishioners to a parent Buddhist temple, called a *danna-dera* ("family temple"), which every year had to guarantee that the parishioner was not a Christian. In 1639, when Portuguese ships were forbidden to visit Japan, the seclusion orders were fully effective both in name and reality. The Dutch and the Chinese were allowed to trade as before, but this trade was restricted and was confined to the island of Dejima at Nagasaki. Iemitsu also allowed a certain amount of trade with Korea and the Ryukyu Islands.

The Shimabara Rebellion

Various views exist as to the influence of national seclusion on the Japanese, but the depth of its impact was probably without parallel in Japanese history. It is an undeniable fact that the vigorous desire of the Japanese to expand overseas prior to the seclusion policy was thenceforth diverted into a negative national character known as *shimaguni konjō,* or insularity. While the seclusion policy was useful in enabling the Tokugawa shogunate to exercise an enduring and stable political dominance for nearly 300 years, this simply resulted in the long continuation of a rigid feudal system to an extent unknown elsewhere in the world. Industry developed and gave rise to a unique popular culture, but it was a limited Japanese feudal culture with no international characteristics.

Status distinction system. Thus, the *bakuhan* system was firmly established in the second half of the 17th century, as determined by domestic and foreign conditions. This establishment of a feudal class structure of warriors, farmers, artisans, and merchants (*shi-nō-kō-shō*) was truly the final consummation of the *bakuhan* system. Distinctions between the statuses of warriors, farmers, artisans, and merchants were strictly enforced, but the distinction between the warriors and the other three statuses was especially strict. Forming barely 7 percent of a total population estimated at 30,000,000, the warriors levied taxes on the farmers, who formed more than 80 percent of the population and who thus provided the economic foundation of the system. To symbolize this society, the warriors wore swords in everyday life, because the system was maintained by their great military power. While peace lasted for some 250 years, it was in fact no more than an armed peace. Another special feature of this society was that, even in family relationships, absolute obedience was demanded from members of the family toward the house head (*kachō*). Among the family members, the status of women was especially low, and the idea of *danson-johi* (respect for the male, contempt for the female) was prevalent.

Relations between superiors and inferiors

The establishment of the *bakuhan* system created a need for a new world view and a system of ethics to support it. In these circumstances, the Shintō and Buddhist ideologies of early medieval society were useless. Only the morals of Confucianism, especially the Chu Hsi studies, could provide an intellectual rationalization for the status-oriented social structure of the *bakuhan* system, with its stratified feudal landowning structure headed by the shogun and its attribution of superior status to warriors over the other social classes. The feudal rulers, in particular, demanded Shushigaku ("Chu Hsi school") because, among the various schools of Confucianism, it was the most systematic as a doctrinal system. The Chu Hsi scholar Hayashi Razan was employed by the *bakufu* and was well suited to the *bakufu* foundation period. He is said to have had a hand in the drafting of all *bakufu* official documents and in the formulation of *bakufu* laws. His political ideas, as seen in his *Honchō hennen-roku* ("Chronological History of Japan") and in the *Honchō tsugan* ("Survey History of Japan"), compiled by his son Gahō, provided a historical justification for the establishment of the Tokugawa shogunate. The role of Chu Hsi studies was to repudiate the revolutionary idea of *gekokujō,* or the overthrow of superiors by inferiors, and to emphasize the idea of *kenshin,* linking this to Confucian moral conceptions. The central moral conceptions of Confucianism were *chū,* or "loyalty," and *kō,* or "filial piety." But in contrast to China, more emphasis was placed on *chū* as a support for feudal lord-vassal relations than on *kō,* which was a family ethic. Chu Hsi studies opposed the world view and logic of Christianity, which gave more importance to God than to the ruler-subject relationship and which also bitterly criticized Buddhism, which had been the ideology of the Middle Ages.

Wang Yang-ming studies (Oyōmeigaku) also held a special place in Confucian circles in the early Edo period. These studies were characterized by a strong subjective idealism, but in other aspects they had practical elements. Nakae Tōju, said to be the father of Japanese Wang Yang-ming studies, was so earnest in performing virtuous acts that he was called the sage of Ōmi. One of his followers, Kumazawa Banzan, transformed Wang Yang-ming studies from a means for individual spiritual training to a method for providing the energy for political reformation.

Commercial economy

Industries, cities, and culture. With the establishment of the *bakuhan* system, the anarchy of the earlier period ended, bringing in its wake a tendency to domestic peace. As a result, industry was promoted and cities developed. This trend took place in the latter half of the 17th century, centred in the Kinki district, where productive power was the most advanced in the whole country. Feudal petty farmers, who went under the general name of *hombyakusho,* now became widespread, and, while paying heavy taxes and performing various kinds of labour services, they sought to enjoy a better standard of living, however slight. In addition to their efforts as cultivators, they reclaimed new lands and produced various commercial crops and handicraft goods for sale in the city markets. Representative of such commercial crops were cotton and rapeseed oil in the Kinki district and silk in east Japan. Communications and transportation also developed for the circulation of such goods. But this circulation of commercial goods among the people was centred on the side roads and waterways rather than on the five great highways constructed by the *bakufu.* As a result of the development of industry and communications, such new-style merchants as wholesalers and brokers came to the fore, and powerful financiers also appeared.

As for the cities, the castle towns of the domain lords were the most numerous, but these gradually came to acquire the character of commercial cities. Purely commercial cities and post towns (towns along highways) also arose throughout the country. The cities of Edo, Ōsaka, and Kyōto, under the direct control of the *bakufu,* were especially developed. When its warrior inhabitants are included, Edo in the early years of the 18th century had a population of more than 1,000,000 and thus became the largest city in the world.

Cultural activities

In this period remarkable artists and masterpieces appeared in the fine arts and crafts. Great artists representative of the culture include Ihara Saikaku in *ukiyo-zōshi* ("tales of the floating world") genre novels, Chikamatsu Monzaemon in *jōruri* ("puppet play") drama, and Matsuo Bashō in haiku. Saikaku was a townsman who spent all his life in Ōsaka. He first aspired to write haikai—humorous renga (linked-verse) poetry from which the more serious haiku was derived—and for more than 30 years he was active as a haikai composer. One of his fortes was in *yakazu-haikai,* which was a competition to compose as many haikai as possible within a fixed period of time. Saikaku set a new record by composing 23,500 haikai in a single day and night—one verse every four seconds. In 1685 Saikaku gave up haikai and next set out to write *ukiyo-zōshi,* producing about 20 masterpieces in succession, beginning with *Kōshoku ichidai otoko* (1682; *The Life of an Amorous Man,* 1964). Between 1624 and 1643 the *ukiyo-zōshi* novel had transformed the *kana-zōshi* (storybooks written in *kana*) into an even more thoroughly townsman form of literature after the latter had replaced the previous *otogi-zōshi* ("fairy-tale books") in popularity. The unique spirit of the age can be seen in the word *ukiyo,* which had meant "sad world" in medieval times. Written with a different calligraphic character in *bakuhan* times, it now meant "floating world" and implied pleasure. The consistent aim of Saikaku's works was to depict, as accurately as possible, human desire for love and gain, taking the realistic viewpoint that "human beings are bundles of desire equipped with arms and legs." His viewpoint also contained a keen criticism of the warriors as men so bound by social status and moral principles that they could not live a free life.

Matsuo Bashō became closely attached to haiku (although the word itself was not commonly used until the 19th century) and fashioned it into up-to-date popular poetry. Bashō came from a warrior family; but after becoming a masterless samurai, or *rōnin,* he devoted himself to the development of haiku as an artistic literary form, while suffering various hardships in his means of livelihood. Bashō found the existing haikai style unsatisfying. He began writing hokku (17-syllable opening verses for renga) as separate poems, developing a new style called *shōfū* or "Bashō style." Bashō proclaimed what he called *makotono* ("true") haiku, seeking the spirit of this poetic form in sincerity and truthfulness. He also brought a new beauty to haiku by the use of small and simple words. He brought about an artistic flowering of the highbrow conceptions of medieval poetry by grafting onto them the commonplace feelings of the Tokugawa people. Rather than repudiating tradition, Bashō's haiku brought it to completion.

During the period 1592–1614 the *Jōrurihime monogatari* (a romantic ballad), drawing on the traditions of the medieval narrative story, was for the first time arranged in the form of dramatic literature accompanied by puppetry and the samisen (a lutelike musical instrument). It continued to develop until the three masters—Takemoto Gidayū as narrator, Chikamatsu Monzaemon as composer, and Tatsumatsu Hachirobei as puppeteer—made *jōruri* the representative form of Tokugawa performing arts.

The idea of *giri*

Gidayū was the son of a farmer, while Chikamatsu, like Bashō, came from a warrior family. Chikamatsu wrote more than 80 *jidaimono,* or historical dramas, and 20 *sewamono,* or domestic dramas of contemporary townsman society, both for *jōruri.* He also wrote more than 30 Kabuki plays. The chief theme running through Chikamatsu's works is the idea of *giri* (or "duty"), which is to be understood not so much as feudal morality enforced from above but rather as the traditional consciousness of honour and dignity in one's motives and position and of social consciousness in human relations. The compositions of Chikamatsu's later years seek the motif of tragedy in the fact that this *giri,* while proof of men's humanity, cannot be thoroughly achieved because of their immorality and lack of principle. Beginning with his *Shinjū ten no Amijima* (1720; *The Love Suicide of Amijima,* 1953), the leading male and female characters in his *sewamono* dramas are unable to keep the postulates of *giri* in this world and so die by *shinjū* (a suicide pact between lovers) in order to realize their love in a future life. Buddhist elements can be seen in this conclusion, as well as the unresolvable contradictions that faced the townsman in Genroku period society (1688–1703).

Kabuki also established its foundations as Tokugawa drama in this period. The *okuni* Kabuki, named for the female dancer Izumo Okuni, which had been popular at the turn of the 17th century, afterward developed from the *yūjo* Kabuki, or courtesan style, to the *wakashū* Kabuki, or young-man style. The *wakashū* Kabuki was prohibited by the *bakufu* moral censors, however, because of widespread homosexuality among its performers, and it developed into the *yarō* Kabuki style, played by adult males distinguished from the *wakashū* by shaven forelocks. At the same time, Kabuki now developed from its previous dancing-act form into a regular drama centred on a dramatic plot with realistic acting. The Kabuki was centred in the cities of Kyōto, Ōsaka, and Edo. In Kyōto and

Ōsaka Kabuki, the speciality was the *wagoto,* a drama with a pronounced comical element, centred on love, whereas the popular form of Edo Kabuki was the *aragoto,* which had developed from the *kinpira jōruri,* based on the theme of the heroic tale of Kinpira—son of Sakata Kintoki and one of the leading followers of Minamoto Yorimitsu—and adopted the manners of the *kabukimono*—the rowdy bucks of the age.

The traditional arts of Nō drama, the tea ceremony, and flower arrangement also reached new stages of development in the period. The tea ceremony (*cha-no-yu*) in particular became popular and was practiced not only by the shogun and domain lords but also by the newly risen merchants, who used their wealth to become eager collectors of tea-ceremony utensils with historical associations. As the tea ceremony became popular, many schools came into existence, including that of Sen-ke (Sen house), the school of Sen Rikyū, fostering the tendency for the art of the tea ceremony to be monopolized by the house heads of the various schools and for the profession of tea master to develop. This "house head" (*iemoto*) system gradually also spread to flower arrangement and to other arts. One result of this was to inhibit further development of these artistic forms.

Distinctive tendencies also arose in the fine arts and crafts. One representative of this is Ogata Kōrin, who brought decorative painting to completion and bequeathed to posterity many splendid masterpieces in gold lacquer (*maki-e*) and other work. The technique of dyeing and weaving was also improved, and, in Kyōto, Miyazaki Yūzen brought the splendid *yūzen-zome* (rice-paste batik method of dyeing) to completion, while the kimono became even more colourful. In Edo, Ukiyo-e drawing in traditional styles was further developed by Hishikawa Moronobu, who depicted not only the usual courtesans and actors but also vividly portrayed various aspects of the lives of ordinary people. Besides his original drawings, he used the Chinese technique of wood-block printing to satisfy popular demand. Famous centres of pottery production also arose at various places throughout the country.

Changes in ceremonies and daily customs

The old ceremonies of the Imperial court and the various forms of deportment developed in the *bakufu* were extended to the common people and shaped the manners of the cities and farming regions. Japanese customs in dress, food, and housing became established at this time. The custom changed from eating twice a day to three times a day; in the cities rice became the normal food, and a rich variety of cakes and sweets was consumed.

The weakening of the bakuhan system. On entering the 18th century, the inner contradictions of the *bakuhan* system came to the surface and began to show signs of weakness. The finance of the *bakufu* and domain lords was based on a rice-using economy, in which executives endeavoured to levy taxes—to be paid in kind, mostly in rice, centred on the yearly crop. Rice and other crops were then transported to the great central cities like Edo and Ōsaka and exchanged into money. This forced the farmer-producers to subsist at a low standard of living and to be as self-sufficient as possible, being able to purchase only iron tools, salt, medicines, and other such goods that they could not produce for themselves. Taxes, however, came to be paid in money to a fair extent, and, because farmers also wanted to widen the scale of their activities and to enjoy a more comfortable life, it was unavoidable that they would show a strong interest in commercial-goods production. But whereas rich farmers profited from commercial-goods production, the vast majority of poor farmers and peasants became more deeply impoverished as a result of their involvement in such production. Squeezed by the merchant intermediaries, who forced them to sell their products cheaply, many were forced to part with their lands.

Thus, as the commercial economy extended into the farming villages, social divisions arose among the farmers. Tax-paying connections became unstable, and the financial difficulties of the warriors, who existed on the taxes paid by the farmers, were naturally aggravated. Because the level of agricultural technology in the feudal period was generally backward, hundreds of thousands of people starved or left their villages during periodic crop failures and famines, and the abandonment of cultivated land also became conspicuous. The samurai class had long since taken up normal residence in the cities. With the development of the urban way of life, they now suffered from increased expenditures. The *bakufu* and the domains tried to suppress commercial-goods production, which destroyed the self-sufficient economy of the farmers. But when all such attempts failed, they encouraged such production, seeking to supplement their finances by monopolizing the farmers' commercial goods and selling them themselves. This policy was the monopoly system, a phenomenon that may be termed the merchantization of the feudal lords. Thus, on top of increased taxes, the farmers were deprived of the profits of their commercial goods. Unable to restrain themselves, they violated the strict legal prohibitions with repeated farmer uprisings (*hyakushō ikki*). Meanwhile, the city poor, who were driven to the border line of starvation by the rising price of rice and other commodities, often rioted, plundering and destroying rice shops and pawnshops; these urban riots were called *uchikowashi* ("destruction").

Rural and urban unrest

Political reform in the bakufu and the domains. Continual political reforms made by the overlords in response to this crisis in the feudal system characterize the latter half of the Tokugawa period. Such reforms began with the eighth shogun, Tokugawa Yoshimune (ruled 1716–45). Yoshimune appointed civilians to official posts in finance and rural administration in order to increase government efficiency. As an emergency policy, he ordered the domain lords to make rice contributions (*agemai*), which he allotted to the *hatamoto* to supplement their stipends. More characteristic was his effort to increase tax yields by opening new lands to cultivation and revising the method of taxation. He also sought to bring fresh air into a stuffy, despotic government by such means as setting up complaint boxes, or *meyasu-bako,* to hear the views of the people and drawing up a part of a legal code, the Kujikata Osadamegaki, which mitigated criminal punishments. He made efforts to control the falling price of rice, earning him the name of "the rice shogun." But when the price of rice rose sharply in a great famine in the 1730s, the common people of Edo attacked the wholesale rice dealers who had cornered the market. This was the first such riot in Edo. Yoshimune's reforms took many twists and turns; but in 1744, the year before his retirement, the receipts of the *bakufu* both in total land taxes and in tax receipts reached their highest level for the entire Edo period. For this reason Yoshimune was called the "great ruler who restored the *bakufu.*" His success, however, was possibly due to the fact that the disturbances had not yet become very grave, whereas the political power of the *bakufu* was still quite strong.

The "rice shogun"

When Yoshimune's son Ieshige became the ninth shogun, government by the personal attendants of the shogun, which Yoshimune's personal rule had prevented, was revived. Chamberlains, or *soba-yōnin,* who handled communications with the senior councillors, gained strong powers of authority as his spokesmen when they won the Shogun's confidence. One such man was Tanuma Okitsugu, who rose from chamberlain to be senior councillor. Okitsugu delighted in bribery, and he was criticized by an opposition group for corrupting the *bakufu* government. But he was an active reformer who further developed some of Yoshimune's programs. Instead of suppressing the activities of the big-city merchants, Okitsugu used them to promote production; and while advancing the development of the commercial-goods economy, he sought to control it. His ready recognition of the commercial and industrial guilds, or *kabu nakama,* seems to have been aimed not so much at gaining contributions (*myōga-kin*) for the treasury as for establishing a controlled commercial-goods circulation, linking the city guilds with the village producers. Okitsugu was criticized by the people for issuing large amounts of debased coinage that caused a rise in prices, but he may have been trying to increase the amount of currency to match the developing circulation of commercial goods.

Okitsugu's progressive political attitude is best revealed in his development of Ezo (present-day Hokkaido) as a

Natural disasters

precaution against the southward advance of the Russians; he even considered trading with Russia. Various natural disasters occurred in his time, however: a great eruption of Mt. Asama (Asama-yama) in 1783 was followed by a great famine in 1783–87, in which large numbers of people starved to death. In addition, derelict land deserted by its cultivators increased, and the custom of *mabiki,* or infanticide, became common among parents unable to rear their children. Some people became landlords by collecting together the lands of poor farmers; others became powerful merchants. The anti-feudal struggle of the farmers also rose to an unprecedented pitch, until, in 1787, a large-scale riot lasted for three days and threatened to reduce Edo to anarchy. Okitsugu had already been dismissed as senior councillor in the previous year, and Matsudaira Sadanobu, grandson of Yoshimune and the lord of Shirakawa domain (in modern Fukushima Prefecture), was indicated as his successor. But Okitsugu's supporters in the *bakufu* made every effort to prevent Sadanobu's appointment, and for more than six months the political situation remained a complete vacuum. The farmers' riot suddenly changed the situation, however, and Sadanobu was appointed senior councillor.

Sadanobu is renowned as the director of the Kansei reforms (1789–1801). He rejected Okitsugu's free and easy administration and instituted a policy of retrenchment. He set out to reduce the high prices in the great cities and had a fund established in Edo under the name of *shichibu tsumikin* (70 percent reserve fund); knowing that land and house rents were high in the shogun's capital because of the heavy *machinyūyō* tax levied on its landlords, he reduced this tax and set aside 70 percent of it as a fund for the relief of the poor. To relieve the hardships of the *bakufu* retainers, he took emergency measures to cancel the debts of the *hatamoto* to the *fudasashi,* the Edo merchants who handled the exchange of their stipends. The farming villages, which were the foundation of the *bakuhan* system, had been devastated in the great famine of the 1780s, so while encouraging the *daikan* (a head official managing the *kuraireichi*) to bring land back into cultivation and to increase the population of the villages, Sadanobu also issued orders that annual rice taxes were to be increased as much as possible.

Sadanobu was a firm admirer of Chu Hsi studies, and he believed that government must be conducted on the basis of Confucian benevolent rule. He established an examination system for promotions among *bakufu* officials and also prohibited all teachings except those of the Chu Hsi school at the Shōheikō, the *bakufu* official college.

Sadanobu's reforms give the impression of an overly severe reaction to Okitsugu's administration; and whereas people at first welcomed them, antipathy gradually increased. Within the *bakufu* also, the O-Ōku, or shogunate domestic quarters, composed of ladies, disliked Sadanobu's reforms and forced him to resign by covert stratagems (the Shogun's favourite mistresses tempted the Shogun to expel him). While Sadanobu was senior councillor, a Russian envoy, Adam Laxman, landed at Nemuro in 1792 and requested trade relations, but the *bakufu* did not give its assent. Sadanobu ordered that plans be drawn up immediately for a coastal defense system centred on Edo Bay (now called Tokyo Bay), while he himself inspected the coastline of Izu, Sagami, and Bōsō. Because of Sadanobu's resignation (1793), these plans were not carried out; but the *bakufu* councillors of this time were the first to react to the footsteps of the foreign nations advancing on Asia, which now came to be heard through the wall of national seclusion.

In conjunction with the *bakufu* programs, reforms were carried out within the various daimyo domains. A distinctive feature of domain reforms at this time, however, was that they tried to apply stronger regulations and control over the commercial-goods economy of the farmers.

Sadanobu was senior councillor during the reign of Tokugawa Ienari, the 11th shogun. Ienari was restrained by Sadanobu's strict political reforms, but when the latter left the *bakufu* council, the Shogun was able to relax. Even so, Ienari was not completely free while the councillors who had supported Sadanobu's reforms were still alive. During the period 1804–31 these men died one after another, and the *bakufu* government became even more lax than in the time of Okitsugu. Mizuno Tadaakira, a senior councillor in this period, had risen as a personal attendant to Ienari and had business ability. But he welcomed bribery, and the other officials followed his lead, greatly increasing the corruption of the *bakufu.* On the surface, things seemed peaceful, but, underneath, the stagnation of the feudal system became even more grave. Especially in the farming villages of Kantō, right in the lap of the *bakufu,* homeless ruffians and gamblers continually caused disturbances. The *bakufu* therefore set up an office called the Kantō Torishimari-deyaku (Supervisors of the Kantō District) to strengthen police control of the area, and it ordered the villages of Kantō to form associations to assist this office.

The Russian attack on Etorofu

The growth of the northern problem. In the early 1800s foreign relations became a fairly pressing problem, and the situation in Ezo fell into confusion. In 1804 another Russian envoy, N.P. Rezanov, came to Nagasaki to request commercial relations. When the *bakufu* refused this request, his men attacked Etorofu Island. Prior to this, the *bakufu* had relieved the Matsumae domain of eastern Ezo and placed it under its direct control; and in 1807 the *bakufu* also took direct control of both eastern and western Ezo for defensive purposes. In 1808 the English warship "Phaeton" made an incursion on Nagasaki, and three years later the Russian naval lieutenant V.M. Golovnin landed on Kunashiri Island and was arrested by the Japanese. When these incidents were resolved, peace continued for a time in the northern regions, and the *bakufu* relaxed its precautions, returning all Ezo to the Matsumae domain in 1821. In the south, English ships often appeared in Japanese waters after the "Phaeton" incident, and the *bakufu,* in 1825, cancelled the Order for the Provision of Firewood and Water (*Shinsui Kyūyo Rei*) and promulgated the Order for the Driving Away of Foreign Ships (*Ikokusen Uchiharai Rei*). While attempting to preserve the iron law of seclusion to the bitter end, *bakufu* policy had no consistency (to foreign ships it was at times offensive and at times not), and it was utterly powerless when it received the full weight of foreign pressure in about the 1840s.

New learning and thought. As this weakening of the *bakuhan* system grew more serious, new movements took place in scholarship and culture. In the latter half of the 17th century the Kogaku (Study of Antiquity) school arose; scholars criticized Chu Hsi studies and advocated a return to the original ideals of Confucianism. This view was developed by Itō Jinsai and Ogyū Sorai. Especially in his work *Seidan,* Sorai insisted that the main reason for the financial distress of the warrior class in both the *bakufu* and the domains was that there was no "system" in things, and that when the warriors left the villages and moved to the cities, they had to buy everything, whereas before they had been able to live self-sufficiently. Various other schools of Confucianism arose, such as Setchūgaku (Eclectic Studies) and Kōshōgaku (Positivistic Studies). Conflict between the various schools became fierce, and the academic world grew confused. The authority of Chu Hsi studies grew weak, and, at the time of the Kansei reforms the *bakufu* attempted to reinvigorate them. It issued an order prohibiting all other schools of Confucianism in the college of the *bakufu,* but it had no marked results. At the same time, studies like Confucianism spread remarkably in the provinces, as can be seen from the establishment of domain schools (*hankō*), principally in the later Edo period, for the education of the domain samurai. Thus, learning and culture arose in the domains, accompanied by a growth of scholarship with local colouring. Among such schools, the Kaitoku-dō in Ōsaka was famous as the "townsmen's university." This school was founded by cooperation between Confucian scholars and rich merchants, and both samurai and merchants sat together to hear lectures. It is not surprising that the unique thinker and scholar Yamagata Bantō should have been produced from this rationalized way of study.

New movements also appeared in Shintō, which, with Confucianism and Buddhism, acted as the ideology of popular education. The Confucian scholar Yamazaki Ansai formulated Shintō from the standpoint of Confucianism and proclaimed the Suika form of Shintō. But in the later Tokugawa period popular interest in Shintō grew progressively stronger, centred on the Ise faith. This tendency was spurred on by lectures that explained Shintō in terms easily understood by the common people. Thereafter, Shingaku (Heart Learning) arose, which explained Confucian, Shintō, and Buddhist teachings to the townsmen and farmers. Its founder was Ishida Baigan. Kokugaku (National Learning) was also established against the same social background. Kamo Mabuchi studied the *Man'yōshū,* the most ancient collection of poetry in Japan, and other ancient writings, urging a return to old ways that had not been defiled by foreign ideas, such as Confucianism and Buddhism. By studying the ancient language of Japan's oldest classic, the *Kojiki* ("Records of Ancient Matters"), his pupil Motoori Norinaga clarified Japan's ancient system of morality, called *kannagara no michi* ("way of the gods"). Inheriting Norinaga's explanation of Shintō, called Fukko (Restoration) Shintō, Hirata Atsutane regarded Japan as the centre of the world; and fiercely upholding the conception of Japan as a divine country (*shinkoku*), he strongly advocated reverence for the Imperial house. Hirata's thought, along with Mitogaku (Mito school), provided the ideological foundation for the "*Sonnō jōi*" ("Revere the Emperor! Drive out the barbarians!") movement of the late Tokugawa period.

Western studies

Studies of European modern science also arose, termed *yōgaku* ("Western learning") or *rangaku* ("Dutch learning"). A great stimulus to the concrete development of Western studies was provided by the publication, in 1774, of the *Kaitai shinsho,* a translation by Sugita Gempaku and others of an anatomical book imported from the Low Countries. Thus, Western studies became progressively more vigorous, centred on medicine. But as the crisis in the feudal system grew more serious, many scholars of Western studies criticized the seclusion policy, making the *bakufu* nervous. The persecution of Watanabe Kazan, Takano Chōei, and other representative scholars of Western studies in the *bansha no goku* incident (1839), which resulted from a conservative plot within the *bakufu,* seriously undermined Western studies in Japan. Thereafter, as consciousness of the foreign threat grew stronger, Western studies came to place heavy emphasis on the field of military technology.

Other philosophers also appeared who repudiated feudal society. Andō Shōeki, in his *Shizen shin'eidō,* portrayed an ideal society in which all people equally engaged in farming, without social distinctions or exploitations. This shows the extent to which the common people were troubled by the contradiction of feudalism. Even if Shōeki is considered exceptional, other men successively appeared with an anti-feudal worldview directly or indirectly influenced by empirical science and Western studies. Miura Baien of Kyushu called his learning Rational Studies (Jōrigaku); it contained a dialectical method of thought. Hiraga Gennai, the son of a foot soldier of the Takamatsu domain, disliked the restricted life of the warrior; he became a masterless samurai (*rōnin*) and thought and acted freely. He advocated that Japan prevent the outflow of gold and silver by promoting domestic production and exchanging these products for foreign goods. Because this view agreed with Tanuma Okitsugu's desire for a production development policy, Hiraga was employed by Okitsugu and sent to Nagasaki. While experimenting with such things as functional dynamos and thermometers, Gennai gave full play to his genius by cultivating sugarcane and carrots, producing Dutch-style pottery, and surveying and developing mines in various provinces of the country. He also produced masterpieces as a dramatist. An attitude of naive materialism grew in his thought.

The work of Shiba Kōkan and Yamagata Bantō also appeared at this time. Kōkan is known as the pioneer of etching in Japan; but he also displayed an evolutionistic attitude, repudiating the feudal status system on the ground that both the emperor and the beggar were similar human beings, thus insisting on human equality. Bantō was chief manager to a rich Ōsaka merchant and had financial ability. In his work *Yume no shiro* ("Instead of Dreams"), he reconstructed the Japanese history in the age of gods from the worldview of natural science. He, too, had a strong feeling for human equality, saying "both in ancient times and at the present, there is no upper or lower among human beings."

Growth of popular knowledge

The common people of the Tokugawa period, by their production and labour, were progressively reared in empirical knowledge, and their self-awareness as human beings became stronger. At the outset of the period only a handful of upper-class farmers, such as the *shōya* or *nanushi,* were literate; by the end of this period, with the exception of the very lowest class, farmers were all at least partly literate. This is also connected with the diffusion of the *terakoya* (temple schools), the educational organs of the common people. But in any case, it reflected the growth in popular knowledge extending more than 300 years. This is also shown by the fact that "village conflicts" (*murakata sōdō*) became more fierce in the later part of this period, as the farmers sought to censure the improper acts of village officials and to make the village more democratic. The fact that the leadership in these conflicts was taken by the middle- and lower-class farmers was a natural phenomenon. In comparison with the present, however, it is undeniable that, in the Tokugawa cities and villages, backward social tendencies were still firmly rooted. The cause of this backwardness may be attributed to the collective abuses brought about by the feudal system, but the role played by such religions as Shintō and Buddhism cannot be overlooked.

Buddhism especially had strong powers of regulation over the lives of the common people. Among the various Buddhist sects in this period, the Jōdo, Jōdo Shin, Zen, and Nichiren made striking advances because their temples were guaranteed in their privileged status by the implementation of the *terauke* ("temple certificate") system of the *bakufu.* Besides their previous role in conducting funeral rites and masses, they now had charge of registration and census; thus grafted onto the livelihood of the people, their operation was generally stabilized. The tendency also became strong for the people to seek *genze riyaku—i.e.,* to pray for happiness during their lifetime, such as for commercial prosperity or restoration of health—and not to wait for happiness after their death. In response to this, various ceremonies were conducted and efforts made to swell the financial income of the temples. Among such ceremonies, the most important were *kaichō* and *tomitsuki. Kaichō* consisted of allowing the people to worship a Buddhist image that was normally kept concealed and not generally displayed. Gradually this came to be performed by moving the image out to other cities and villages. *Tomitsuki* was an officially authorized lottery, and in Edo the raffles at such temples as the Yanaka Tennō-ji, the Yushima Tenjin, and the Meguro Fudō (better known as the Ryūsen-ji) were famous. Among the Buddhist priests who profited from the trend of the times, there were some who led loose private lives, providing the Confucian scholars with examples for demanding that Buddhism be stamped out.

Various sorts of popular faiths flourished in the cities and villages. Shugen-dō had, as its special characteristic, prayers by itinerant monks (*yamabushi*) for curing illness or bringing happiness, but in its teachings, while centred on Buddhism, it also contained beliefs drawn from Shintō and elsewhere to meet the religious feelings of the people. A new faith in healing spirits arose from the view that human suffering could be cured only by men who had suffered the same hardships themselves, and in the late Tokugawa period a development took place toward faith in living gods (*ikigami*) who could respond to the various requests of the common people and who became revered as sect founders (*kyōso*). These sects included the Kurozumi-kyō, founded by Kurozumi Munetada, the Konkō-kyō of Kawate Bunjirō, and the Tenri-kyō of Nakayama Miki. People like Nakayama Miki reflected the confused social conditions of the late Tokugawa period, and their advocacy of *yo-naoshi,* or relief of the world by social reform, had political undertones. Influenced by the popularity of the cult of Shintō shrines, periodic pilgrimages to Ise, called *okage-mairi* or *nuke-mairi,* became popular. Pilgrimages were also made to Ise by groups of several hundreds of thousands of common people. Among such pilgrims, there were those who had political hopes for *yo-naoshi* in common with the faiths of the sect founders.

The maturity of Edo culture. In the early 19th century the city culture that had arisen in Edo was brought to full maturity in learning and craftsmanship. This Edo culture was supported by rich townsmen and warriors, and it was also widespread among the townsmen generally. Literary styles took various forms; representative authors are Santō Kyōden in the *sharebon,* or genre novel, Jippensha Ikku in the *kokkeibon,* or comic novel, and Takizawa Bakin in the *yomihon,* or regular novel. They examined in detail such things as the townsman's way of life, customs, conceptions of beauty, and ways of thinking. In the world of art, Ukiyo-e came to completion in both form and content and established its position as popular art. *Nishiki-e,* wood-block printing in many colours, was invented by Suzuki Harunobu and entered its golden age with the prints of Kabuki actors by Tōshūsai Sharaku and of courtesans by Kitagawa Utamaro. In the last years of the Edo period, the masters of wood-block landscape prints, Andō Hiroshige and Katsushika Hokusai, created new boundaries and even influenced foreign art. As a result of links with the city culture, scholarship arose even in local towns and villages, where crafts and products with distinctive local colouring were supported by landlords and merchants. A national culture emerged from these city and local cultures and became the foundation of a modern culture that developed during and after the Meiji period.

Literary styles of the Edo period

But signs of stagnation and corruption also appeared in some aspects of Edo culture—a reflection of the crisis in the *bakuhan* system. The crisis had reached new levels by the third decade of the 19th century. A great famine, lasting several years, dealt a savage blow to the impoverished villages, and farmer uprisings and city riots reached unprecedented peaks. In 1836 an uprising took place in the Gunnai district of Kai Province (Yamanashi Prefecture), then under direct *bakufu* control, and spread westward to the Kuninaka plain across the Sasago Pass, by this time numbering more than 50,000 adherents. Dividing into two groups, insurgents attacked the residences of nearly 500 prominent men and for a time reduced the centre of Kai to anarchy. How deeply the *bakufu* was shocked by this can be seen in the sentencing of 562 persons to crucifixion for their part in the uprising. A year later in the city of Ōsaka, also under direct *bakufu* control, Ōshio Heihachirō, a former city official, raised a revolt to overthrow the officials and rich merchants and to relieve the poor; although the uprising was speedily suppressed, the *bakufu* was greatly astonished that a former faithful official would lead a revolt. In addition to these domestic disturbances, the European powers began to press more heavily upon Japan. The Opium War (1839–42) arose between the Ch'ing dynasty of neighbouring China and the British, and foreign gains following this war filled the *bakufu* authorities with a sense of crisis. Tokugawa Nariaki, the lord of the Mito domain, urged the *bakufu* to make definite political reforms; he called the domestic resistance of the common people the domestic anxiety and the pressure of the foreign nations the foreign anxiety.

Internal and external threats

The Tempō reform and the collapse of the bakuhan system. Faced by this threatening situation in both domestic and foreign affairs, the chief senior councillor, Mizuno Tadakuni, instituted a reform program; this Tempō (the name of the year period from 1830 to 1844) reform lasted only from 1841 to 1843, however. Tadakuni set in order the regulations for the government officials and encouraged them to practice frugality and the literary and military arts. He also aimed to restore the farming villages devastated by the great famine. Tadakuni was stricter than earlier reformers in that he planned to force temporary residents in Edo to return to their home villages and to restrict the commercial-goods production of the farmers to make them produce rice. In order to lower the price of commodities in the cities, he applied detailed regulations to the townsmen. He further ordered that the *kabu nakama,* or merchant and artisan guilds, be dissolved because he regarded them as the cause of rising commodity prices. Tadakuni planned to reclaim by military means the Imba Swamp (Imba-nuema; in modern Chiba Prefecture) so that food supplies could be conveyed to Edo from the provinces of Hitachi and Kazusa (Ibaraki and Chiba prefectures) if Edo Bay were blockaded by foreign ships. Plans for the defense of Edo Bay took concrete form and included the seven islands of Izu. Tadakuni also promulgated a land-requisition (*agechi*) order to bring daimyo and *hatamoto* domains surrounding Edo and Ōsaka under direct *bakufu* control; the main object of this was the defense of Edo, but it was also a plan to supplement the finances of the *bakufu.* The *agechi* order was finally withdrawn, however, in the face of fierce opposition from the daimyo, *hatamoto,* and people of the domains concerned, and, as a direct result of this failure, Tadakuni was driven from his seat among the senior councillors in 1845. Tadakuni predicted that because of his reforms the Tokugawa shogunate would keep its government for another 30 years, and it was in fact almost 30 years after his reforms that the *bakufu*'s downfall took place (1867). In the same Tempō period, administrative reforms were carried out in many of the domains, and the reforms of the powerful domains in southwestern Japan, especially Chōshū and Satsuma, are noteworthy. In these domains middle- and lower-class samurai came forward as reformers, replacing the previous conservative officials. Adopting the slogan "Enrich the country, strengthen the military" ("*Fukoku kyōhei*"), these new officials were able to institute policies that improved domain finances; the resulting surpluses were used to modernize the military armaments of the domains. The way was thus gradually being prepared for the emergence of the leaders of the Meiji Restoration (1868) and of modern Japan.

In 1845, when Abe Masahiro replaced Mizuno Tadakuni as head of the *rōjū* (senior councillors), various movements appeared that can be called reactions to the Tempō reform. One of these involved the commissioners for the shogun's capital, with close connections with the Edo merchants, and tried to restore the *kabu nakama* guilds. In 1851 an order was finally issued for the revival of the *nakama.* But the guilds included those (sometimes called *kumiai*) that had arisen since the dissolution of the Tempō era *kabu nakama,* and their controlling power over city markets was thus extremely restricted. The confrontation of the city merchants with the village producers and local merchants over monopoly of commercial-goods circulation routes had grown fiercer, and the former had been forced to yield further.

Thus, the domestic reaction to the Tempō reform was comparatively calm, and the major stumbling block facing the *bakufu* was the foreign problem. The Netherlands, the only European power trading with Japan, saw that, if Britain succeeded in forcing Japan to open the country, it would lose its monopoly; so the Dutch now planned to seize the initiative in opening Japan and to thus turn the situation to their own advantage. In 1844 the king of The Netherlands, William II, sent a diplomatic mission urging the *bakufu* to open the country, but Abe and the *bakufu* rulers refused this suggestion. Visits by foreign ships, however, increased progressively. In 1844, 1845, and 1846, British and French warships visited the Ryukyu Islands and Nagasaki to request commercial relations. In response, the *bakufu* in 1845 established the new office of Kaibo-gakari for coastal defense and various diplomatic posts. The defense system of Edo Bay was also revived, the number of domains on guard duty was increased, and new gun emplacements were built. In 1848 it was decided not to revive the order to drive away foreign ships, which had been rescinded during the Tempō reform, but that extensive military preparations should be made.

Foreign rivalries in the opening of Japan

Rumours had long circulated among the various foreign countries that the United States government would send an expeditionary fleet to Japan. In 1846 Comdr. James Biddle of the American East Indian fleet appeared with two warships in Uraga Harbour (Uraga-kō) and held consultations on the question of commercial relations. When refused by the *bakufu,* he left empty-handed. The United States, however, eagerly desired ports for fuel and provisions for its Pacific merchant and whaling ships and would not give up trying to open Japan. But the *bakufu* had for so many years kept its place as overlord of the political regime by strictly maintaining the ancestral law of seclusion that it could not muster up the resolution to step forward and open the country. The opening of Japan was thus postponed until the last possible moment and had to be effected unilaterally by foreign pressure, backed by massive naval strength; this pressure was initiated by the squadron of U.S. warships commanded by Matthew C. Perry that entered Uraga-kō in July 1853. (K.Ma.)

JAPAN SINCE 1850

The Meiji Restoration. The term restoration is commonly applied to the political changes that returned power to the throne during the reign of Mutsuhito, who took the reign name Meiji ("enlightened rule," 1868–1912). The term *ōsei-fukkō* ("restoration of Imperial rule") made it possible to interpret sweeping changes as being traditional in motivation. Actually, the Meiji changes constituted a social and political revolution that began before 1868, and political innovations ended only with the promulgation of a constitution in 1889.

Fall of the Tokugawa. The arrival of the foreigners in the 1850s provided a new issue for domestic politics and a new measure for the effectiveness of the feudal administration. When it became clear that the Shogun was unable to protect Japan from the barbarians and that his concessions to them were made in spite of their known repugnance to the Imperial court in Kyōto, the two shogunal boasts of loyalty to and protection for the court proved spurious. The slogan "*Sonnō jōi*" ("Revere the Emperor! Drive out the barbarians!") was first raised by men who sought to influence shogunal policy and then taken up by others who wanted to embarrass the Tokugawa. The Shogun's ratification of the Harris Treaty (1858) and of others that followed was carried out in the face of strong opposition from the Kyōto court, and it brought to the surface antagonisms that had developed during the long years of peace and study. They centred in the Tokugawa house of Mito, which had done much to sponsor Confucian scholarship. The Mito daimyo made vigorous attempts to involve the Kyōto court in affairs of the shogunate with a view to establishing a nationwide program of preparedness. For this he was punished by the head of the Edo council of elders, Ii Naosuke. In 1860 Ii was assassinated by men from Mito and Satsuma, an act that inaugurated years of violence. Many of those who took part were young samurai from Edo; their swords availed little against the foreigners' guns, but they took a heavy toll of political enemies.

Reaction against foreign intervention

Years of extremism followed. The Tokugawa shogunate, anxious to rally support among its feudatories and to help them to prepare their defenses, relaxed its controls and regulations. In many fiefs young enthusiasts tried to push their feudal superiors into a less cautious and more strongly antiforeign position. It soon became obvious to most that expelling the foreigners by force was impossible. Antiforeign acts provoked stern countermeasures and diplomatic indemnities, which tightened the foreign hold on the country. After the bombardments of Kagoshima in 1863 and Shimonoseki a year later, there could be no doubt of the foreigners' military superiority. Thereafter, the slogans of antiforeignism and exclusion were used chiefly as a means of obstructing and embarrassing the shogunate. The Edo policymakers were forced to make surface concessions to the antiforeign elements, which aroused the hostility and distrust of the treaty powers. After the arrival of the British minister Sir Harry Parkes in 1865, Great Britain, in particular, began to tire of negotiating with a shogunate that stood between it and the Kyōto court and began to consider ways of dealing directly with the latter. It gradually became clear that ultimate authority lay in Kyōto.

In some fiefs the young extremists found themselves unable to budge their superiors from their conservative positions. From Chōshū (now part of Yamaguchi Prefecture), foreign shipping in Shimonoseki Strait was shelled in 1863. This drew the bombardment of Shimonoseki the following year. Samurai opinion grew so vehement that after the fief authorities submitted to Tokugawa discipline in 1864, a swift military coup brought to power, as the daimyo's counsellors, a group of men who had led the radical antiforeign movement. But they were no longer blindly antiforeign; several had secretly travelled to England. Their aims were national—to overthrow the shogunate and create a new government headed by the emperor. The same men developed militia units based on Western training methods and arms. Chōshū became the centre for discontented young samurai from other fiefs who were impatient with their leaders' caution. In 1866 Chōshū allied itself with Satsuma in expectation of a Tokugawa attempt to crush all *tozama* daimyo opponents and erect a centralized despotism with French help.

Anti-Tokugawa alliance

The Tokugawa armies were successfully repulsed at Chōshū in 1866, causing the shogunate to lose more power and prestige. The death of the shogun Iemochi in 1866 brought to power, as the last shogun, Yoshinobu, who was aware of the pressing need for national unity. He spurned suggestions that he seek French help to put down his enemies. When he was urged by a lord of Tosa to resign his powers, he did so rather than risk a full-scale assault by Satsuma and Chōshū, confident that as lord of eastern Japan he would emerge as an important figure in whatever new political organization should develop. But the young Meiji emperor, who had succeeded to the throne in 1867, was guided by several nobles in close touch with leaders of Satsuma and Chōshū, and the last shogun was manoeuvred into a choice between giving up his land, which would risk revolt from his vassals, or appearing disobedient, which would justify punitive measures. Yoshinobu's armies advanced on Kyōto, only to be defeated. Satsuma, Chōshū, and Tosa units, now the Imperial army, advanced on Edo, which was surrendered without a battle; fighting continued to the north until the summer of 1869, but the Tokugawa cause was doomed. In January 1868 the principal lords were summoned to Kyōto to learn of the restoration of Imperial rule. During the next year the capital was moved to Edo (renamed Tokyo), and the building of the modern state began.

From feudal to modern state. The Meiji government was dominated by the Satsuma, Chōshū, and court figures who had outmanoeuvred the Shogun. They were convinced that Japan would need a unified national government in order to achieve military and material equality with the Western powers. Most of them, like Kido Kōin and Itō Hirobumi of Chōshū and Saigō Takamori and Ōkubo Toshimichi of Satsuma, were young samurai of modest rank, but they did not represent in any sense a class interest. Indeed, their measures destroyed that class. In order to gain backing for their policies, they enlisted leaders of fiefs with which they had worked—from Tosa, Saga, Echizen—and maintained their cooperation with such court nobles as Iwakura Tomomi and Sanjō Sanetomi.

The cooperation of the impressionable young emperor was essential. It was taken for granted that Western strength depended on constitutionalism, which produced national unity; on industrialization, which produced material strength; and on a well-trained military. The new slogan of the day became "*Fukoku Kyōhei*" ("Enrich the country, strengthen the military"). Knowledge was to be sought in the West, the goodwill of which was essential if the unequal treaties were to be revised. Therefore, a number of missions to the West were organized. In 1871 Iwakura Tomomi led a large number of his fellow government leaders to visit Europe and the United States. The experience gained abroad strengthened convictions already formed as to measures of modernization that would be required.

Missions to the West

Abolition of feudalism. The Meiji leaders began with measures to lessen the feudal decentralization on which they blamed much of Japan's weakness. In 1869 the leaders of the Satsuma, Chōshū, Tosa, and Saga domains persuaded their daimyo to return their lands to the throne; other lords hastened to follow suit. The court took steps to regularize and make uniform administration in the fiefs, but it appointed the former lords as new governors. In 1871 the governor daimyo were summoned to Tokyo, and feudalism was declared abolished. The approximately 300 fiefs became 72 prefectures and three metropolitan districts; this number was later reduced by one-third. For the most part, the daimyo lost contact with administration, and, although they were rewarded with titles in a new European-style peerage, set up in 1884, their political importance was slight.

It was necessary to end the complex system of social stratification that had existed under feudalism; yet, it was difficult to make arrangements for the samurai, who numbered, with dependents, almost 2,000,000. In 1869 the old hierarchy was replaced with a new and simpler division whereby court nobility and feudal lords were termed aristocracy (*kazoku*); upper and middle samurai, *shizoku;* other samurai, *sotsuzoku* (a rank soon abolished); and all others, commoners (*heimin*), including the previously unlisted pariah groups. The samurai were given pensions equal to a part of their old income. When the regime found these pensions too heavy for its treasury to carry, the pensions were changed to interest-bearing but nonconvertible bonds. During the same years the distinctive hairstyle of the samurai (and those of farmers and merchants) was discouraged; the wearing of swords, the former badge of class, was later banned.

Samurai revolts

Many of the bonds were soon squandered, because few warriors had had occasion to develop commercial aptitude, and the inflation that accompanied government expenditures lessened their value greatly. In 1873, moreover, a nationwide conscription was instituted, depriving the samurai of their traditional monopoly of military service. There were a number of samurai revolts, the most serious in the southwest, which had led in the restoration movement and where warriors previously had reason to expect the greatest rewards. Some revolts, as in Chōshū, expressed discontent against administrative measures that deprived samurai of their importance, while in Saga the dissidents championed a foreign war to employ samurai.

The last and greatest revolt came in Satsuma (1877), led by the restoration hero Saigō Takamori. The new conscript levies were hard pressed to defeat Saigō, and the government had to enlist former samurai and empty its military academies in order to put down the revolt. But the revolts merely expressed regional discontents and were never coordinated. Even in the case of the Satsuma war, the loyalties of most of the Satsuma men in the central government remained with the Imperial cause.

In 1873 land surveys were begun to determine the amount and value of land on the basis of average yield in recent years, and a tax in money of 3 percent of the value was then set as the land tax. Out of the same surveys came certificates of ownership of land for farmers, who were also released from feudal controls. The land measures involved basic changes, and there was widespread confusion and uncertainty among the farmers, frequently expressed by short-lived revolts and demonstrations. The establishment of private ownership, along with measures to promote new technology, fertilizers, and seeds, soon produced a rise in recorded agricultural output. The land tax, supplemented by printed money, was the principal source of the government's income for several decades.

Although hard pressed for money, the government also began a program of industrialization, seen as essential for national strength. Aside from military industries and strategic communications, it was carried out in private hands, although the government set up pilot plants to provide encouragement. Trade and manufacturing benefitted from the new national market and legal security, although unequal treaties made it impossible to protect industries with tariffs until 1911.

Rise of the *zaibatsu*

In the 1880s fear of excessive inflation resulted in a decision to sell most of the new plants to private investors—usually people who had close relations with government officials. A small number of individuals came to dominate many enterprises; they were known as the *zaibatsu,* or financial cliques. With tremendous opportunities and few competitors, the same firms appeared in enterprise after enterprise. Their aims were close to those of the government leaders, and there were often close friendships between them. The House of Mitsui, for instance, had close relations with Meiji leaders, while that of Mitsubishi was founded by a colleague of the restoration leaders.

Equally important for building a modern state was the development of national loyalties. True national unity required the propagation of new loyalties among the masses, previously inarticulate and powerless. The early restoration government, influenced by a Shintō revival, elevated a bureau of Shintō, the state cult, to the highest position in the new political hierarchy and strove to replace Buddhism with a strong cult of the national deities. Christianity was legalized in 1873, with great reluctance, at the urging of the Iwakura mission, and thereafter it seemed important to bolster traditional outlooks without risking foreign condemnation by forcing a state religion upon the Japanese. The education system proved an ideal vehicle for ideological orientation. A system of universal education was announced in 1872. For a time its organization and philosophy were Western inspired; but during the 1880s, as the government leaders saw their countrymen turning to Western ideas and learned of a new nationalist orientation of schooling in Europe, the Japanese system was altered to include emphasis on "ethics," and in 1890 the Imperial Rescript on Education laid out the lines of Confucian and Shintō ideology, which constituted the moral content of later Japanese education. Thus, loyalty to the emperor, who was hedged about with Confucian teaching and Shintō reverence, became the centre of a citizen's ideology. Meanwhile, to avoid charges of indoctrination, the state distinguished between this secular cult and actual religion; in this way, the leaders could permit "religious freedom" while requiring a form of worship as the patriotic duty of all Japanese subjects. This uniform system of mass education was also utilized to project into the nation at large the ideal of samurai loyalty that had been the heritage of the ruling class.

Charter oath of 1868

Constitutional movement. It was widely believed that constitutions provided much of the unity that gave Western countries their strength, and Japanese leaders were eager to bring themselves abreast of the world in this respect. A government plan (1868) experimented with a two-chamber house, but it proved unworkable because the government leaders preferred to have their own way. The Emperor's charter oath (April 1868), however, committed the government to seek knowledge and wisdom throughout the world, to abandon customs of the past, to allow all subjects to fulfill their proper aspirations, and to allow popular opinion to influence their decisions.

To these statements of intent were added protests from below. A democratic movement grew out of a split in the leadership group over government policy. Itagaki Taisuke and other leaders of the Tosa faction combined with members of the Saga fief in 1873. Their demands for a punitive expedition against Korea, the obscurantist government of which had insulted Japanese envoys, had been refused because domestic reforms were to come first, and they resigned their positions. Instead of championing the old order, however, Itagaki and his friends called for a popular assembly so that future decisions would reflect the will of the people (by which they initially meant their fellow samurai) and thus preserve unity. Itagaki and his Tosa followers developed discussion and mutual-help groups and, gradually growing in political confidence and ability, organized themselves on a national basis as the Liberal Party (Jiyūtō) in 1881. It should be noted that the movement had only a narrow social and regional base at this time and that its purposes were to promote effective national unity rather than tolerance of diversity and dissent.

When the remaining Meiji leaders were asked to submit their opinions on constitutional problems in 1881, Ōkuma Shigenobu, a Saga leader, revealed a relatively liberal draft instead of first submitting it for the scrutiny of his colleagues. He also revealed sensational evidence of corruption in the disposal of government assets on the island of Hokkaido. Ōkuma was forced out of the government and he organized the Progressive Party (Kaishintō) in 1882. Itagaki's Liberal Party had a predominantly rural backing of former samurai and village leaders, many of whom objected to government taxation policies; Ōkuma's party had an urban base and attracted support in the business and journalistic worlds.

The Emperor promised that a constitution would be insti-

Economic distress of the late 19th century

tuted in 1889; the parties were urged to await the Imperial decisions quietly. The constitution was prepared behind the scenes by a commission headed by Itō Hirobumi. The period of constitution writing coincided with one of intense economic distress as the government sought to stem the inflation caused by the spending of the 1870s. But deflationary measures caused hardship in the countryside and provided a situation in which party agitation could easily kindle direct action. Several instances of this and severe government repression in the form of police and press controls forced the parties to dissolve temporarily in 1884. Itagaki travelled to Europe and returned more than ever convinced of the need for national unity in the face of Western condescension.

Itō Hirobumi also travelled to Europe for help in preparation of the new constitution. In Germany he found what seemed an appropriate balance of imperial power and constitutional forms that seemed to offer modernity without sacrificing effective control. As a balance to a popularly elected house, Itō first organized a new European-style peerage in 1884. The government leaders, military commanders, and former daimyo were given titles and readied for future seats in a house of peers. A Cabinet system was installed in 1885, and a privy council, designed to judge and safeguard the constitution, was set up in 1888. Itō resigned as premier to head the council.

The constitution was completed by 1889, and elections for the lower house were held to prepare for the initial diet, which met in 1890. The constitution took the form of a gracious grant by the Emperor, and it could be amended only upon Imperial initiative. Its provisions were couched in general terms. Rights and liberties were granted "except as regulated by law." If the diet refused to approve a budget, the previous year's could be followed. The emperor was "sacred and inviolable"; he commanded the armies, made war and peace, and dissolved the lower house at will. Effective power thus lay with the executive, which could claim to represent the Imperial will. The education rescript of 1890 was to guarantee that future generations accept the Imperial will and authority without question. In spite of its antidemocratic features, the constitution provided a much greater area for dissent than had previously existed. The lower house could initiate legislation. Private property was inviolate, and freedoms, even when subject to legislation, were greater than none at all. The budgetary arrangements meant that increased support for the military could be had only with Diet approval. Initially, a tax qualification of 15 yen limited the electorate to about 500,000; this was lowered in 1900 and 1920, and in 1925 universal manhood suffrage came into effect. The government leaders had difficulty controlling and manipulating the lower house, despite their power of dissolution and their resources for intimidation and bribery, thus illustrating that the constitution had altered the political picture. And the party leaders' cooperation with their erstwhile enemies when given a reasonable amount of prestige and patronage showed their large areas of agreement with the Meiji oligarchs.

End of the Restoration

The constitution ended the Meiji Restoration and revolution. The government leaders soon retired behind the scenes to influence the political world as elder statesmen (*genrō*) and acted to maintain and conserve the balance of ideological and political institutions they had worked out.

Imperial Japan. *Foreign affairs.* Achieving equality with the Western powers had been one of the major goals since the beginning of the Meiji period. Treaty reform, designed to end the foreigners' judicial and economic privileges provided by extraterritoriality and fixed customs rates, had been attempted as early as the Iwakura mission of 1871; but the Western powers refused to consider it until Japanese legal institutions had been brought into line with those of the West. Japanese attempts at compromise arrangements in the 1880s were denounced by the press and opposition groups in Japan. The treaty provisions for extraterritoriality were formally changed in 1894, after the completion of the Meiji institutional reforms; tariff autonomy came into effect in 1911.

Asian matters took second place to internal problems during most of the Meiji period. In 1874 a punitive expedition was launched against Formosa to chastise the aborigines for murdering Ryukyuan fishermen. This lent support to the Japanese claim to the Ryukyus, which had been under Satsuma influence in Tokugawa times; the islands were incorporated into Japan in 1879 despite Chinese protests. Adventures in Korea, however, although espoused by nationalists and, on occasion, by liberals, were avoided by the government, which was conscious of its need for internal reform and foreign approval. The matter was complicated by a growing Chinese readiness to resist Japanese interference in the affairs of Korea, China's most important tributary state. The Chinese were alert to the danger of Japanese gains. Incidents in 1882 and 1884 that might have led to war with China and Korea were instead settled by compromise. In 1885 China and Japan agreed that neither would send troops to Korea without first informing the other.

Sino-Japanese War (1894–95)

By the early 1890s Chinese influence in Korea was clearly becoming predominant. In 1894 Korea requested Chinese assistance in putting down a rebellion. When the Chinese informed Tokyo of this, Japan quickly rushed troops to Korea and, after the rebellion was crushed, showed no inclination to withdraw. Hostilities between Chinese and Japanese forces broke out first at sea and then in Korea in July–August 1894. The Japanese navy sank or captured much of the northern Chinese fleet, and a peace treaty was negotiated at Shimonoseki between Japan and China on April 17, 1895. Both powers recognized the independence of Korea; China ceded Formosa, the Pescadores Islands, and the Liaotung Peninsula, granted Japan all rights enjoyed by European powers, and made significant new economic concessions; new treaty ports were opened, and Japan received an indemnity of 200,000,000 taels in gold in two installments. A subsidiary treaty of commerce (1896) gave Japan freedom to engage in trade, manufacture, and industry in China's treaty ports and provided for tax exemption within China for all goods so manufactured. Japan thus marked its own emancipation from unequal treaties by imposing even harsher terms on its neighbour. But the European powers were not yet prepared to welcome Japan as a full equal in the imperialist scramble in China. Germany, France, and Russia forced Japan to return the Liaotung Peninsula to China. In 1898 Russia forced China to grant it the lease of that peninsula with its important naval base at Port Arthur. The war thus demonstrated that the Japanese could not maintain Asian military victories without Western sufferance. Nevertheless, the war proved a tremendous source of prestige for Japan and brought the Tokyo government much internal support; it also strengthened the hand of the army in national affairs.

Russo-Japanese War (1904–05)

Instead of accepting Japanese leadership, Korea sought the help of the Russians as a counterweight. During the Boxer Rebellion in China (1900), Japanese troops took a major part in the allied expedition that rescued foreign nationals in Peking, but Russia occupied south Manchuria, thereby strengthening communications with Korea. Realizing the need of protection against a possible combination of European enemies, the Japanese government began talks that led to an Anglo-Japanese Alliance (1902). Each signatory agreed to aid the other in the event of an attack by two or more powers, while remaining neutral if the other was at war with a single power. The Tokyo government was thus prepared to take a firmer line with respect to Russian advances in Manchuria and Korea. In 1904 Japanese ships attacked the Russian fleet at Port Arthur without the formality of a declaration of war. Japanese arms were everywhere successful; the most spectacular victory was in Tsushima Strait, where the ships of Adm. Togō Heihachirō destroyed the Russian Baltic fleet. But Japanese armies were strained to their utmost, and it was with relief that Japan accepted the U.S. Pres. Theodore Roosevelt's offer of good offices for the negotiations that led to peace, signed at Portsmouth, New Hampshire, September 5, 1905. Japanese primacy in Korea was recognized, and Russia surrendered to Japan its economic and political interests in south Manchuria (including the Liaotung Peninsula) as well as the southern half of the island of Sakhalin. The victory over Russia altered the balance of power in Asia, and Japan's ability to cope with a great European power accelerated the development of nationalist movements in Asia. Within Japan, however, the failure to secure a Russian indemnity to cover the costs of the war made the treaty unpopular.

After the conclusion of the war, Japanese leaders now had a free hand to guide the course of reform in Korea, and Korean resistance was met with force. Itō Hirobumi, sent to Korea as resident general, forced through treaties that gave Korea little more than protectorate status and forced the abdication of the Korean king. In 1909 Itō was assassinated, and the following year Korea was formally annexed to Japan. Korean liberties and resistance were crushed under military rule. By the end of the Meiji period, Japan had thus achieved equality with the West and had, in fact, become the strongest military and imperialist power in Asia.

Japan had abundant opportunity to use its new power in the years that followed. During World War I the Western powers were fully occupied in Europe. Japan took part in the war in compliance with the Anglo-Japanese Alliance, but generally it limited its participation to the seizure of the German Pacific Islands and the German holdings on the Shantung Peninsula. When China pressed for return of these, the Japanese government presented the so-called Twenty-one Demands in January 1915. China reluctantly agreed to extend the duration of the Manchurian leases and to joint control of steelworks and ironworks in central China. The German Shantung holdings were to be settled by agreement between Japan and Germany at the time of the peace treaty; subsequently, Japan agreed to hand back the territory in return for further commercial privileges. China promised not to alienate harbours in Fukien province to any other power without Japanese approval. But the Chinese resisted group V of the Twenty-one Demands, which would have reduced China to the status of a Japanese ward. Japan had gained abundant opportunity for the exploitation of Manchuria, but the ill feeling aroused by the negotiations, together with Chinese chagrin at failure to recover its losses in the Treaty of Versailles, cost Japan any hope of Chinese friendship. Subsequent Japanese sponsorship of corrupt warlord regimes in Manchuria and North China helped to confirm the anti-Japanese nature of modern Chinese nationalism.

The Twenty-one Demands

Japanese behaviour when the Allies intervened in Siberia in 1918 after the Bolshevik Revolution furthered the impression of Japanese rapacity. One of the principal reasons for a disarmament conference in Washington, D.C. (1922), was an attempt to lessen Japanese influence. A network of treaties was worked out that placed restraints on Japanese ambitions while guaranteeing Japanese security. Japan, Great Britain, the United States, and France concluded a four-power pact that replaced the Anglo-Japanese alliance; a five-power pact (with Italy) for disarmament set limitations for capital ship construction on a ratio of five for Great Britain and the United States to three for Japan. Parallel guarantees against fortifying advanced bases assured Japan of safety in Pacific waters. A nine-power pact would, it was hoped, protect China from further unilateral demands. Japan subsequently agreed to retire from Shantung, and, shortly afterward, Japanese armies withdrew from Siberia and northern Sakhalin. In 1925 a treaty with the Soviet Union extended recognition and ended active hostilities.

Thus, by the mid-1920s Japan's great surge forward in the Pacific had ended; this brought hope that a new quality of moderation and reasonableness, based on the absence of irritating reminders of inferiority and weakness, might characterize Japanese policy.

Lower-house opposition

Constitutional government. The inauguration of constitutional government in 1890 saw a vigorous and often obstreperous opposition in the lower house of the Diet, and it was probably general determination to prove that parliamentary institutions could work in Japan that forced the party and government leaders to cooperate sufficiently to make the system work. The first Cabinets, led by Yamagata Aritomo, Matsukata Masayoshi, and Itō, attempted to maintain the principle that the government, which in their view represented the emperor, should be aloof from parties and that it was the duty of the lower house to approve government requests. This policy failed because the parties desired to increase their power and patronage and therefore sought Cabinets responsible to the lower house. It was only the Sino-Japanese War that produced the kind of unity the constitution makers had envisaged. In the years that followed, the oligarchs formed alliances with the two parties, usually exchanging a Cabinet seat or two for support in the lower house. These arrangements proved unsatisfactory as party leaders soon raised their sights. In 1898 Itagaki and Ōkuma combined forces to form a single party, the Kenseitō, and, because this ruled out successful administration by a nonparty Cabinet, they were allowed to form a government. But their alliance was of short duration, as long-standing animosities and jealousies enabled antiparty forces among the bureaucracy and oligarchy to force their resignation within a few months.

A discernible division now developed among the dwindling group of Meiji leaders. Yamagata Aritomo dominated the army and much of the bureaucracy. During the two years he held power after the fall of the Kenseitō Cabinet, he strengthened legal and institutional safeguards against rule by political parties and secured an Imperial ordinance that service ministers should be career officers on the active list; this gave the army or navy power to break a Cabinet. Partly in reaction, Itō Hirobumi, also of Chōshū, formed his own political party in 1900, the Rikken Seiyūkai, enlisting most of the former followers of Itagaki. Thereafter, practical political goals of power and patronage softened the hostility between oligarchs and politicians.

After 1901 both Itō and Yamagata retired from active participation in politics; until 1913 Cabinets were led by their protégês Saionji Kimmochi and Katsura Tarō. Basic decisions of politics and policy, however, continued to be made by the core group of elder statesmen, who advised the Emperor on all important decisions and selected prime ministers by rotating power between the two principal factions. Saionji was the last to be recruited into this extraconstitutional body.

Changing political patterns

With the death or enfeeblement of the first generation of leaders, the pattern of political manipulation changed. No subsequent group could match the prestige the Meiji leaders had enjoyed. The Meiji emperor died in 1912 and was succeeded by a son who took the reign name Taishō ("great righteousness," reigned 1912–26); but mental illness prevented him from approximating his father's fame. The growth in prestige and power of businessmen found expression in their control of the political parties and resulted in an increasing role for professional party politicians. The *genrō*'s last attempt to seat Katsura in 1912 ended in failure, while his successor, Adm. Yamamoto Gonnohyōe, was discredited by scandals in naval procurement. Ōkuma Shigenobu emerged from retirement to head a Cabinet during World War I and was succeeded by a military Cabinet under Gen. Terauchi Masatake. In 1918, however, discontent with Terauchi's reactionary posture and administrative incompetence combined with the rising power of the party professionals to bring about the appointment of Hara Takashi (Hara Kei) as prime minister. Hara was the first nontitled person to hold that office, and his appointment marked the first party Cabinet. His assassination in 1921 cut short his cautious efforts to reduce the power of the military and the bureaucracy and to extend the franchise. After several short-lived Cabinets, a successful party Cabinet was organized in 1924 by Katō Takaaki. The army was reduced in size; moderate social legislation was enacted; and universal manhood suffrage extended the franchise to 14,000,000 voters. Meanwhile, Japan avoided stronger steps in China's civil war and pursued a conciliatory course with Russia, despite demands from nationalists, who utilized alleged outrages in China and the discriminatory U.S. Immigration Act of 1924 to warn of the futility of appeasing or cooperating with other powers.

But, as the parties grew in power, they tended to look to bureaucrats for leadership. The businessmen who supported the parties and the bureaucrats who led them shared a fear of the social movements that followed in-

Reaction to social movements

dustrialization and the importation of foreign ideas. A growing labour movement had already been checked by a special police law introduced in 1900. This was strengthened under Katō in 1925, as conservatives generally began to fear subversion in labour and tenant movements. A small Communist party was organized by a group of intellectuals in 1922, and a general interest in Marxist thought contributed to more fears of subversion. Under the Meiji constitution, party governments had to make their peace with the military, with the House of Peers, and with the conservatives close to the throne; whatever ideas for reform they had therefore had to be worked out with the utmost caution. Frequently, the Diet found itself virtually powerless, and this encouraged corruption and disorders in the chamber, which did little to win popular respect for the machinery of representative government. There were no institutional changes that enabled a government to be firmly based on popular support. The Meiji Constitution was so ambiguous in its provisions for the executive that the party prime ministers could achieve little unless they secured, through compromise, the cooperation of forces antagonistic to democratic government.

Social change. Changes in the social and intellectual scene outstripped those in the political. Many of them were related to the development of industry. After the Treaty of Shimonoseki the government utilized the Chinese indemnity to subsidize the Yawata Iron and Steel Works, which were established in 1897 and began production in 1901. Yawata depended on China for its ores. After 1900 Japan's population exceeded the capabilities of domestic food production so that there was need to import food as well. Growing textile and other consumer-goods industries expanded to meet Japanese needs and to earn credits required for the import of raw materials. Heavy industry was encouraged by government-controlled banks, which provided needed capital. Strategic industries, notably steel and the principal rail trunk lines, were in government hands, but most new growth was in the private sector, albeit somewhat concentrated in the *zaibatsu* financial and industrial giants.

Social protest

The enlarged urban population produced movements of social inquiry and protest. In 1895 the industrial labour force numbered about 400,000. Several efforts to organize socialist movements speedily met with police repression. Peace preservation laws were passed in 1900 and 1925, and in 1928 it became a capital crime to agitate against private property or Japanese state policy (*kokutai*). In 1903 a small group organized the *Heimin shimbun* ("Commoner's Newspaper"); it published *The Communist Manifesto* and opposed the Russo-Japanese War in the name of the workers of Russia and Japan before being forced to cease publication. The labour and Socialist movements gained strength after World War I, but leadership was usually theoretical and doctrinaire, with little real contact with the workers. Police repression and the difficulties of organizing a labour force of diverse industrial empires such as those of Mitsui and Mitsubishi also retarded the labour movement. Meanwhile, the increasing confidence and power of management came to influence and at times control the political parties. The Katō Cabinet of 1924–26 was sometimes referred to as a Mitsubishi Cabinet.

In the countryside the principal reflection of the new trade patterns was an additional impetus to silkworm production to augment the farmers' income. Farm villages also provided the bulk of the labourers for the new industries, and farm daughters were found in many textile plants. The early 20th century was not a time of agricultural prosperity. Farmers were handicapped by growing fragmentation of holdings and increasing tenantry. The rising number of tenants resulted in the growth of tenant organizations, especially during and after World War I. Government efforts to encourage voluntary reform brought only a law for mediation of disputes in 1924. But a financial panic in 1927 aggravated rural conditions and indebtedness even before the collapse of the U.S. silk market in 1929 spelled disaster for the farmers and workers alike.

The most lasting social changes were found in the great metropolitan centres, where a growing labour force and new middle-income groups were concentrated. The Tokyo–Yokohama area was devastated by the great Kantō earthquake in 1923, and its reconstruction as a modern metropolis symbolized the growth and orientation of the urban society. The currents of enthusiasm during and after World War I were uniformly international and largely U.S. in inspiration. Western music, dancing, and sports became popular, and rising standards of living and expectation produced the need for more and better higher education. The participation of women in office work in the new enterprises and the rise of a feminist movement, however unsuccessful, marked the beginning of changes in the family system.

The educated class grew in numbers and in vigour. Currents of thought included Western-style democracy and the new radicalism of the Soviet Union; the Marxist influence went far beyond the ranks of the struggling Communist Party—which was, in any event, soon crushed by the police. Political liberalism was championed by the University of Tokyo figure Yoshino Sakuzō, who formed a group of students and intellectuals the title of which—Shinjinkai (New Peoples Association)—symbolized the self-conscious break with tradition. Minobe Tatsukichi, a distinguished constitutional theorist, introduced the idea that the emperor was an organ of the state and not the sole source of sovereignty. Such men faced sharp criticism and had, in time, to resign their positions, but they had great influence and symbolized and stimulated advanced currents of thinking.

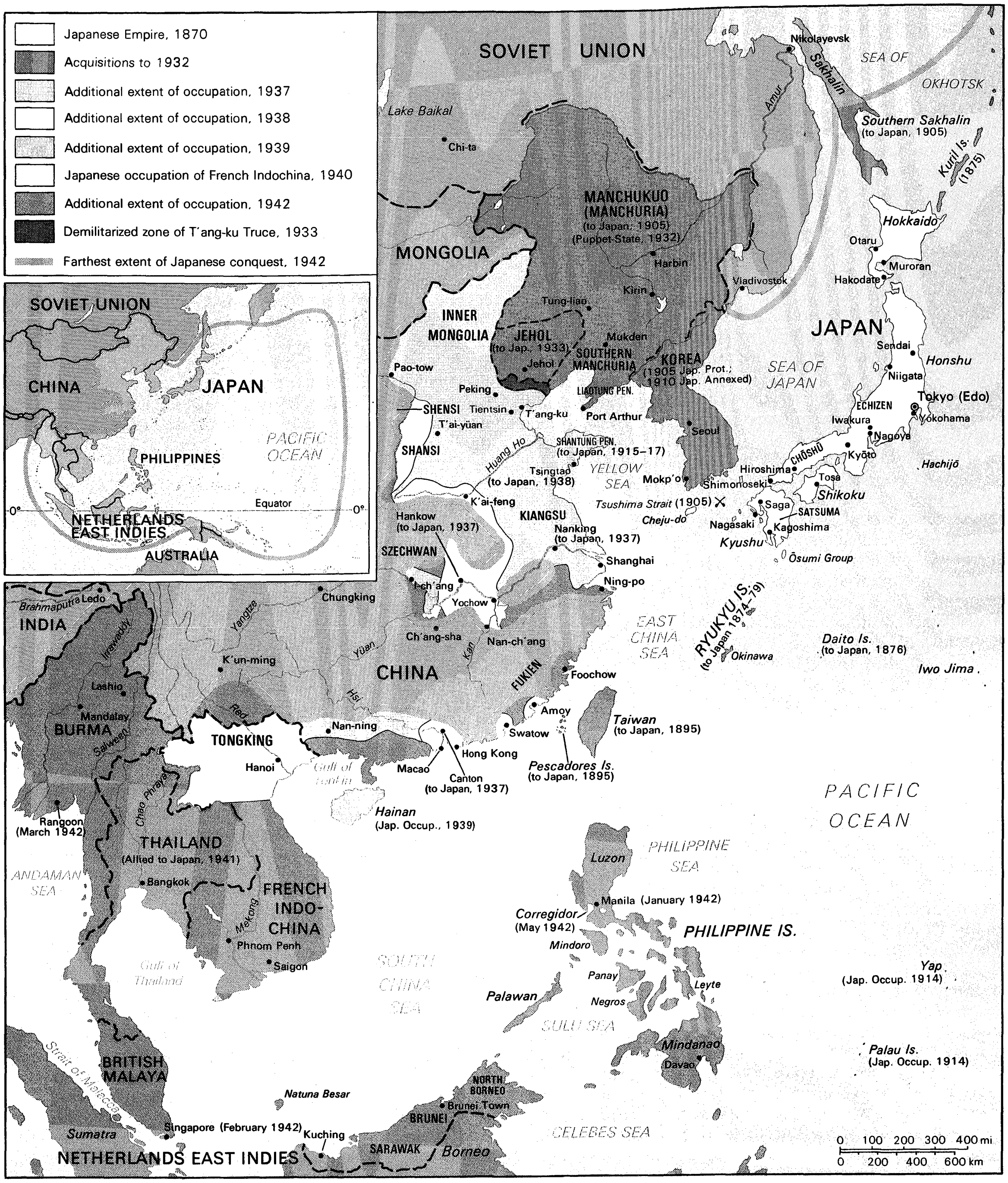
Attempts to break with tradition

The base for these new currents was precarious. Politically and institutionally, no advances—beyond the universal manhood suffrage of 1925—were scored, while, under the peace-preservation laws of 1928, a special police corps was established to seek out "dangerous thoughts." Economically, the urban classes were dependent upon the continuance of the favourable trade patterns of the 1920s. When the Great Depression at the end of the decade wrecked Japan's foreign markets and removed the possibility of the villagers' augmenting rice income with that of silk and when the irresponsibility and occasional corruption of Diet representatives contrasted with poverty elsewhere in Japanese society, many were prepared to listen to charges that the political-party government, dominated by selfish *zaibatsu* interests, had neglected Japan's markets in China, imperilled morality and decency at home, and allowed subversive trends to flourish, while the politicians reaped personal gains.

The rise of the militarists. The notion that expansion through military conquest would solve Japan's economic problems gained currency during the Great Depression of 1929. A key argument advanced to support it was that Japan's population had grown from 30,000,000 at the time of the Meiji Restoration to almost 65,000,000 in 1930; each year the problem grew worse, and the imports of needed foodstuffs increased. It was also argued that emigration to many areas was cut off because of discrimination against Oriental peoples. Efforts made by Japan and China to secure a racial-equality clause in the League of Nations covenant had been frustrated by Western statesmen who feared the anger of their constituents. So the argument ran that no recourse could be expected without resort to force.

Military distrust of party government

To these economic and racial arguments was added the military's distrust of party government. The Washington conference had allowed a smaller ratio of naval strength than the navy had desired, and the government of Prime Minister Hamaguchi Osachi in 1930 accepted and gained approval of the London Naval Conference limitations of cruiser strength over military objections. The Katō government had cut the army strength. Many service leaders had also bridled under Japan's moderation during the Chinese Kuomintang northern expedition in 1926 and 1927, and they would have preferred a much stronger stand. The Seiyūkai Cabinet under Prime Minister Tanaka Giichi reversed that policy by intervening in Shantung in 1927 and 1928. Tanaka was forced out in 1929 and replaced by Hamaguchi, under whom the policy of moderation returned. It seemed to many that such vacillation earned Japan ill will and expensive boycotts in China without gaining any advantage.



Japanese expansion in the late 19th and 20th centuries.

Many military leaders resented the restrictions that civilian governments had placed upon them, and their power was considerable. It would be wrong to attribute such views to all or even most of the high command, but enough army officers in particular held this position to furnish a possible focus for dissatisfaction among other groups in Japanese society. The idea of the frugal, selfless samurai was peculiarly useful as a contrast to the stock characterization of the selfish party politician.

These economic pressures and political misgivings were exploited by civilian ultranationalists who opposed parliamentary government as "un-Japanese." Since Meiji times a number of rightist organizations had formed, dedicated to the theme of internal "purity" and external expansion. They sought to preserve what they thought was unique in the Japanese spirit and fought against excessive Western influences. Some originated in the Meiji period, when nationalists had felt obliged to work for a "fundamental settlement" of differences with Russia; the Kokuryūkai (Black Dragon Society, popularly the Amur River Society) was one such, while others, such as the Seisantō (Productivity Society), were keyed to labour and social problems. The Kokusūikai (National Purity Society) worked to preserve national purity, while the Ketsumeidan (League of Blood) was terrorist. They opposed political parties, big business, acculturation, and Westernization. By allying with other rightists, they alternately terrorized and intimidated their presumed opponents. A number of business leaders and political figures lost their lives, and the assassins' success in publicizing and dramatizing the virtues they claimed to embody had a considerable importance in the ethos of the troubled 1930s. It is clear, however, that the terrorists never had as much influence as they claimed or as the West believed.

The principal force against parliamentary government was provided by junior military officers. Largely from rural backgrounds, distrustful of their senior leaders, ignorant of political economy, and contemptuous of the urban luxuries of politicians, the officers were ready marks for rightist theorists. Many of them were animated by goals that were national-socialist in character. Kita Ikki, a former Socialist and former member of the Kokuryūkai, wrote in his outline for the reconstruction of Japan that the Meiji Constitution should be set aside in favour of a revolutionary regime advised by "national patriots" and headed initially by a military government, which should nationalize major forms of property, limit wealth, end party-government and peerage systems, and prepare to grasp the leadership of a revolutionary Asia. Kita helped persuade a number of young officers to take part in the violence of the 1930s, in large measure designed to create disorder so great that military government would follow.

Aggression in Manchuria

The Kwantung Army, which invested the Kwantung (Liaotung) Peninsula and patrolled the South Manchurian Railway zone, provided a rich harvest of officers keenly aware of Japan's continental interests and prepared to take steps to further them. They hoped to place the civilian government in an untenable position and to force its hand. The Tokyo terrorists similarly sought to change foreign as well as domestic policies. The pattern of direct action in Manchuria began with the murder in 1928 of Marshal Chang Tso-lin, the warlord ruler of Manchuria. The action, though not authorized by the Tanaka government, helped bring about its fall. Tanaka's Cabinet, however, dared not investigate and punish those responsible, and this convinced extremist officers that their lofty motives would make retribution impossible. The succeeding government of Prime Minister Hamaguchi showed intentions of restraining military activists and powers, however, and the next plots centred around plans for replacing civilian government altogether; Hamaguchi was mortally wounded by an assassin in 1930. In March 1931 a coup involving highly placed army generals, planned to terrorize civilian politicians into a grant of martial law, was abandoned because of disagreement among the principals.

Terrorism and revolts

On September 18, 1931, came the Manchurian Incident, which launched aggression in East Asia. A Kwantung Army charge that Chinese soldiers had tried to bomb a South Manchurian Railway train (which arrived at its destination safely) resulted in a speedy and unauthorized capture of Mukden, followed by the occupation of all Manchuria. The civilian government in Tokyo could not stop the army, and even army headquarters was not always in full control of the field commanders. Prime Minister Wakatsuki Reijirō gave way, in December 1931, to Inukai Tsuyoshi. Inukai's plans to stop the armies by Imperial intervention were frustrated. In 1932 naval officers took the lead in extremism; a terrorist attack in Tokyo in May took the life of Inukai but failed to secure a proclamation of martial law. The army now announced that it would accept no party Cabinet. To forestall its desires for power, the last *genrō,* Saionji, suggested retired Adm. Saitō Makoto as prime minister. Plotting continued, culminating in a revolt of a regiment about to leave for Manchuria. On February 26, 1936, several outstanding statesmen (including Saitō) were murdered; Prime Minister Okada Keisuke escaped when the assassins mistakenly shot his brother-in-law. For more than three days the rebel unit held much of downtown Tokyo. When the revolt was put down on February 29, the ringleaders were quickly arrested and executed. The influence of the young extremists, the Imperial Way faction (Kōdō-ha), now gave way before that of the more cautious Control faction (Tōsei-ha), which had less sweeping plans for internal reform but shared many of the foreign-policy goals of the young fanatics.

The only possible source of prestige sufficient to thwart the military lay with the throne. The senior statesmen, however, were cautious lest they imperil the Imperial institution itself. The young emperor Hirohito had succeeded to the rule in 1926, taking as his reign title Shōwa. His outlook was more progressive than that of his predecessors; he had travelled in the West, and his interests lay in marine biology (of which the ultranationalists disapproved in one whose role it was to embody the Japanese mystique). The palace advisers feared that a strong stand by the Emperor would only widen the search for victims and might lead to dethronement of the monarch. As international criticism of Japan's aggression grew, many Japanese rallied to the support of their soldiers.

Creation of Manchukuo

The road to World War II. Each advance by the military extremists gained them a new compromise concession by more moderate elements in the government and brought greater foreign hostility and distrust. Rather than attempt to thwart the military, the government agreed to reconstitute Manchuria as the "independent" state of Manchukuo. The last Manchu emperor of China, P'u-i was first declared regent and then enthroned as emperor in 1934. Actual control lay with the Kwantung Army, however; all key positions were held by Japanese, with surface authority for cooperative Chinese and Manchu. A League of Nations committee recommended in October 1932 that Japanese troops be withdrawn, Chinese sovereignty in Manchuria recognized, and a large measure of autonomy granted to Manchuria. The League called upon member states to withhold recognition from the new puppet state. In March 1933 Japan formally withdrew from the world body. Thereafter, Japan poured technicians and capital into Manchukuo, exploiting its rich resources to establish the base for the heavy-industry complex that was to undergird the new order in East Asia.

In northern China, boundary areas were consolidated in order to enlarge Japan's economic sphere. In early 1932 the Japanese Navy precipitated an incident at Shanghai in order to end a boycott of Japanese goods there; but Japan was not yet prepared to challenge other powers for control of central China, and a League of Nations commission arranged terms for a withdrawal in May 1932. Frustrated naval officers returned to Tokyo to carry out the violence that killed Inukai on May 15. A move southward from Manchuria into Jehol in January 1933 led to the T'ang-ku Truce in May, whereby a demilitarized zone was set up between Peking and the Great Wall. This brought the fighting to a temporary close. In 1934, Japan made it clear that it would brook no interference in its China policy and that Chinese attempts to procure technical or military assistance elsewhere would bring Japanese opposition.

Further external ambitions, however, had to wait for the resolution of domestic crises. The military revolt in Tokyo in February 1936 marked the high point of the extremist faction and the consolidation of power by the Control faction within the army. Finance minister Takahashi Korekiyo, whose policies had brought Japan out of its economic depression, was killed, and his opposition to further inflationary spending was thus stilled. When further efforts by the palace advisers to defer full power for the military failed, the leadership went to the popular but ineffective Konoe Fumimaro, scion of an ancient court family (June 1937). In this same period Chiang Kai-shek was kidnapped by Chinese border armies at Sian in December 1936, and he formulated an agreement to consolidate Nationalist and Communist efforts into an anti-Japanese front. To this was added evidence that the Japanese people were not yet prepared to renounce their parliamentary system. In the spring of 1937, general elections showed a startling strength for a new Social Mass Party, which received 36 seats out of 466, and a heavy majority for the two parties (the Seiyūkai and its rival the Minseitō), which had combined forces against the government and its policies. The time seemed ready for new efforts by civilian leaders, but the field armies anticipated them.

The Marco Polo Bridge Incident

On July 7, 1937, Japanese troops engaged Chinese units near Peking. Soon dubbed the Marco Polo Bridge Incident, it rekindled warfare between China and Japan. Japanese armies took Nanking, Hankow, and Canton despite vigorous Chinese resistance; to the north, Inner Mongolia and the provinces of Shansi and Shensi were invaded but not fully invested. On discovering that the Nationalist government, which had retired to Chungking in Szechwan, refused to compromise, the Japanese installed a more cooperative regime at Nanking in 1940.

Japan had signed the Anti-Comintern Pact with Germany in November 1936 and later with Italy. This was replaced by the Tripartite Pact in September 1940, by which Japan was recognized as the leader of a new order for Asia, and the three signatories agreed to assist each other if any one was attacked by a power not then at war. This was directed against the United States, since the Soviets and Nazis were then allied; the Soviet Union was invited to join in the pact later in 1940.

Japanese relations with the Soviet Union were considerably less cordial than those with Germany. The Soviets consented, however, to sell their Chinese Eastern Railway holdings to the South Manchurian Railway in 1935, thereby strengthening Manchukuo. In 1937 the Soviet Union signed a nonaggression pact with China, and in 1938 and 1939 Russian and Japanese armies tested each other in two full-scale battles along the border of Manchukuo. The Soviet-Nazi pact of August 1939, however, was followed by a neutrality pact between the Soviet Union and Japan in April 1941.

The German–Japanese tie was never a close or effective one. Both parties were limited in their cooperation by distance, distrust, and claims of racial superiority. The Japanese were uninformed about Nazi plans for attacking the Soviet Union, and the Germans were not told of Japan's plans to attack Pearl Harbor. Nor, despite formal statements of rapport, did Japan's state structure approach the totalitarianism of the Nazis. A national-mobilization law (1938) gave the Konoe government sweeping economic and political powers, and in 1940, under the second Konoe Cabinet, the Imperial Rule Assistance Association was established to merge the political parties into one central organization; yet, the institutional structure of the Meiji Constitution was never altered, and the wartime governments never achieved full control over interservice competition. The Imperial Rule Assistance Association never succeeded in mobilizing all segments of national life around a leader. The emperor remained but a symbol, albeit an increasingly military one, and no *Führer* could compete without endangering the national polity. Wartime social and economic thought retained important vestiges of an agrarianism and familism that were in essence premodern rather than totalitarian.

Relations with the democratic powers

Japan's relations with the democratic powers deteriorated steadily. The United States and Great Britain did what they could to assist the Chinese Nationalist cause. The Burma Road permitted the transport of minimal supplies to Nationalist forces. Constant Japanese efforts to close this route were successful briefly in 1940, when the British felt they could not risk a second war. But anti-Japanese feeling had strengthened in the United States, especially after the sinking of a U.S. gunboat, the "Panay," in the Yangtze River in 1937. In 1939 U.S. Secretary of State Cordell Hull denounced the 1911 treaty of commerce with Japan, and thus embargoes became possible in 1940. Pres. Franklin D. Roosevelt's efforts to rally public opinion against aggressors included efforts to stop Japan, but, even after the outbreak of war in Europe in 1939, public opinion in the United States was averse to courting war by stronger measures.

The European war presented the Japanese with tempting opportunities. After the Nazi attack on Russia (1941), the Japanese were torn between German urgings to join the war against the Soviets and their natural inclination to seek richer prizes from the colonial powers to the south. In 1940 Japan had occupied northern Indochina in an attempt to block access to supplies of the Chinese Nationalists, and in July 1941 it announced a joint protectorate with Vichy France over the whole colony. The way was prepared for further moves in Southeast Asia.

The United States reacted to the occupation of Indochina by freezing Japanese assets and declaring an embargo on oil to Japan. The government was faced with the alternatives of withdrawing from at least Indochina and possibly China or seizing the sources of oil production in the Netherlands East Indies. Negotiations with Washington were carried on under the second Konoe Cabinet. Konoe was willing to withdraw from Indochina, and he sought a personal meeting with Roosevelt, hopeful of some U.S. concessions or favour with which he might convince his military leaders. But the U.S. State Department refused to agree to a meeting without prior Japanese concessions. Pressed by his war minister, Gen. Tōjō Hideki, Konoe resigned in October 1941 to be succeeded by Tōjō. Secretary of State Hull refused to agree to Japan's "final offer": Japan would withdraw from Indochina after China had come to terms in return for U.S. promises to resume oil shipments, cease aid to China, and unfreeze Japanese assets. With Japan's decision for war made, the negotiators received instructions to continue to negotiate. Preparations for the opening strike against the U.S. fleet at Pearl Harbor were already in motion. The Japanese military elected to try to establish, through a "new order in East Asia," a co-prosperity sphere in which Japan, as the centre of an industrial bloc comprising Manchuria, Korea, and North China, would draw from the rich colonies of Southeast Asia the raw materials it needed, while inspiring them to friendship and alliance by destruction of their previous masters. But, in practice, "East Asia for the Asiatics," Japan's slogan, turned out to mean "East Asia for Japan."

Preparations for war

World War II and defeat. The attack on Pearl Harbor (December 7, 1941) achieved complete surprise and success. It also unified U.S. opinion and determination to see the war through to a successful finish. The Japanese had expected that, once they fortified their new holdings, a reconquest would be so expensive in lives and treasure that it would discourage the "soft" democracies. Instead, the U.S. fleet was rebuilt with astonishing speed, and the chain of defenses was breached before these riches could be effectively tapped by Japan.

Early successes

The first years of the war brought Japan great success. Japanese troops occupied Manila in January 1942, although Corregidor held out until May; Singapore fell in February, the Netherlands Indies and Rangoon in early March. The Allies had difficulty maintaining communication lines to Australia, and the loss of the British battleships "Repulse" and "Prince of Wales," added to the U.S. Pacific fleet disaster, seemed to promise the Japanese Navy freedom of action. Tōjō grew in confidence and popularity and began to style himself somewhat in the manner of a Fascist leader. But the U.S. Navy had not been permanently driven from the South Pacific. The Battle of Midway in June 1942 cost the Japanese fleet aircraft carrier strength it could ill afford to lose, and the battle for Guadalcanal in the Solomons ended with Japanese withdrawal in February 1943.

After Midway, Japanese naval leaders came secretly to the conclusion that Japan's outlook for victory was poor. When the fall of Saipan in July 1944 brought U.S. bombers within range of Tokyo, the Tōjō Cabinet was replaced by that of Koiso Kuniaki. Koiso formed a supreme war direction council designed as a link between the Cabinet and the high command. It was becoming evident that Japan was losing the war, but no group had a program acceptable to the military leaders. There were also grave problems about breaking the news to the Japanese people, who had been told only of victories. Great fire-bombing raids in 1945 brought destruction to every major city except the old capital of Kyōto; but the generals were still determined to continue the war, confident that a major victory or a protracted battle would be the best way of gaining honourable terms. The Allied talk of unconditional surrender provided a good excuse for continuing the fight.

In February 1945 the Emperor met with a group of senior statesmen to discuss steps that might be taken. When U.S. landings were made on Okinawa in April, the Koiso government fell. The problem of the new premier, Adm. Suzuki Kantarō, was not whether to end the war but how best to do so. The first plan advanced was to ask the Soviet Union, with which Japan was still at peace, to intercede with the Allies. The Soviet government, however, was planning to enter the Pacific war, and reply was delayed while Soviet leaders took part in the Potsdam Conference in July. The Potsdam Declaration of July 26 offered the first ray of light with its statement that Japan would not be "enslaved as a race nor destroyed as a nation."

The atomic bombs

On August 6 and 9 the atomic bombs took their toll of life in Hiroshima and Nagasaki. On August 8 the Soviet Union declared war and on the 9th marched into Manchuria, where the Kwangtung Army could offer only slight resistance. The Japanese government attempted to gain as its sole condition for surrender a qualification concerning the maintenance of the Imperial institution; after the Allies agreed to respect the will of the Japanese people, the Emperor insisted on surrender. The Pacific war came to an end on August 14. The formal surrender was signed on September 2 in Tokyo Bay aboard the USS "Missouri."

Military extremists made an unsuccessful attempt to prevent the radio broadcast of the Emperor's announcement to the nation. There were a number of suicides among the military officers and nationalists who felt themselves dishonoured, but the Emperor's prestige and personal will, once expressed, sufficed to bring an orderly transition. To increase the appearance of direct rule, the Suzuki Cabinet was replaced by that of Prince Higashikuni Naruhiko.

Investigators concluded that neither atomic bomb nor Soviet entry was central to the decision to surrender, although they probably helped to advance the date. It was decided that submarine blockade of the Japanese islands had brought economic defeat by preventing exploitation of Japan's new colonies, sinking merchant tonnage, and convincing Japanese leaders of the hopelessness of the war. Bombing brought the consciousness of defeat to the people. Destruction of the Japanese Navy and Air Force jeopardized the home islands. Japan's largest armies, however, were never defeated, and this was responsible for the army's eagerness to fight on. Occupation found Japan's cities destroyed, its stockpiles exhausted, and its plants gutted. The government stood without prestige or respect. An alarming shortage of food and rising inflation threatened what remained of national strength. The time was ripe for changes.

After World War II. *SCAP and its objectives.* Gen. Douglas MacArthur, Supreme Commander, Allied Powers (SCAP), received his orders for the occupation of Japan through U.S. military channels; a Far Eastern Commission made up of Pacific war Allies was to make policy in Washington and provide consultation through an Allied Council for Japan, which sat in Tokyo. In fact the occupation became an American affair, and SCAP grew into a large headquarters. SCAP worked through the Japanese government. In the early years it provided direct instructions frequently, but with time suggestions were made more discreetly.

Purposes of the occupation

Occupation purposes had been held out in general terms in the Potsdam Declaration, with its promises of freedoms and statements of intent to remove undemocratic tendencies; those purposes were defined more precisely in a document that was worked out by the U.S. departments of state, war, and navy. Its emphases were on demilitarization, so that Japan would not again become a danger to peace; on democracy, so that (although the U.S. was not to impose any particular form of government) a responsible Japanese government would guard individual rights; and on encouragement of the Japanese to develop an economy that would be adequate for peacetime needs.

MacArthur responded enthusiastically to the idea of a demilitarized and democratic Japan and utilized the complex pattern of authority under which he functioned to ward off interference from Washington or from the Allies. He rushed constitutional reform to anticipate outside suggestions and first ignored and then delayed moves for partial Japanese rearmament after the Cold War changed U.S. priorities. The occupation measures created an open historical situation in which new forces could and did rise; SCAP measures proved lasting in cases where they coincided with trends already present within Japanese society, and those measures were vital to Japan's recovery as a free society and economy.

The early months of the occupation saw SCAP move swiftly to remove the principal supports of the militarist state. The armed forces were demobilized; State Shintō was disestablished; nationalist organizations were abolished and their members removed from important posts. Also removed from active roles were all persons prominent in wartime organizations and politics, including commissioned officers of the armed services and all high executives of the principal industrial firms. In Tokyo an international tribunal tried General Tōjō and other war leaders, sentencing seven to death, 16 to life imprisonment, and two to shorter terms. Millions of Japanese were repatriated from the former colonies and from Southeast Asia. The Home Ministry, which had controlled wartime Japan through its appointive governors and national police, was abolished, and the Education Ministry was deprived of its sweeping powers to control compulsory education. Because central control and military influence were being attacked by a military government that needed centralized powers in order to be effective, the occupation's role was often contradictory. Geography and economic rationality reinforced the logic of centralization, and many of the moves toward decentralization were modified or reversed a half decade later.

New constitution

Political reform. SCAP informed leading Japanese citizens that constitutional reform should receive first attention. Between October 1945 and February 1946 a Cabinet committee headed by Matsumoto Jōji prepared revisions of the Meiji Constitution, but the changes were few and superficial. MacArthur's government section rushed a new draft and submitted it to the Japanese government as a basis for further deliberations. Despite the misgivings of conservative statesmen, it was approved by the Emperor and submitted for amendment to the first postwar Diet, which had been elected in April 1946 (in these elections women had voted for the first time). The constitution, slightly modified, was promulgated on November 3, 1946, and went into effect on May 3, 1947. Its preface stated the intention of the Japanese people to ensure peaceful cooperation with all nations and the blessings of liberty for themselves and their descendants. The constitution included a 31-article bill of rights, and Article 9 renounced war as a "sovereign right of the nation" and pledged that "land, sea, and air forces, as well as other war potential, will never be maintained." The Emperor was described as the "symbol of the state and of the unity of the people, deriving his position from the will of the people with whom resides sovereign power." Earlier, on January 1, 1946, the Emperor had renounced claim to divinity. The constitution provided for a bicameral Diet, with the greatest power for a House of Representatives, the members serving four-year terms. The old peerage was dissolved and the House of Peers replaced by a House of Councillors, the members serving six-year terms. The prime minister was to be chosen by the Diet from its members, and an independent judiciary had the right of judicial review.

The new constitution thus reversed the Meiji pattern and contributed to responsible government by specifying the locus of executive authority. Despite its hasty preparation and foreign inspiration, it gained wide public support. Although the ruling conservatives desired to revise it after Japan regained its sovereignty in 1952, and an official commission favoured changes in 1964, the decreasing likelihood of mobilizing the two-thirds majority of the Diet necessary to secure approval for changes gradually made the possibility seem moot. In 1982, however, the election as prime minister of Nakasone Yasuhiro, who had long called for revision of Article 9 and preparation of a more "Japanese" constitution, once again revived discussions of revision.

By the time a peace treaty went into effect in 1952, elements of the political pattern had already changed, and subsequent governments showed their ability to modify by administrative actions a constitution that remained unchanged. Decentralization in some fields had proved expensive and inefficient. The police, for instance, while less centralized than in the days of the Home Ministry, had returned to a substantially national organization. Despite the announced goals of local decentralization, changing patterns of communications and administration had shown the logic of incorporating many small units of administration into larger units, a trend particularly marked in the countryside, where villages and towns merged to form a more rational tax structure. Article 9 had been compromised by a decision taken by SCAP to form the National Police Reserve of 75,000 men in 1950, during the Korean War. The force, later (1954) renamed the Self-Defense Forces, came to number about 240,000 by 1980.

Nevertheless, the basic principles of the constitution of 1947 enjoyed support among all factions in Japanese politics. Executive leadership was a chief asset of the new institutions. With the abolition of the competing forces that beset the premiers of the 1930s, the postwar prime ministers found themselves in charge of the administration and, with rearmament, of the armed forces as well. Thus, responsible leadership gradually replaced the ambiguous claims of Imperial rule of earlier days.

Economic and social changes. SCAP's political democratization was reinforced by economic and social changes designed to create interest groups that would use their new rights to protect the new political and economic structure. Changes in the countryside, in industry, and in social legislation all had the same purpose of breaking or weakening the old pattern of hierarchic control that had distinguished the "family-state" ideal of the Meiji leaders.

In agriculture the occupation established a program of land reform to convert tenants into owners. Tenancy had risen after World War I, and only ineffectual measures had been taken against it by the prewar government. Japan's wartime governments made important changes in land relationships. In their attempts to achieve national unity and equal sacrifice, they created agricultural associations to collect all rice. Absentee landlords received a lower rate of payment, and the tenant's relations with his landlord became much less important. Moreover, peasant sons in the armed services were able to send home part of their pay, while the shortage of labour made it possible for many members of farm families to secure gainful employment in factories. Thus, important preliminaries in rural well-being, not least among them the opportunities of the black market, took place prior to SCAP's instructions to the Japanese government to prepare a land-reform plan.

SCAP land reform

The government plan seemed inadequate to occupation authorities, and in the spring of 1946 a SCAP plan was drawn up; it became law in October. By its terms village and prefectural land commissions were elected with tenant, owner-farmer, and landlord representation to select land for purchase and eligible purchasers from among tenants. The government then bought the land at pre-inflation prices and sold it to the tenant. Four years later the reform had changed the ownership of more than two-thirds of Japan's cultivated acreage, and advantageous tax and price arrangements enabled the majority of new owners to pay for their land. The average family holding remained about 2½ acres (one hectare). In view of the larger population and the change in laws covering primogeniture, there was increased fragmentation of land, but the reform helped produce a striking rise in rural prosperity.

Initial Allied plans contemplated exacting heavy reparations from Japan, but the unsettled state of other Asian countries that were to have been recipients brought reconsideration. Except for Japanese assets overseas and a small number of war plants, reparations were very nearly limited to those worked out between Japan and its Asian victims after the peace treaty signed in San Francisco in 1951.

SCAP and the *zaibatsu*

Similar moderation marked the course of planning for deconcentration of the great *zaibatsu* firms. They at first were considered Japan's chief potential war makers but later came to be seen as essential elements in economic recovery. Of 1,200 concerns marked for investigation and possible dissolution in 1948, only 28 were broken up by SCAP, though the major units of the *zaibatsu* empires—holding companies—were dissolved and their securities made available for public purchase. New legislation sought to enforce fair trading and to guard against return to monopolies. Taxes on the profits of war wiped out many large fortunes and affected all large concentrations, while capital levy, inheritance, and graduated income taxes were designed to equalize the tax burden. The removal of wartime leaders from the business world prevented any action by the senior executives of 250 concerns during most of the occupation years. By 1950, extensive changes, although far short of those initially proposed, had taken place in the industrial world. The large banks, however, were not broken up and proved to be the centres for a measure of reconsolidation in the years after the occupation ended.

The balance of economic power was also affected by measures that produced a strong labour movement, which contested with management for political and economic primacy. After the Home and Welfare ministries were dissolved, a new Labour Ministry was established in 1947. All political prisoners, including the core of the Japan Communist Party, were released in the early months of the occupation. Most of those released turned their attention to organizing the labour movement, hoping to use it as a path to power. Laws on trade unions and labour relations, modelled on New Deal legislation in the United States, were passed, and soon a strong union movement appeared, led by men with political ambitions. When a general strike was announced for February 1947, with the avowed purpose of overthrowing the government, SCAP issued an injunction against it. Thereafter, occupation policy was concerned with reconstruction and no longer exclusively with liberation, and steps against inflation, political radicalism, and Communist control of labour unions followed. Under the Socialist Cabinet of Prime Minister Katayama Tetsu, labour education was emphasized, and in July 1948 SCAP ordered the government to take steps to deprive government workers—including those in communications unions—of the right to strike. A new labour organization, the General Council of Trade Unions of Japan (Sōhyō), was sponsored as a counterweight and gradual replacement for the Congress of Industrial Labour Unions of Japan (Sambetsu Kaigi), which had become dominated by the left. After 1951 Sōhyō too became increasingly anti-government and anti-American. Although some occupation measures deprived labour of useful weapons for fighting its way to power, the political strength of organized labour, expressed through the Japan Socialist Party (JSP), remained significantly different from what it had been before the war. The government's failure to carry through a police law designed to curb labour radicalism and sabotage in 1958 demonstrated the powerful support labour held in the Diet, the press, and public opinion.

Labour unions

The postwar social legislation saw energies and hopes long repressed by the Japanese government spring to full flower. The civil code, which reinforced the power of the male head of the family with numerous legal supports, was rewritten to allow for equality between the sexes and joint inheritance rights. Women were given the right to vote and to sit in the Diet. The abolition of the peerage, created in Meiji days, symbolized the modernization of society.

In the years after the peace treaty, which became effective on April 28, 1952, there were a number of changes in the pattern of occupation reforms. The Japanese government continued the emphasis on economic reconstruction that the occupation had inaugurated in 1948. The great labour offensives of the late 1940s had failed in their political objectives, but strikes throughout the 1950s continued to hamper industry. A new pattern of enterprise unions gradually brought peace to industry, while national federations became dominated by government workers. Many of the great firms united around banks whose credit was essential to their operation, and the government's power to allocate foreign currency, inherited from SCAP, gave it power to set and influence industrial policy. Although the land-reform system remained unaltered, rapid industrialization relieved pressure on the countryside as more and more Japanese moved to the cities. Electoral representation remained substantially unchanged, giving agricultural constituencies disproportionate influence in the politics of trade liberalization. Social legislation, however, had created such strong interest groups throughout Japanese society that substantial reversal of postwar changes did not take place. Japan soon entered a period of sustained prosperity and growth, and most Japanese credited this to the egalitarian reforms of the postsurrender years. For many, the new order was symbolized by the marriage of Crown Prince Akihito and a commoner, Shōda Michiko, in 1959.

International relations. Japan's return to international relations at the end of the occupation found it stripped of its conquests and deprived of some of its own territory. The Republic of China on Taiwan, the People's Republic of China on the mainland, the Republic of Korea (South Korea), and the Democratic People's Republic of Korea (North Korea) possessed military establishments far larger than Japan's Self-Defense Forces. International relations were not destined to be conducted on the pacifist lines envisioned by Article 9 of the constitution of 1947. The United States maintained its occupancy of Okinawa and the Ryukyus, while the Soviet Union occupied the entire

Kuril chain and reclaimed southern Sakhalin. The Korean War, which broke out in June 1950, increased the urgency of a peace treaty. Arrangements were worked out between the principal non-Communist allies before and during the command of Gen. Matthew B. Ridgway, who succeeded MacArthur as supreme commander in April 1951.

The San Francisco Conference that convened in September 1951 to sign the Japanese peace treaty ratified arrangements that had been worked out earlier by John Foster Dulles under the direction of Secretary of State Dean Acheson. Japan recognized the independence of Korea and renounced all rights to Taiwan and the Pescadores, the Kurils, and southern Sakhalin and gave up its rights in the Pacific islands to which it had held mandate under the League of Nations. The Soviet Union attended the San Francisco Conference, but it failed to make its objections to the treaty heard and consequently did not become a signatory. This enabled Japan to retain the hope of regaining at least the Kuril islands closest to Hokkaido—territory that it had not seized in war—through diplomatic efforts.

The San Francisco peace treaty recognized Japan's "right of individual and collective self-defense," which was exercised through a U.S.–Japan security pact whereby U.S. forces would remain until Japan could "assume responsibility for its own defense." Japan agreed not to grant similar rights to a third power without U.S. approval. U.S. assistance was extended to the Japanese defense forces, while U.S. units, except for air detachments and naval bases, were gradually removed to Okinawa.

Reparation treaties

The peace treaty made no arrangement for reparations for Japan's Pacific war victims but provided that Japan should negotiate with the countries concerned. Consequently, effective resumption of relations with the nations of Asia came only after treaties covering reparations had been worked out. These were signed with Burma in 1954, the Philippines in 1956, and Indonesia in 1958. In 1956 Japan restored diplomatic relations with the Soviet Union but without working out a formal peace treaty. In December 1956, with the Soviet Union no longer invoking a veto, Japan was admitted to the United Nations and subsequently became active in UN meetings and specialized agencies. It also became a contributing member of the Colombo Plan group of nations for economic development in South and Southeast Asia, of the General Agreement on Tariffs and Trade (GATT), and of the Organization for Economic Cooperation and Development (OECD). For many Japanese, their country's return to international status and eminence was symbolized by its holding the Olympic Games in 1964 and an international fair (Expo 70) at Ōsaka in 1970. Japan played a leading role in the creation of the Asian Development Bank in 1965–66.

At the time of the San Francisco treaty, Prime Minister Yoshida Shigeru had intended to delay committing Japan to either of the two Chinas, and the absence of both governments from San Francisco made this seem possible. But John Foster Dulles convinced Yoshida that the treaty would meet opposition in the U.S. Senate unless some assurance was given that Japan would recognize the Republic of China on Taiwan; thus, Tokyo soon negotiated a peace treaty with that regime but a treaty that did not prejudice possible subsequent negotiations with Peking. A lively trade developed between Japan and Taiwan, and Japanese contributions to the economy of Taiwan were considerable. The treaty also encouraged the development within Japan's Liberal-Democratic Party (LDP) of a so-called Taiwan lobby. Because Japan's relations with Peking remained tenuous, Chiang Kai-shek was for a time able to hold the Japanese government to its commitments by threatening to cut off Taiwan trade if Tokyo considered developmental loans to the mainland.

Chinese–Japanese trade

Mainland trade relationships developed slowly in the absence of political ties. In 1953 an unofficial trade pact was signed between private Japanese groups and authorities of the People's Republic, and the list of goods under embargo for mainland trade was gradually shortened. As late as 1972, however, 167 items remained on that list. In addition to unofficial agreements with Japanese firms designated as friendly by Peking, there developed in the 1960s an informal, semiofficial "memorandum" trade that became increasingly important. The Peking government made skillful use of trade for political purposes, in the hope of embarrassing or weakening Japan's conservative governments, and intervals of ideological tension on the mainland were usually reflected in declining trade with Japan. In 1958 China's Great Leap Forward campaign resulted in a temporary closure of all trade with Japan, and in the mid-1960s the Great Proletarian Cultural Revolution resulted in a severe decline. Nevertheless, Japan gradually became China's most important trading partner.

In 1971 Pres. Richard M. Nixon's announcement of a forthcoming visit to Peking produced a rapid growth in Japanese willingness to compromise ties with Taiwan in favour of closer relations with Peking. Peking also indicated new interest in formal relations with Japan, subject to the revocation of Japan's treaty with Taiwan. The People's Republic was admitted to the United Nations in 1971; in September 1972 Prime Minister Tanaka Kakuei reached agreement with Peking on steps to normalize relations, and simultaneously Japan severed its ties with Taiwan, replacing its embassy with a nonofficial office. Japan then vigorously pursued trade opportunities with the People's Republic, and in August 1978 the two countries concluded a Treaty of Peace and Friendship that bound both to "perpetual peace and friendship" and pledged them to oppose "hegemony" from whatever source and to foster economic and cultural relations. That same year, an eight-year agreement for a total of $20,000,000,000 in industrial contracts was signed. As Peking's program of modernization seemed to falter, however, Japanese investors moderated their expectations and showed renewed interest in Taiwan. Moderation in Peking and Taiwan made it possible to pursue both objectives. Japan also developed massive interests in South Korea, with which it had signed a treaty in 1965 that provided for reparations and opened the way to trade and investment.

Japanese–Chinese Treaty of Peace and Friendship

Postwar politics. After the surrender in 1945, Japanese politics at first returned to the pattern that had been interrupted by the militarist domination of national life. Extremists of the right were discredited by their identification with the lost war. Their major figures were removed from office or arrested, and until 1952, when all but those convicted by the international tribunal were permitted to resume their careers, little rightist organization was possible. Thereafter, some figures of the 1930s reemerged, but the rightists lacked unity and could offer no program of leadership in Asia. They were handicapped by a decline of support in the military and business sectors. Most important, rightist ideology found few listeners among the postwar generation accustomed to new freedoms. Except for a few spectacular incidents, such as the murder of the Socialist leader Asanuma Inajirō in 1960, rightist activities were limited to efforts to revive national holidays, such as February 11 (Foundation Day, for Emperor Jimmu, a campaign that succeeded in 1966), and demonstrations against the Soviet Union and China.

The left fared better for a time. With the release of political prisoners after the war, and with the repeal of the peace-preservation laws that had hampered political organization in prewar days, prominent Communist Party leaders returned to action. Land reform deprived them of an issue they had used elsewhere in Asia, but the postwar years, with their confusion and economic hardship, provided a favourable climate for Communists. Their high point at the polls came in the general election of 1949, when Communists placed 35 candidates in the House of Representatives and received nearly 10 percent of the vote.

On the outbreak of the Korean War in 1950, SCAP ordered Communist leaders removed from politics. Most chose to go underground, reappearing after the occupation ended. By that time popular sympathy with the Communist cause had declined markedly. The steady rise of living standards, the uncooperative attitude of the Soviet Union in negotiations over the Kuril Islands and in fishing-treaty discussions, a popular distaste for Communist opposition to the Imperial institution, and widespread dislike of the extremist tactics shown by leftist labour unions combined to create an unpromising climate for the Communists.

While not reversed, these trends were modified in the late

1960s. The spectacular economic growth of that decade produced great urban migrations that provided promising settings for mass organization and politics, and both conservative-religious (*e.g.,* Kōmeitō, or Clean Government Party, the political arm of the Sōka-gakkai movement) and radical political movements grew in strength. The conservative government's policy of giving priority to export development over social-welfare measures, though partly justified by a rising standard of living in which all groups shared, did little to alleviate the difficult conditions for many of the new recruits to the urban labour force. Communist leaders exploited their possibilities skillfully. In the mid-1960s they broke publicly with Peking to establish an autonomous and somewhat nationalist image, and their student organizations followed a relatively moderate line during anarchic disruptions at the universities in the late 1960s. These policies helped produce large pluralities in many urban elections, as in contests for the House of Councillors, where voter constituencies were large, but the Japan Communist Party remained far from power and found it difficult to establish satisfactory coalition arrangements with the more strongly pro-Peking Japan Socialist Party and other "reformist" elements.

The vicissitudes of right and left made it natural for the prewar moderates to dominate postwar politics. Career diplomats and bureaucrats possessed the command of English that enabled them to work with SCAP authorities, and, because they had been out of action since the 1930s, they had not become liable to the purge that removed militarists from office. Thus, figures of the 1920s and '30s reemerged, as did also the remnants of the party organizations of those years. The liaison agency, staffed largely by former diplomats, assumed immediate importance. The Cabinet that emerged shortly after the arrival of U.S. forces was headed by Shidehara Kijūrō, who was replaced in May 1946 by Yoshida Shigeru; both were diplomats. In 1947 and 1948 there was an interval of rule under Katayama Tetsu, a Socialist who headed a coalition Cabinet but who was unable to carry out a Socialist program. In 1948 Ashida Hitoshi held office for five months, after which Yoshida returned as prime minister and remained until December 1954, setting a record for modern Japanese prime ministers. Yoshida negotiated the peace treaty and the security pact in 1951 and set Japan's postsurrender course of close cooperation with the United States.

Formation of the Liberal-Democratic Party

Hatoyama Ichirō became a candidate for Yoshida's position, and the Liberal Party was split between their respective followers as a result. The San Francisco treaty and security pact split the Socialist Party into two factions also: the left opposing both the treaty (because it did not include the Communist countries) and the security pact with the United States, the right wing favouring the treaty while opposing the security pact. In October 1955 the Socialists reunited, and a month later the Liberals and Democrats reunited to form the Liberal-Democratic Party, which thereafter was the dominant party.

After independence. The Korean War marked the turn from depression to recovery for Japan. As the staging area for the UN effort in South Korea, the country profited from the many services it provided.

The return of independence in 1952 found the Japanese economy in the process of growth and change. Sustained prosperity and high growth rates changed all sectors of life. The countryside, where farmers had benefitted from land reform, began to feel the effects of small-scale mechanization and a consistent migration to industrial centres. Agricultural yields rose as improved strains of crops and modern technology were introduced, as household appliances appeared in remote villages, and as the cities' changing, more diversified patterns of food consumption provided a market for more cash crops, truck (market) garden fruits and vegetables, and meat products. Population control slowed the birth rate, and steady industrial growth brought full employment and even a labour shortage.

Structural changes in the economy

Particularly in the 1960s, the structure of the Japanese economy changed in order to concentrate on products of highly advanced technology, emphasizing Japan's need for stable, advanced trading partners instead of its earlier developed Asian markets for inexpensive textiles. Improvements in transportation—*e.g.,* cargo-handling methods and bulk transport by large ore carriers—were removing the disadvantage of the greater distances over which Japan's new materials were moving. Most important, a large and growing domestic market was rendering invalid earlier generalizations about Japan's need for cheap labour and captive Asian markets for inexpensive exports. Japan's economy grew rapidly from the 1960s to the "oil shock" of 1973 and thereafter continued to grow, though at a slower rate. Its output shifted with world currents, and its industrial development made it a world leader in shipbuilding, electronics, steel, automobiles, and high technology.

Japanese leaders attempting to raise the national income felt their options in international affairs severely restricted by the alliance with the United States, Taiwan, and South Korea, which prevented closer ties with the Soviet Union, China, and North Korea. Prosperity was a universal goal, but international politics proved sharply divisive. Public-opinion polls showed firm agreement against military power and continued horror of atomic or nuclear developments; and it was agreed that the Self-Defense Forces could not be utilized for international or UN causes.

U.S.–Japanese treaty revision

Restoration of relations with the Soviet Union and membership in the United Nations, both in 1956, were the principal efforts and achievements of Hatoyama Ichirō, who succeeded Yoshida in 1954. Hatoyama was followed by Ishibashi Tanzan in December 1956 and by Kishi Nobusuke in 1957. Kishi, who had been named, though not tried, as a war criminal because of his membership in the Tōjō Cabinet, was much criticized for his war record, but he continued the policies of cooperation with the United States that Yoshida had initiated. In 1958 the Peking government, occupied with its Great Leap Forward, closed all trade contacts with Japan. At the same time, revisions in the U.S.–Japan Treaty of Mutual Cooperation and Security were being discussed. A proposed treaty revision, to be in force for 10 years, alarmed many Japanese who had felt only slightly involved in the original agreement negotiated at the time of independence. Issues were further complicated by plans for a state visit by Pres. Dwight D. Eisenhower. Originally planned as a follow-up to a visit to Moscow, the visit changed drastically after the Soviet Union shot down a U.S. U-2 reconnaissance plane on May 1, 1960. Eisenhower's trip abroad had begun as a symbol of peace and coexistence but became a strategic tour of U.S. Pacific allies and bases. Its critics charged that it was designed to sustain the falling popularity of the Kishi government. After the Kishi Cabinet used its majority to force the treaty revisions through the Diet, opposition to the Prime Minister, the treaty, and the Eisenhower visit increased steadily. Gigantic student demonstrations shook Tokyo day after day. The treaty survived, but Eisenhower's visit was cancelled, and Kishi resigned in July 1960. In 1970, when the treaty had run its course, both governments were reluctant to see a repetition of the events of 1960 and agreed to invoke its provisions indefinitely, subject to one year's revocation by either party.

Ikeda, Satō, and their successors. Kishi was followed by Ikeda Hayato in 1960. A specialist in economic policy, Ikeda's goal of doubling national income in 10 years was more than met, as Japan's economy grew at rates of over 10 percent annually, the highest in the industrialized world. The U.S. administration of Pres. John F. Kennedy caught the imagination of many Japanese, and Kennedy's designation of the popular scholar Edwin O. Reischauer as ambassador further improved Japanese–American relations. By the late 1960s the unpopularity of the war in Vietnam threatened to disturb U.S. relations again.

The era of Satō

Satō Eisaku, who succeeded Ikeda in 1964, continued Ikeda's policies and proved an able and resourceful figure; he remained in office until 1972, thus eclipsing Yoshida's record. In 1965 Japan entered treaty relations with South Korea. At home, Satō worked for continued economic growth and tried to free Japan from reminders of defeat—the Bonin and Ryukyu islands, which had been left under U.S. occupation by the San Francisco treaty of peace, and U.S. military bases on Okinawa, essential to U.S. commitments to South Korea and an important link to U.S. forces in Vietnam. In the late 1960s, Japanese opposition to the Vietnam war made these bases highly objectionable. Satō's government secured the return of the Bonin Islands in 1967, and the retrocession of the Ryukyus became effective in 1972. U.S. bases on Okinawa were retained but were subject to the restrictions that affected other U.S. bases in Japan. Satō resigned in 1972 and in 1974 was awarded the Nobel Prize for Peace for his role in maintaining Japan's policy against nuclear weapons.

Tanaka scandals

Tanaka Kakuei, Satō's successor, seemed to promise a new stage of Japanese strength. One of his first acts was to travel to Peking, on the heels of U.S. Pres. Richard M. Nixon, to officially recognize the People's Republic. Tanaka reacted to increasing public concern with problems of pollution and overcrowding by calling for the redistribution of industry throughout the Japanese islands. Soon he was being charged with worsening inflation as land prices soared. More serious was the effect of the petroleum crisis of 1973 on a country completely dependent on imported oil. Outbreaks of panic buying by consumers brought reminders of the essential fragility of Japan's economic position; the rapid rise in the price of oil marked the end of an era of relatively cheap and abundant resources. Japan experienced the world recession of the 1970s, and its recovery seemed slower because of the previous years of exuberant growth. The Tanaka era ended in 1974 with a scandal based on irregularities in the accumulation of his private fortune. Shortly afterward Tanaka was implicated in improper use of official influence to bring about the selection by a Japanese airline of airplanes manufactured by Lockheed Aircraft Corp. (M.B.J.)

Tanaka was succeeded by Miki Takeo, the leader of a small faction in the governing Liberal-Democratic Party who won favour while larger factions were in disarray. Miki was determined to pursue the Lockheed affair, and a lengthy investigation and prosecution of Tanaka and his associates was begun. Tanaka was arrested in July 1976 and indicted on a charge of bribery; throughout the protracted court proceedings and even after his conviction in October 1983, Tanaka was a political power and remained so until 1987. Elections to the Diet in December 1976 brought a temporary end to the Liberal-Democrats' absolute majority, and Miki resigned. The new prime minister, Fukuda Takeo, had rich experience in many branches of government and was considered a specialist in economic policy. The problems of the Japanese economic turndown proved difficult, however, and combined with party factional differences to bring about Fukuda's defeat in 1978. He was succeeded by Ōhira Masayoshi, who announced his intention to continue Fukuda's foreign policies.

Ōhira died in office in 1980 and was succeeded by Suzuki Zenko, who in 1982 gave way to Nakasone Yasuhiro. Like their predecessors, each was a faction leader in the ruling party, but, unlike the others, Nakasone was a much more visible and outspoken leader, particularly on the matter of increasing Japan's defensive capability. In characteristic fashion, he took the unprecedented action of having his term extended by one year, and he became only the second postwar prime minister to name his successor, Takeshita Noboru, who took office in 1987.

In 1988 scandal again rocked the government when it was disclosed that the Recruit Cosmos Co., a data and real-estate conglomerate, had given money and sold stock not yet available to the public at prices well below expected market value to many bureaucrats and politicians—including Nakasone, Takeshita, and other prominent Liberal-Democrats—in order to gain influence. Largely as a result of the scandal and of a new and unpopular consumption (value-added) tax, the party's public approval plummeted, and in June 1989 Takeshita resigned. He was replaced by foreign minister Uno Sōsuke, who proclaimed as his main goal the restoration of public confidence in the party but was himself implicated in a sex scandal. In July the opposition parties outpolled the Liberal-Democrats in elections to the upper house and for the first time ever gained a majority in that body; in August, after Uno had resigned, the upper house selected Doi Takako, leader of the JSP, to be prime minister. Although Doi's nomination was overruled by the lower house—which chose Liberal-Democrat Kaifu Toshiki—it was the first time that a woman had been considered for prime minister and reflected a growing influence by women in Japanese politics.

The Shōwa era of Hirohito came to an end with his death in January 1989 and was followed by the Heisei ("Achieving Peace") era of his son Akihito. Hirohito's 62-year reign was the longest in recorded Japanese history.

After the 1970s Japan produced large surpluses in Japanese-American trade and became the largest or second largest trading partner of virtually every country with which it traded. It invested heavily in other countries, and its manufacturers built plants throughout the world.

Relations with China

Within Japanese politics and opinion, the long-standing polarization over Japan's treaty with the United States was made obsolete by the developments in Japanese relations with China—*i.e.,* vigorous trading and the signing of the Treaty of Peace and Friendship referred to above. Although opponents of the treaty had long argued that the American tie and the Japanese Self-Defense Forces threatened to cut Japan off from China, Peking saw both the treaty and forces as constraints against its Soviet rival and no longer criticized them. In fact, friendship with China had become compatible with alliance with the United States. The ideology of class warfare had lost relevance in the atmosphere of general affluence and middle-class consciousness. Decades of resistance to further rearmament were weakening because of resentment of the Soviet stance on the northern islands and because of increased U.S. pressure for Japan to have a greater share of the region's defense. At the same time, the new mood of national confidence made it difficult for leaders to secure agreement for further liberalization of imports into Japan, and charges of protectionism produced resentment of U.S. pressures. Japan's prosperity depended upon world trade, but protectionist sentiments grew against Japanese goods.

Thus, the Japanese became aware that the shibboleths of the postsurrender decades were obsolete. Japan was not poor but wealthy, not weak but a power in the international economy, and not isolated but the largest trading partner of almost every country in the world. This awareness was reflected in a surge of introspection in which writers discussed the role for their country and the nature of their society. What remained was a growing consensus around general principles, summed up by Prime Minister Fukuda in 1978: Japan should adhere to its decision not to become a major military power and should promote friendly cooperation with all nations and work to accept growing responsibility within the international community. Japan remained unusually dependent upon the stability of the world economy, but that stability in turn was more dependent upon the quality of Japanese participation than it had ever been before. (M.B.J./Ed.)

For later developments in the history of Japan, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 934, 96/10, and 975, and the *Index.*

BIBLIOGRAPHY

Physical and human geography: Two good gazetteers of Japan are *Nihon chimei jiten,* compiled by AKIRA WATANABE, 4 vol. (1954–56); and *Nihon chimei daijiten,* compiled by AKIRA WATANABE *et al.,* 7 vol. (1967–68). The *Nippon: A Charted Survey of Japan,* an industrial digest issued annually since 1936, contains physical and economic statistical data with explanations. Representative geographical works in English include *Regional Geography of Japan,* trans. from the Japanese, 6 vol. (1957), the proceedings of a conference of the International Geographical Union; and ROBERT B. HALL, JR., *Japan: Industrial Power of Asia,* 2nd ed. (1976), which contains a brief analysis of postwar industrial development. Geomorphology is covered by AKIRA WATANABE, "Landform Divisions of Japan," *Bull. Geogr. Surv. Inst., Tokyo,* 2:81–94 (1950–51); and by TORAO YOSHIKAWA, SOHEI KAIZUKA, and YOKO OTA, *The Landforms of Japan* (1981); climatology is dealt with by EIICHIRŌ FUKUI (ed.), *The Climate of Japan* (1977). TAKESHI MATSUI, "General Characteristics of the Soil Geography of Japan," *Pedorojisto/Pedologist,* 12:25–36 (1968), is useful; and *Japanese Cities: A Geographical Approach* (1970), published by the ASSOCIATION OF JAPANESE GEOGRAPHERS, discusses many pertinent subjects. BUREAU OF STATISTICS, *Statistical Handbook of Japan* (annual), contains official information, and *Japan Statistical*

Yearbook (annual) provides the best readily available information source on Japan as a whole. The MINISTRY OF FOREIGN AFFAIRS OF JAPAN, *Japan in Transition: One Hundred Years of Modernization* (1968, reissued 1975), gives general descriptions and is illustrated. HISAO AONO and SHŌHEI BIRUKAWA (eds.), *Nihon chishi* (1967–80), is a comprehensive series dealing with Japan's regional geography by prefectures. EDWARD A. ACKERMAN, *Japan's Natural Resources and Their Relation to Japan's Economic Future* (1953), is a basic source for understanding the economic development. For general background on the economy, see also GEORGE C. ALLEN, *Japan's Economic Recovery* (1958), and *Japan's Economic Expansion* (1965, reprinted 1969); ROBERT J. BALLON (ed.), *Doing Business in Japan,* 2nd ed., rev. (1968); RICHARD K. BEARDSLEY (ed.), *Studies on Economic Life in Japan* (1964); ALICE H. COOK, *An Introduction to Japanese Trade Unionism* (1966); WILLIAM W. LOCKWOOD (ed.), *The State and Economic Enterprise in Japan* (1965, reprinted 1969); YUTAKA MATSUMURA, *Japan's Economic Growth, 1945–1960* (1961); and TAKAFUSA NAKAMURA, *The Postwar Japanese Economy* (1981; originally published in Japanese, 1980). Detailed current information on various aspects of the economy may be found in publications of the Japanese government, such as the BANK OF JAPAN, *Economic Statistics of Japan* (annual); and the MINISTRY OF AGRICULTURE AND FORESTRY, *Abstract of Statistics on Agriculture, Forestry, and Fisheries* (annual). *Kodansha Encyclopedia of Japan,* 9 vol. (1983), provides a comprehensive compilation of information on Japan's history and its modern physical, social, political, and cultural environment. RYUZIRO ISIDA, *Geography of Japan* (1961, reprinted 1969), is concerned with major aspects of the physical, economic, and cultural environment. GLENN T. TREWARTHA, *Japan: A Geography* (1965), also deals with geographical aspects in detail; as does THE ASSOCIATION OF JAPANESE GEOGRAPHERS, *Geography of Japan* (1980). A detailed field survey of village life from geographical, historical, and social viewpoints is made in RICHARD K. BEARDSLEY, JOHN W. HALL, and ROBERT E. WARD, *Village Japan* (1959, reprinted 1969). Japan's cities are treated in detail in RONALD P. DORE, *City Life in Japan: A Study of a Tokyo Ward* (1958, reprinted 1965). J.D. BISIGNANI, *Japan Handbook* (1983), is a comprehensive travel guide. FREDERICA M. BUNGE (ed.), *Japan: A Country Study,* 4th ed. (1983), is a comprehensive work covering geographical, economic, social, and cultural aspects. See also *Grand Atlas of Japan* (1985), published by Heibonsha; and *The National Atlas of Japan* (1977), published by the GEOGRAPHICAL SURVEY INSTITUTE.

History: Among various critical guides in Western languages, JOHN W. HALL, *Japanese History: New Dimensions of Approach and Understanding,* 2nd ed. (1966), is a convenient and skilled presentation of scholarly work, and *Japanese History: A Guide to Japanese Reference and Research Materials* (1954, reprinted 1973); and HERSCHEL WEBB, *Research in Japanese Sources: A Guide* (1965), survey the extent of works on Japanese history by Japanese scholars. Additions to published materials will be found, though not annotated, in the annual bibliographical issue of the *Journal of Asian Studies.*

In addition to general surveys and monographs, the following three compendiums by groups of recognized scholars are suggested for advanced students who are accustomed to reading Japanese: KŌTA KODAMA (ed.), *Zusetsu Nihon bunka shi taikei* ("Illustrated Compendium of Japanese Cultural History"), 2nd ed., 14 vol. (1965–68), a collection of essays arranged in chronological order and with numerous illustrations for the general reading public, the last volume devoted to bibliography and related materials; SABURŌ IENAGA (ed.), *Iwanami Kōza: Nihon rekishi* ("Iwanami Lectures on Japanese History"), 23 vol. (1962–64), a collection of articles contributed by many authors and arranged in chronological order by periods, well-documented and directed to somewhat advanced readers; and KAWADE SHOBO (ed.), *Nihon no rekishi* ("A History of Japan"; 1965–67), composed of 27 volumes written by different authors with 4 volumes devoted to illustrations, in chronological order.

Authoritative accounts in English include JOHN K. FAIRBANK, EDWIN O. REISCHAUER, and ALBERT M. CRAIG, *East Asia: Tradition and Transformation,* rev. ed. (1973); and H. PAUL VARLEY, *Japanese Culture,* 3rd ed. (1984). Though dated, JAMES MURDOCH, *A History of Japan,* 3 vol. (1903–26, reprinted in 6 vol., 1964), is a pioneer work that gives a detailed political history. A later survey is GEORGE B. SANSOM, *A History of Japan,* 3 vol. (1958–63, reissued 1978). Shorter excellent interpretive works include SANSOM's *Japan: A Short Cultural History,* rev. ed. (1976), and *The Western World and Japan* (1950, reissued 1977); EDWIN O. REISCHAUER, *Japan: The Story of a Nation,* 3rd ed. (1981), and *The Japanese* (1977, reprinted 1981); SABURŌ IENAGA, *History of Japan,* 7th ed. (1963, reissued 1969); MIKISO HANE, *Japan: A Historical Survey* (1972); BRADLEY SMITH, *Japan: A History in Art* (1964); MITSUSADA INOUE, *Introduction to Japanese History: Before the Meiji Restoration,* rev. ed. (1968); WILLIAM G. BEASLEY, *The Modern History of Japan,* 3rd ed. (1981); and JOHN W. HALL, *Japan from Prehistory to Modern Times* (1970). The JAPAN. NATIONAL COMMISSION FOR UNESCO, *Japan: Its Land, People and Culture,* 3rd ed. (1973), is a comprehensive reference work.

For further reading, selected monographic works on Japan (in English) are listed below, grouped according to their historical approach. (*Ancient*): GERARD J. GROOT, *The Prehistory of Japan* (1951, reprinted 1972); C. MELVIN AIKENS and TAKAYASU HIGUCHI, *Prehistory of Japan* (1982), an inventory of archaeological sites and finds; J. EDWARD KIDDER, JR., *Japan Before Buddhism,* rev. ed. (1966); ROBERT KARL REISCHAUER, *Early Japanese History: c. 40 B.C.–A.D. 1167,* 2 vol. (1937, reissued 1967); IVAN I. MORRIS, *The World of the Shining Prince: Court Life in Ancient Japan* (1964, reissued 1979). (*Medieval and early modern*): G. CAMERON HURST III, *Insei: Abdicated Sovereigns in the Politics of Late Heian Japan, 1086–1185* (1976); JOHN W. HALL, *Government and Local Power in Japan, 500–1700* (1966, reissued 1980); KAN'ICHI ASAKAWA (ed.), *The Documents of Iriki, Illustrative of the Development of the Feudal Institutions of Japan,* rev. ed. (1955, reissued 1974); MINORU SHINODA, *The Founding of the Kamakura Shogunate, 1180–1185* (1960); JEFFREY P. MASS, *Warrior Government in Early Medieval Japan* (1974), and, with JOHN W. HALL (eds.), *Medieval Japan: Essays in Institutional History* (1974); JOHN W. HALL and TOYODA TAKESHI (eds.), *Japan in the Muromachi Age* (1977), the outcome of a binational conference held in 1973; DELMER M. BROWN, *Money Economy in Medieval Japan* (1951); YI-T'UNG WANG, *Official Relations Between China and Japan, 1368–1549* (1953); HELEN CRAIG MCCULLOUGH, *The Taiheiki: A Chronicle of Medieval Japan* (1959, reprinted 1976); CHARLES R. BOXER, *The Christian Century in Japan, 1549–1650* (1951, reprinted 1974), *Fidalgos in the Far East, 1550–1770,* 2nd ed. (1968), and *Jan Compagnie in Japan, 1600–1817,* 2nd ed. (1950, reprinted 1968); GEORGE ELISON, *Deus Destroyed: The Image of Christianity in Early Modern Japan* (1973); H. PAUL VARLEY, *The Ōnin War* (1967); JOHN W. HALL, NAGAHARA KEIJI, and KOZO YĀMAMURA (eds.), *Japan Before Tokugawa* (1981), the outcome of a 1977 conference; CONRAD TOTMAN, *Politics in the Tokugawa Bakufu, 1600–1843* (1967); JOHN W. HALL and MARIUS B. JANSEN (eds.), *Studies in the Institutional History of Early Modern Japan* (1968); SUSAN B. HANLEY and KOZO YAMAMURA, *Economic and Demographic Change in Preindustrial Japan, 1600–1868* (1978); HAROLD BOLITHO, *Treasures Among Men: The Fudai Daimyo in Tokugawa Japan* (1974); RONALD P. DORE, *Education in Tokugawa Japan* (1965); CHARLES D. SHELDON, *The Rise of the Merchant Class in Tokugawa Japan, 1600–1868* (1958, reprinted 1973); THOMAS C. SMITH, *The Agrarian Origins of Modern Japan* (1959, reissued 1966); JOHN W. HALL, *Tanuma Okitsugu, 1719–1788: Forerunner of Modern Japan* (1955, reprinted 1982); DONALD KEENE, *The Japanese Discovery of Europe, 1720–1830,* rev. ed. (1969); JOHN A. HARRISON, *Japan's Northern Frontier* (1953); GEORGE A. LENSEN, *The Russian Push Toward Japan* (1959, reissued 1971); WILLIAM G. BEASLEY (ed. and trans.), *Select Documents on Japanese Foreign Policy, 1853–1868* (1955, reprinted 1967); GRACE E. FOX, *Britain and Japan, 1858–1883* (1969); MARIUS B. JANSEN, *Sakamoto Ryōma and the Meiji Restoration* (1961, reprinted 1971); ALBERT M. CRAIG, *Chōshū in the Meiji Restoration* (1961, reissued 1967). (*Modern*): CHITOSHI YANAGA, *Japan Since Perry* (1949, reprinted 1975); E. HERBERT NORMAN, *Japan's Emergence as a Modern State* (1940, reprinted 1973); WILLIAM W. LOCKWOOD, *The Economic Development of Japan: Growth and Structural Change,* exp. ed. (1968, reprinted 1974); THOMAS C. SMITH, *Political Change and Industrial Development in Japan: Government Enterprise, 1868–1880* (1955, reissued 1974); WILLIAM CHAMBLISS, *Chiaraijima Village: Land Tenure, Taxation and Local Trade, 1818–1884* (1965); JOHANNES HIRSCHMEIER, *The Origins of Entrepreneurship in Meiji Japan* (1964); MARIUS B. JANSEN (ed.), *Changing Japanese Attitudes Toward Modernization* (1965, reprinted 1982); KENNETH B. PYLE, *The New Generation in Meiji Japan: Problems of Cultural Identity, 1885–1895* (1969); ROBERT A. WILSON, *Genesis of the Meiji Government in Japan, 1868–1871* (1957, reprinted 1980); JOSEPH PITTAU, *Political Thought in Early Meiji Japan, 1868–1889* (1967); GEORGE AKITA, *Foundations of Constitutional Government in Modern Japan, 1868–1900* (1967); ROGER F. HACKETT, *Yamagata Aritomo in the Rise of Modern Japan 1838–1922* (1971); ROBERT E. WARD (ed.), *Political Development in Modern Japan* (1968, reissued 1973); HILARY CONROY, *The Japanese Seizure of Korea, 1868–1910* (1960, reissued 1974); JOHN A. WHITE, *The Diplomacy of the Russo-Japanese War* (1964); SHUMPEI OKAMOTO, *The Japanese Oligarchy and the Russo-Japanese War* (1970); JAMES I. NAKAMURA, *Agricultural Production and the Economic Development of Japan, 1873–1922* (1966); TETSUO NAJITA, *Hara Kei in the Politics of Compromise, 1905–1915* (1967); PETER DUUS, *Party Rivalry and Political Change in Taishō Japan* (1968); ROBERT A. SCALAPINO, *Democracy and the Party Movement in Prewar Japan* (1953, reissued 1975); DELMER

M. BROWN, *Nationalism in Japan* (1955, reprinted 1971); AKIRA IRIYE, *After Imperialism: The Search for a New Order in the Far East, 1921–1931* (1965, reprinted 1978); JAMES B. CROWLEY, *Japan's Quest for Autonomy* (1966); FRANCIS C. JONES, *Japan's New Order in East Asia* (1954, reprinted 1978); JOHN M. MAKI, *Japanese Militarism* (1945); RICHARD STORRY, *The Double Patriots: A Study of Japanese Nationalism* (1957, reprinted 1973); SADAKO OGATA, *Defiance in Manchuria: The Making of Japanese Foreign Policy, 1931–1932* (1964, reprinted 1984); ROBERT J.C. BUTOW, *Tōjō and the Coming of the War* (1961, reissued 1969), and *Japan's Decision to Surrender* (1954); HERBERT FEIS, *The Road to Pearl Harbor* (1950, reissued 1971); THOMAS R.H. HAVENS, *Valley of Darkness: The Japanese People and World War Two* (1978). (*Postwar years*): BARON E.J. LEWE VAN ADUARD, *Japan: From Surrender to Peace* (1953); SUPREME COMMANDER FOR THE ALLIED POWERS, GOVERNMENT SECTION, *Political Reorientation of Japan: September 1945 to September 1948*, 2 vol. (1949, reprinted 1970); GEORGE C. ALLEN, *Japan's Economic Policy* (1980); SHIGERU YOSHIDA, *The Yoshida Memoirs*, trans. from the Japanese (1961, reissued 1973); FRANK GIBNEY, *Japan: The Fragile Superpower*, rev. ed. (1979); HUGH T. PATRICK and HENRY ROSOFSKY (eds.), *Asia's New Giant: How the Japanese Economy Works* (1976); CHALMERS JOHNSON, *MITI and the Japanese Miracle: The Growth of Industrial Policy, 1925–1975* (1982).

Japanese Literature

Both in quantity and quality, Japanese literature ranks as one of the major literatures of the world, comparable in age, richness, and volume to English literature, though its course of development has been quite dissimilar. The surviving works comprise a literary tradition extending from the 7th century AD to the present; during all of this time there was never a "dark age" devoid of literary production. Not only do poetry, the novel, and the drama have long histories in Japan, but some literary genres not so highly esteemed in other countries—including diaries, travel accounts, and books of random thoughts—are also prominent. A considerable body of writing by Japanese in the Chinese classical language, of much greater bulk and importance than comparable Latin writings by Englishmen, testifies to the Japanese literary indebtedness to China. Even the writings entirely in Japanese present an extraordinary variety of styles, which cannot be explained merely in terms of the natural evolution of the language. Some styles were patently influenced by the importance of Chinese vocabulary and syntax; but others developed in response to the internal requirements of the various genres, whether the terseness of haiku (a poem in 17 syllables) or the bombast of the dramatic recitation.

This article is divided into the following sections:

GENERAL CONSIDERATIONS

Problems in reading Japanese literature

The difficulties of reading Japanese literature can hardly be exaggerated; even a specialist in one period is likely to have trouble deciphering a work from another period or genre. Japanese style has always favoured ambiguity, and the particles of speech necessary for easy comprehension of a statement are often omitted as unnecessary or as fussily precise. Sometimes the only clue to the subject or object of a sentence is the level of politeness in which the words are couched; for example, the verb *mesu* (meaning "to eat," "to wear," "to ride in a carriage," etc.) designates merely an action performed by a person of quality. In many cases, ready comprehension of a simple sentence depends on a familiarity with the background of a particular period of history. The verb *miru,* "to see," had overtones of "to have an affair with" or even "to marry" during the Heian period in the 10th and 11th centuries, when men were generally able to see women only after they had become intimate. The long period of Japanese isolation in the 17th and 18th centuries also tended to make the literature provincial, or intelligible only to persons sharing a common background; the phrase "some smoke rose noisily" (*kemuri tachisawagite*), for example, was all readers of the late 17th century needed to realize that an author was referring to the Great Fire of 1682 that ravaged the shogunal capital of Edo (the modern city of Tokyo).

Despite the great difficulties arising from such idiosyncrasies of style, Japanese literature of all periods is exceptionally appealing to modern readers, whether read in the original or in translation. Because it is prevailingly subjective and coloured by an emotional rather than an intellectual or moralistic tone, its themes have a universal quality almost unaffected by time. To read a diary by a court lady of the 10th century is still a moving experience, because she described with such honesty and intensity her deepest feelings that the modern reader forgets the chasm of history and changed social customs separating her world from his own.

The "pure" Japanese language, untainted and unfertilized by Chinese influence, contained remarkably few words of an abstract nature. Just as English borrowed such words as morality, honesty, justice, and the like from the Continent, the Japanese borrowed these terms from China; but if the Japanese language was lacking in the vocabulary appropriate to a Confucian essay, it could express almost infinite shadings of emotional content. A Japanese poet who was dissatisfied with the limitations imposed by his native language or who wished to describe unemotional subjects—whether the quiet outing of aged gentlemen to a riverside or the poet's awareness of his insignificance as compared to the grandeur of the universe—naturally turned to writing poetry in Chinese. From the 16th century on, many words that had been excluded from poetry because of their foreign origins or their humble meanings, following the dictates of the codes of poetic diction established in the 10th century, were adopted by the practitioners of the haiku, originally an iconoclastic, popular verse form. For the most part, however, the Japanese writers, far from feeling dissatisfied with the limitations on expression imposed by their language, were convinced that virtuoso perfection in phrasing and an acute refinement of sentiment were more important to poetry than the voicing of intellectually satisfying concepts.

Effect of the language on literary forms

The Japanese language itself also shaped poetic devices and forms. Because it lacks a stress accent, meaningful rhymes (all words end in one of five simple vowels), or quantity, poetry was distinguished from prose mainly in that it consisted of alternating lines of five and seven syllables; however, if the intensity of emotional expression was low, this distinction alone could not save a poem from dropping into prose. The difficulty of maintaining a high level of poetic intensity may account for the preference for short verse forms that could be polished with perfectionist care. But however moving a tanka (verse in 31 syllables) is, it clearly cannot fulfill some of the functions of longer poetic forms; and there are no Japanese equivalents of *Paradise Lost, The Rape of the Lock,* or *Tintern Abbey.* Instead, the poets devoted their efforts to perfecting each syllable of their compositions, expanding the content of a tanka by suggestion and allusion and prizing shadings of tone and diction more than originality or boldness of expression.

The fluid syntax of the prose affected not only style but content as well. Japanese sentences are sometimes of inordinate length, responding to the subjective turnings and twistings of the author's thought; and the writers considered smooth transitions from one statement to the next, rather than structural unity, the mark of excellent prose. The longer works accordingly betray at times a lack of overall structure of the kind associated in the West with Greek concepts of literary form but consist instead of episodes linked chronologically or by other associations. The difficulty experienced by Japanese writers in organizing their impressions and perceptions into sustained works may explain the development of the diary and travel account, genres in which successive days or the successive stages of a journey provide a structure for otherwise unrelated descriptions. Japanese literature contains some of the world's longest novels and plays; but its genius is most strikingly displayed in the shorter works, whether the tanka, the haiku, the Nō plays, or the poetic diaries.

An acute literary sensibility, fostered especially by the traditions of the court, encouraged the creation of "codes" of poetic practice and of a considerable body of criticism, extending back to the 10th century, that was usually composed by the leading poets or dramatists themselves. These codes exerted an inhibiting effect on new forms of literary composition, but they also helped to preserve a distinctively aristocratic tone.

Relation to Chinese and Korean literature

Japanese literature absorbed much direct influence from China, but the characteristic literary works are strikingly dissimilar. The tradition of feminine writing, especially of such introspective works as diaries, gave a colouring to Japanese prose quite unlike the more objective, masculine Chinese writings. Although the Japanese have been criticized for their imitations of Chinese examples (even by some Japanese), the Japanese novel in fact antedates any Chinese novels by centuries; and the theatre developed quite independently. Because the Chinese and Japanese languages are unrelated, the poetry naturally took different forms, although Chinese poetic examples and literary theories were often in the minds of the Japanese poets. Japanese and Korean are probably related languages, but Korean literary influence was negligible, though Koreans served an important function in transmitting Chinese literary and philosophical works to Japan. Poetry and prose written in the Korean language were unknown to the Japanese until relatively modern times.

From the 8th to the 19th century Chinese literature enjoyed greater prestige among educated Japanese than their own; but a love for the Japanese classics, especially those composed at the court in the 10th and 11th centuries, gradually spread among the entire people and influenced literary expression in every form, even the songs and tales composed by humble people totally removed from the aristocratic world portrayed in classical literature.

HISTORY

Origins. The first writing of literature in Japanese was occasioned by influence from China. The Japanese were still comparatively primitive and without writing when, in the first four centuries AD, knowledge of Chinese civilization gradually reached them. They rapidly assimilated much of this civilization, and the Japanese scribes adopted Chinese characters as a system of writing, although an alphabet (if one had been available to them) would have been infinitely better suited to the Japanese language. The characters, first devised to represent Chinese monosyllables, could be used only with great ingenuity to represent the agglutinative forms of the Japanese language. The ultimate results were chaotic, giving rise to one of the most complicated systems of writing ever invented. The use of Chinese characters enormously influenced modes of expression and led to an association between literary composition and calligraphy lasting many centuries.

Early writings. The earliest Japanese texts were written in Chinese because no system of transcribing the sounds and grammatical forms of Japanese had been invented. The oldest known inscription, on a sword that dates from about AD 440, already showed some modification of normal Chinese usage in order to transcribe Japanese names and expressions. The most accurate way of writing Japanese words was by using Chinese characters not for their meanings but for their phonetic values, giving each character a pronunciation approximating that used by the Chinese themselves. In the oldest extant works, the *Kojiki* (712; "Records of Ancient Matters") and *Nihon shoki*, or *Nihon-gi* (720; "Chronicles of Japan"), more than 120 songs, some perhaps dating back to the 5th century AD, are given in phonetic transcription, doubtless because the Japanese attached great importance to the sounds themselves. In these two works, both officially commissioned "histories" of Japan, many sections are written entirely in Chinese; but parts of the *Kojiki* were composed in a complicated mixture of language that made use of the Chinese characters sometimes for their meaning and sometimes for their sound. (D.Ke.)

The Kojiki and Nihon shoki as collections of myths. Most of the surviving Japanese myths are recorded in these two works. They tell of the origin of the ruling class and were apparently aimed at strengthening its authority. Therefore, they are not pure myths but have much political colouring. They are based on two main traditions: the Yamato Cycle, centred around the sun goddess Amaterasu Ōmikami, and the Izumo Cycle, in which the principal character is Susanoo (or Susanowo) no Mikoto, the brother of Amaterasu.

Myths and genealogies

Genealogies and mythological records were kept in Japan, at least from the 6th century AD and probably long before that. By the time of the emperor Temmu (7th century), it became necessary to know the genealogy of all important families in order to establish the position of each in the eight levels of rank and title modelled after the Chinese court system. For this reason, Temmu ordered the compilation of myths and genealogies that finally resulted in the *Kojiki* and *Nihon shoki.* The compilers of these and other early documents had at their disposal not only oral tradition but also documentary sources. A greater variety of sources was available to the compiler of the *Nihon shoki.* While the *Kojiki* is richer in genealogy and myth, the *Nihon shoki* adds a great deal to scholarly understanding of both the history and the myth of early Japan. Its purpose was to give the newly Sinicized court a history that could be compared with the annals of the Chinese.

Cosmology

The purpose of the cosmologies of the *Kojiki* and *Nihon shoki* is to trace the Imperial genealogy back to the foundation of the world. The myths of the Yamato Cycle figure prominently in these cosmologies. In the beginning, the world was a chaotic mass, an ill-defined egg, full of seeds. Gradually, the finer parts became heaven (Yang), the heavier parts earth (Yin). Deities were produced between the two: first, three single deities, and then a series of divine couples. According to the *Nihon shoki,* one of the first three "pure male" gods appeared in the form of a reed that connected heaven and earth. A central foundation was now laid down for the drifting cosmos, and mud and sand accumulated upon it. A stake was driven in, and an inhabitable place was created. Finally, the god Izanagi (He Who Invites) and the goddess Izanami (She Who Invites) appeared. Ordered by their heavenly superiors, they stood on a floating bridge in heaven and stirred the ocean with a spear. When the spear was pulled up, the brine dripping from the tip formed Onogoro, an island that became solid spontaneously. Izanagi and Izanami then descended to this island, met each other by circling around the celestial pillar, discovered each other's sexuality, and began to procreate. After initial failures, they produced the eight islands that now make up Japan. Izanami finally gave birth to the god of fire and died of burns. Raging with anger, Izanagi attacked his son, from whose blood such deities as the god of thunder were born. Other gods were born of Izanami on her deathbed. They presided over metal, earth, and agriculture. In grief, Izanagi pursued Izanami to Yomi (analogous to Hades) and asked her to come back to the land of the living. The goddess replied that she had already eaten food cooked on a stove in Yomi and could not return. In spite of her warning, Izanagi looked at his wife and discovered that her body was infested with maggots. The angry and humiliated goddess then chased Izanagi from the underworld. When he finally reached the upper world, Izanagi blocked the entrance to the underworld with an enormous stone. The goddess then threatened Izanagi, saying that she would kill a thousand people every day. He replied that he would father one thousand and five hundred children for every thousand she killed. After this, Izanagi pronounced the formula of divorce.

Izanagi then returned to this world and purified himself from the miasma of Yomi no Kuni. From the lustral water falling from his left eye was born the sun goddess Amaterasu Ōmikami, ancestress of the Imperial family. From his right eye was born the moon god Tsukiyomi no Mikoto and from his nose, the trickster god Susanoo. Izanagi gave the sun goddess a jewel from a necklace and told her to govern heaven. He entrusted the dominion of night to the moon god. Susanoo was told to govern the sea. According to the *Kojiki,* Susanoo became dissatisfied with his share and ascended to heaven to see his older sister. Amaterasu, fearing his wild behaviour, met him and suggested that they prove their faithfulness to each other by bringing forth children. They agreed to receive a seed from each other, chew it, and spit it away. If gods rather than goddesses were born, it would be taken as a sign of the good faith of the one toward the other. When Susanoo brought forth gods, his faithfulness was recognized, and he was permitted to live in heaven.

Susanoo trickster myth

Susanoo, becoming conceited over his success, began to play the role of a trickster. He scattered excrement over the dining room of Amaterasu, where she was celebrating the ceremony of the first fruits. His worst offense was to fling into Amaterasu's chamber a piebald horse he had "flayed with a backward flaying" (a ritual offense).

Enraged at the pranks of her brother, the sun goddess hid herself in a celestial cave, and darkness filled the heavens and the earth. The gods were at a loss. Finally, they gathered in front of the cave, built a fire, and made cocks crow. They erected a sacred evergreen tree, and from its branches they hung curved beads, mirrors, and cloth offerings. A goddess named Amenouzume no Mikoto then danced half-nude. Amaterasu, hearing the multitudes of gods laughing and applauding, became curious and opened the door of the cave. Seizing the opportunity, a strong-armed god dragged her out of the cave.

The Izumo Cycle

The myths of the Izumo Cycle then begin to appear in the narration. Having angered the heavenly gods and having been banished from heaven, Susanoo descended to Izumo, where he rescued Princess Marvellous Rice Field (Kushiinada Hime) from an eight-headed serpent. He then married the Princess and became the progenitor of the ruling family of Izumo. The most important member of the family of Susanoo was the god Ōkuninushi no Mikoto, the great earth chief, who assumed control of this region before the descent to earth of the descendants of the sun goddess.

Before long, Amaterasu, the leader of the celestial gods—the gods of Izumo were known as earthly gods—asked Ōkuninushi to turn over the land of Izumo, saying that "the land of the plentiful reed-covered plains and fresh rice ears" was to be governed by the descendants of the heavenly gods. After the submission of Izumo, Amaterasu made her grandson Ninigi no Mikoto (*ninigi* is said to represent rice in its maturity) descend to earth. According to the *Nihon shoki,* Amaterasu handed Ninigi some ears of rice from a sacred rice field and told him to raise rice on earth and to worship the celestial gods. The grandson of the sun goddess then descended to the peak of Takachiho (meaning high thousand ears) in Miyazaki, Kyushu. There he married a daughter of the god of the mountain, named Konohana-sakuya Hime (Princess Blossoms of the Trees).

When Ninigi's wife became pregnant and was about to give birth, all in a single night, he demanded proof that the child was his. She accordingly set fire to her room, then safely produced three sons. One of them, in turn, became the father of the legendary first emperor, Jimmu, who is considered to mark the watershed between the "age of the gods" and the historical age; but Jimmu's eastern expedition and conquest of the Japanese heartland was also a myth. (N.M./D.Ke.)

Origin of the tanka in the Kojiki. The myths in the *Kojiki* are occasionally beguiling, but the only truly literary parts of the work are the songs. The early songs lack a fixed metrical form; the lines, consisting of an indeterminate number of syllables, were strung out to irregular lengths, showing no conception of poetic form. Some songs, however, seem to have been reworked—perhaps when the manuscript was transcribed in the 8th century—into what became the classic Japanese verse form, the tanka (short poem), consisting of five lines of five, seven, five, seven, and seven syllables. Various poetic devices employed in these songs, such as the *makura kotoba* ("pillow word"), a kind of fixed epithet, remained a feature of later poetry.

Altogether, some 500 primitive songs have been preserved in various collections. Many describe travel, and a fascination with place-names, evident in the loving enumeration of mountains, rivers, and towns with their mantic epithets, was developed to great lengths in the gazetteers (*fudoki*) compiled at the beginning of the 8th century. These works, of only intermittent literary interest, devote considerable attention to the folk origins of different place-names, as well as to other local legends.

The significance of the Man'yōshū. A magnificent anthology of poetry, the *Man'yōshū* (compiled after 759; "Collection of Ten Thousand Leaves"), is the single great literary monument of the Nara period (710–784), although it includes poetry written in the preceding century, if not earlier. Most of the 4,500 or so poems are tankas; but the masterpieces of the *Man'yōshū* are the 260 *chōka* ("long poems"), ranging up to 150 lines in length and cast in the form of alternating lines in five and seven syllables followed by a concluding line in seven syllables. The amplitude of the *chōka* permitted the poets to treat themes impossible within the compass of the tanka—whether the death of a wife or child, the glory of the Imperial family, the discovery of a gold mine in a remote province, or the hardships of military service.

Kakinomoto Hitomaro

The greatest of the *Man'yōshū* poets, Kakinomoto Hitomaro, served as a kind of poet laureate in the late 7th and early 8th centuries, accompanying the sovereigns on their excursions and composing odes of lamentation for deceased members of the Imperial family. Modern scholars have suggested that the *chōka* may have originated as exorcisms of the dead, quieting the ghosts of recently deceased persons by reciting their deeds and promising that they will never be forgotten. Some of Hitomaro's masterpieces describe the glories of princes or princesses he may never have met so convincingly as to transcend any difference between "public" expressions of grief and his private feelings. Hitomaro's *chōka* are unique in Japanese poetry thanks to their superb combination of imagery, syntax, and emotional strength; they are works of truly masculine expression. He showed in his tanka, however, that he was also capable of the evocative, feminine qualities typical of later Japanese poetry.

The *chōka* often concluded with one or more *hanka* ("envoys") that resumed central points of the preceding poem. The *hanka* written by the 8th-century poet Yamabe Akahito are so perfectly conceived as to make the *chōka* they follow at times seem unnecessary; the concision and evocativeness of these poems, identical in form with the tanka, are close to the ideals of later Japanese poetry. Nevertheless, the supreme works of the *Man'yōshū* are the *chōka* of Hitomaro, Ōtomo Tabito, Ōtomo Yakamochi (probably the chief compiler of the anthology), and Yamanoue Okura. The most striking quality of the *Man'yōshū* is its powerful sincerity of expression. The poets were certainly not artless songsmiths exclaiming in wonder over the beauties of nature, a picture that is often painted of them by sentimental critics; but their emotions were stronger and more directly expressed than in later poetry. The corpse of an unknown traveller, rather than the falling of the cherry blossoms, stirred in Hitomaro an awareness of the uncertainty of human life.

The *Man'yōshū* is exceptional in the number of poems composed outside the court, whether by frontier guards or persons of humble occupation. Perhaps some of these poems were actually written by courtiers in the guise of commoners, but the use of dialect and familiar imagery contrasts with the strict poetic diction imposed in the 10th century. The diversity of themes and poetic forms also distinguishes the *Man'yōshū* from the more polished but narrower verse of later times. In Okura's famous "Dialogue on Poverty," for example, two men—one poor and the other destitute—describe their miserable lots, revealing a concern over social conditions that would be absent from the classical tanka. Okura's visit to China early in the 8th century, as the member of a Japanese embassy, may account for Chinese influence in his poetry. His poems are also prefaced in many instances by passages in Chinese stating the circumstances of the poems or citing Buddhist parallels.

The *Man'yōshū* was transcribed in an almost perversely complicated system that used Chinese characters arbitrarily, sometimes for meaning and sometimes for sound. The lack of a suitable script probably inhibited literary production in Japanese during the Nara period. The growing importance, however, of Chinese poetry as the mark of literary accomplishment in a courtier may also have interrupted the development of Japanese literature after its first flowering in the *Man'yōshū.*

Eighteen *Man'yōshū* poets are represented in the collection *Kaifūsō* (751), an anthology of poetry in Chinese composed by members of the court. These poems are little more than pastiches of ideas and images borrowed directly from China; the composition of such poetry reflects the enormous prestige of Chinese civilization at this time.

Classical literature: Heian period (794–1185). The foundation of the city of Heian-kyō (later known as Kyōto) as the capital of Japan marked the beginning of a period of great literary brilliance. The earliest writings of the period, however, were almost all in Chinese because of the continued desire to emulate the culture of the continent. Three Imperially sponsored anthologies of Chinese poetry appeared between 814 and 827, and it seemed for a time that writing in Japanese would be relegated to an extremely minor position. The most distinguished writer of Chinese verse, the 9th-century poet Sugawara Michizane, gave a final lustre to this period of Chinese learning by his erudition and poetic gifts; but his refusal to go to China when offered the post of ambassador, on the grounds that China no longer had anything to teach Japan, marked a turning point in the response to Chinese influence.

The first major work in *kana*

Poetry. The invention of the *kana* phonetic syllabary, traditionally attributed to the 9th-century Shingon priest and Sanskrit scholar Kūkai, enormously facilitated writing in Japanese. Private collections of poetry in *kana* began to be compiled about 880; and in 905 the *Kokinshū* ("Collection from Ancient and Modern Times"), the first major work of *kana* literature, was compiled by the poet Ki Tsurayuki and others. This anthology contains 1,111 poems divided into 20 books arranged by topics, including six books of seasonal poems, five books of love poems, and single books devoted to such subjects as travel, mourning, and congratulations. The two prefaces are clearly indebted to the theories of poetry described by the compilers of such Chinese anthologies as the *Shih Ching* and *Wen hsüan,* but the preferences they express would be shared by most tanka poets for the next 1,000 years. The preface by Tsurayuki, the oldest work of sustained prose in *kana,* enumerated the circumstances that move men to write poetry; he believed that melancholy, whether aroused by a change in the seasons or by a glimpse of white hairs reflected in a mirror, provided a more congenial mood for writing poetry than the harsher emotions treated in the *Man'yōshū.* The best tanka in the *Kokinshū* captivate the reader by their perceptivity and tonal beauty, but these flawlessly turned miniatures obviously lack the variety of the *Man'yōshū.*

Skill in composing tanka became an asset in gaining preference at court; it was also essential to a lover, whose messages to his mistress (who presumably could not read Chinese, still the language employed by men in official documents) often consisted of poems describing his own emotions or begging her favours. In this period the tanka almost completely ousted the *chōka* because the shorter poems were more suited to the lover's billet-doux or to competitions on prescribed themes.

For the poets of the *Kokinshū* and the later court anthologies, originality was less desirable than perfection of language and tone. The critics, far from praising novelty of effects, condemned deviations from the standard poetic diction (established by the *Kokinshū*) of some 2,000 words and insisted on absolute adherence to the poetic conventions. Although these restrictions saved Japanese poetry from lapses into bad taste or vulgarity, they froze it for centuries in prescribed modes of expression. Only a skilled critic can distinguish a tanka of the 10th century from one of the 18th century. The *Kokinshū* set the precedent for later court anthologies, and a knowledge of its contents was indispensable to all poets as a guide and source of literary allusions.

Love poetry occupies a prominent place in the *Kokinshū,* but the joys of love are seldom celebrated; instead, the poets wrote in the melancholy vein prescribed in the preface, describing the uncertainties before a meeting with the beloved, the pain of parting, or the sad realization that an affair had ended. The invariable perfection of diction, unmarred by any indecorous cry from the heart, may sometimes make one doubt the poet's sincerity. This is not true of the great *Kokinshū* poets of the 9th century—Ono Komachi, Lady Ise, Ariwara Narihira, and Tsurayuki himself—but even Buddhist priests, who presumably had renounced carnal love, wrote love poetry at the court competitions; and it is hard to detect any difference between such poems and those of sincere lovers.

The preface of the *Kokinshū* lists judgments on the principal poets of the collection. This criticism is unsatisfying to a modern reader because it is so terse and unanalytical; but it nevertheless marks a beginning of Japanese poetic criticism, an art that developed impressively during the course of the Heian period.

The literary diary

Prose. Ki Tsurayuki is celebrated also for his *Tosa nikki* (935; *The Tosa Diary*), the account of his homeward journey to Kyōto from the province of Tosa, where he had served as governor. Tsurayuki wrote this diary in Japanese, though men at the time normally kept their diaries in Chinese (perhaps it was in order to escape reproach for adopting this unmanly style that he pretended a woman in the governor's entourage was the author). Events of the journey are interspersed with the poems composed on various occasions. The work is affecting especially because of the repeated, though muted, references to the death of Tsurayuki's daughter in Tosa.

Tosa nikki is the earliest example of a literary diary. Although Tsurayuki pretended that it was written by a woman, the later Heian diarists who wrote in the Japanese language were, in fact, court ladies; their writings include some of the supreme masterpieces of the literature. *Kagerō nikki* (*The Gossamer Years*) describes the life between 954 and 974 of the second wife of Fujiwara Kaneie, a prominent court official. The first volume, related long after the events, is in the manner of an autobiographical novel; even the author confesses that her remembrances are probably tinged with fiction. The second two volumes approach a true diary, with some entries apparently made on the days indicated. The writer (known only as "the mother of Michitsuna") describes, with many touches of self-pity, her unhappy life with her husband. She evidently assumed that readers would sympathize, and often this is the case, though her self-centred complaints are not endearing. In one passage, in which she gloats over the death of a rival's child, her obsession with her own griefs shows to worst advantage; yet her journal is extraordinarily moving precisely because the author dwells exclusively on universally recognizable emotions and omits the details of court life that must have absorbed the men.

Other diaries of the period include the anecdotal *Murasaki Shikibu nikki* ("The Diary of Murasaki Shikibu"; Eng. trans., *Murasaki Shikibu: Her Diary and Poetic Memoirs*), at once an absorbing literary work and a source of information on the court life the author (Murasaki Shikibu) described more romantically in her masterpiece *Genji monogatari* (*c.* 1010; *The Tale of Genji*), and *Izumi Shikibu nikki* (*The Diary of Izumi Shikibu*), which is less a diary than a short story liberally ornamented with poetry.

These "diaries" are closely related in content and form to the *uta monogatari* ("poem tales") that emerged as a literary genre later in the 10th century. *Ise monogatari* (*c.* 980; *Tales of Ise*) consists of 143 episodes, each containing one or more poems and an explanation in prose of the circumstances of composition. The brevity and often the ambiguity of the tanka gave rise to a need for such explanations, and when these explanations became extended or (as in the case of *Ise monogatari*) were interpreted as biographical information about one poet (Ariwara Narihira), they approached the realm of fiction.

Along with the poem tales, there were works of religious or fanciful inspiration going back to *Nihon ryōiki* (822; *Miraculous Stories from the Japanese Buddhist Tradition*), an account of Buddhist miracles in Japan compiled by the priest Kyōkai. These stories, written in Chinese, were probably used as a source of sermons by the priests with the intent of persuading ordinary Japanese, incapable of reading difficult works of theology, that they must lead virtuous lives if they were not to suffer in hell for present misdeeds. No such didactic intent is noticeable in *Taketori monogatari* (10th century; *Tale of the Bamboo Cutter*), a fairy tale about a princess who comes from the moon to dwell on earth in the house of a humble bamboo cutter; the various tests she imposes on her suitors, fantastic though they are, are described with humour and realism.

Early development of the novel

The first lengthy "novel," *Utsubo monogatari* ("The Tale of the Hollow Tree"), was apparently written between 956 and 983 by Minamoto Shitagō, a distinguished courtier and scholar. This uneven, ill-digested work is of interest chiefly as an amalgam of elements in the poem tales and fairy tales; it contains 986 tanka, and its episodes range from early realism to pure fantasy.

The contrast between this crude work and the sublime *Genji monogatari* is overwhelming. Perhaps the difference is best explained in terms of the feminine traditions of writing, exemplified especially by the diaries, which enabled Murasaki Shikibu to discover depths in her characters unsuspected by the male author of *Utsubo monogatari.* The *Genji monogatari* is the finest work not only of the Heian period but of all Japanese literature and merits being called the first important novel written anywhere in the world. *Genji monogatari* was called a work of *mono no aware* ("a sensitivity to things") by the great 18th-century literary scholar Motoori Norinaga; the hero, Prince Genji, is not remarkable for his martial prowess or his talents as a statesman but as an incomparable lover, sensitive to each of the many women he wins. The story is related in terms of the successive women Genji loves; each of them evokes a different response from this marvellously complex man. The last third of the novel, describing the world after Genji's death, is much darker in tone; and the principal figures, though still impressive, seem no more than fragmentations of the peerless Genji.

The success of *Genji monogatari* was immediate. The author of the touching *Sarashina nikki* (mid-11th century; "Sarashina Diary"; Eng. trans., *As I Crossed a Bridge of Dreams*) described how as a girl she longed to visit the capital so that she might read the entire work (which had been completed some 10 years earlier). Imitations and derivative works based on *Genji monogatari,* especially on the last third of it, continued to be written for centuries, inhibiting the fiction composed by the court society.

Makura no sōshi (*c.* 1000; *The Pillow Book of Sei Shōnagon*) is another masterpiece of the Heian period that should be mentioned with *Genji monogatari.* Japanese critics have often distinguished the *aware* of *Genji monogatari* and the *okashi* of *Makura no sōshi. Aware* meant sensitivity to the tragic implications of a moment or gesture, *okashi,* the comic overtones of perhaps the same moment or gesture. The lover's departure at dawn evoked many wistful passages in *Genji monogatari,* but in *Makura no sōshi* the author noted with unsparing exactness the lover's fumbling, ineffectual leavetaking, and his lady's irritation. Murasaki Shikibu's *aware* can be traced through later literature—sensitivity always marked the writings of any author in the aristocratic tradition—but Sei Shōnagon's wit belonged to the Heian court alone.

The Heian court society passed its prime by the middle of the 11th century, but it did not collapse for another 100 years. Long after its political power had been usurped by military men, the court retained its prestige as the fountainhead of culture. But in the 12th century, literary works belonging to a quite different tradition began to appear. *Konjaku monogatari* (early 12th century; "Tales of Now and Then"), a massive collection of religious and folktales, drawn not only from the Japanese countryside but also from Indian and Chinese sources, described elements of society that had never been treated in the court novels. These stories, though crudely written, provide glimpses of how the common people spoke and behaved in an age marked by warfare and new religious movements. The collection of folk songs *Ryōjin hishō,* compiled in 1179 by the emperor Go-Shirakawa, suggests the vitality of this burgeoning popular culture even as the aristocratic society was being threatened with destruction.

Medieval literature: Kamakura, Muromachi, and Azuchi-Momoyama periods (1192–1600). *Kamakura period (1192–1333).* The warfare of the 12th century brought to undisputed power military men (samurai) whose new regime was based on martial discipline. Though the samurai expressed respect for the old culture, some of them even studying tanka composition with the Kyōto masters, the capital of the country moved to Kamakura. The lowered position of women under this feudalistic government perhaps explains the noticeable diminution in the importance of writings by court ladies; indeed, there was hardly a woman writer of distinction between the 13th and 19th centuries.

Poetry

The court poets, however, remained prolific: 15 Imperially sponsored anthologies were completed between 1188 and 1439; and most of the tanka followed the stereotypes established in earlier literary periods.

The finest of the later anthologies, the *Shin kokinshū* (*c.* 1205), was compiled by Fujiwara Sadaie, or Teika, among others, and is considered by many as the supreme accomplishment in tanka composition. The title of the anthology—"the new *Kokinshū*"—indicates the confidence of the compilers that the poets represented were worthy successors of those in the 905 collection; they included (besides the great Teika himself) Teika's father, Fujiwara Toshinari (Shunzei), the priest Saigyō, and the former emperor Go-Toba. These poets looked beyond the visible world for symbolic meanings. The brilliant colours of landscapes filled with blossoms or reddening leaves gave way to monochrome paintings; the poet, instead of dwelling on the pleasure or grief of an experience, sought in it some deeper meaning he could sense if not fully express. The tastes of Teika especially dominated Japanese poetic sensibility, thanks not only to his poetry and essays on poetry but to his choices of the works of the past most worthy of preservation.

Prose

Teika is credited also with a novel, *Matsura no miya monogatari* ("Tale of Matsura Shrine"). Though it is unfinished and awkwardly constructed, its dreamlike atmosphere lingers in the mind with the overtones of Teika's poetry; dreams of the past were indeed the refuge of the medieval romancers, who modelled their language on the *Genji monogatari,* though it was now archaic, and borrowed their themes and characters from the Heian masterpieces. Stories about wicked stepmothers are fairly common; perhaps the writers, contrasting their neglect with the fabled lives of the Heian courtiers, identified themselves with the maltreated stepdaughters; and the typical happy ending of such stories—the stepdaughter in *Sumiyoshi monogatari* is married to a powerful statesman and her wicked stepmother humiliated—may have been the dream fulfillment of their own hopes.

Various diaries describe travels between Kyōto and the shogun's capital in Kamakura. Courtiers often made this long journey in order to press claims in lawsuits, and they recorded their impressions along the way in the typical mixture of prose and poetry. *Izayoi nikki* ("Diary of the Waning Moon"; Eng. trans. in *Translations from Early Japanese Literature*) tells of a journey made in 1277 by the nun Abutsu. A later autobiographical work that also contains extensive descriptions of travel is the superb *Towazu-gatari* (*c.* 1307; "Uninvited Remarks"; Eng.

trans., *The Confessions of Lady Nijō*) by Lady Nijō, a work (discovered only in 1940) that provides a final moment of glory to the long tradition of introspective writing by women at court.

Although these writings in the aristocratic manner preserved much of the manner of Heian literature, works of quite different character were even more prominent in the medieval period. There are many collections of Buddhist and popular tales, of which the most enjoyable undoubtedly is the *Uji shūi monogatari* (*A Collection of Tales from Uji*)—a compilation over a period of years of some 197 brief stories. Although the incidents described in these tales are often similar to those found in *Konjaku monogatari,* they are told with considerably greater literary skill.

Gunki monogatari established as a literary genre

An even more distinctive literary genre of the period is the *gunki monogatari,* or war tale. The most famous, *Heike monogatari* (*The Tale of the Heike*), was apparently first written at the court about 1220, probably by a nobleman who drew his materials from the accounts recited by priests of the warfare between the Taira (Heike) and the Minamoto (Genji) families in the preceding century. The celebrated opening lines of the work, a declaration of the impermanence of all things, also states the main subject, the rise and fall of the Taira family. The text, apparently at first in three books, was expanded to 12 in the course of time, as the result of being recited with improvisations by priest-entertainers. This oral transmission may account not only for the unusually large number of textual variants but also for the exceptionally musical and dramatic style of the work. Unlike the Heian novelists, who rarely admitted words of Chinese origin into their works, the reciters of the *Heike monogatari* employed the contrasting sounds of the imported words to produce what has been acclaimed as the great classic of Japanese style. Although the work is curiously uneven, effective scenes being followed by dull passages in which the narrator seems to be stressing the factual accuracy of his materials, it is at least intermittently superb; and it provided many later novelists and dramatists with characters and incidents for their works.

Heike monogatari was by no means the earliest literary work describing warfare; and other writings, mainly historical in content, were graced by literary flourishes uncommon in similar Western works. *Ōkagami* (*c.* 1120?; "The Great Mirror"; Eng. trans., *Ōkagami*), the most famous of the "mirrors" of Japanese history, undoubtedly influenced the composition of *Heike monogatari,* especially in its moralistic tone. *Hōgen monogatari* (Eng. trans., *Hōgen monogatari*) and *Heiji monogatari* (partial Eng. trans. in *Translations from Early Japanese Literature*) chronicle warfare that antedates the events described in *Heike monogatari* but were probably written somewhat later.

War tales continued to be composed throughout the medieval period. The *Taiheiki* ("Chronicle of the Great Peace"; Eng. trans., *Taiheiki*), for example, covers about 50 years, beginning in 1318, when the emperor Go-Daigo ascended the throne. Though revered as a classic by generations of Japanese, it possesses comparatively little appeal for Western readers, no doubt because so few of the figures come alive.

Characters are more vividly described in two historical romances of the mid- to late 14th century, *Soga monogatari,* an account of the vendetta carried out by the Soga brothers, and *Gikeiki* ("Chronicle of Gikei"; Eng. trans., *Yoshitsune*), describing the life of Minamoto Yoshitsune. Though inartistically composed, these portraits of resourceful and daring heroes caught the imaginations of the Japanese; and their exploits are still prominent on the Kabuki stage.

Another important variety of medieval literature was the reflective essays of Buddhist priests. *Hōjō-ki* (1212; *The Ten Foot Square Hut*) by Kamo Chōmei is a hermit's description of his disenchantment with the world and his discovery of peace in a lonely retreat. The elegiac beauty of its language gives this work, brief though it is, the dignity of a classic. Chōmei was also a distinguished poet, and his essay *Mumyōshō* (*c.* 1210–12; "Nameless Notes") is perhaps the finest example of traditional Japanese poetic criticism.

A later priest, Yoshida Kenkō, writing during the days of warfare and unrest that brought an end to the Kamakura shogunate in 1333, the brief restoration of Imperial authority under the emperor Go-Daigo from 1333 to 1335, and the institution of the Ashikaga shogunate in 1338, barely hints at the turmoil of the times in his masterpiece *Tsurezure-gusa* (*c.* 1330; *Essays in Idleness*); instead, he looks back nostalgically to the past, seeking out the survivals of happier days. Kenkō's aesthetic judgments, often based on a this-worldly awareness rather surprising in a Buddhist priest, gained wide currency, especially after the 17th century, when *Tsurezure-gusa* was widely read.

The Muromachi (1338–1573) and Azuchi-Momoyama (1574–1600) periods. In the 15th century a poetic form of plebeian origins displaced the tanka as the preferred medium of the leading poets. Renga (linked verse) had begun as the composition of a single tanka by two people and was a popular pastime even in remote rural areas. One person would compose the first three lines of a tanka, often giving obscure or even contradictory details in order to make it harder for the second person to complete the poem intelligibly. Gradually, renga spread to the court poets, who saw the artistic possibilities of this diversion and drew up "codes" intended to establish renga as an art. These codes made possible the masterpieces of the 15th century, but their insistence on formalities (*e.g.,* how often a "link" on the moon might appear in 100 links, and which links must end with a noun and which with a verb) inevitably diluted the vigour and freshness of the early renga, itself a reaction against the excessively formal tanka. Nevertheless, the renga of the great 15th-century master Iio Sōgi and his associates are unique in their shifting lyrical impulses, moving from link to link like successive moments of a landscape seen from a boat, avoiding any illusion that the whole was conceived in one person's mind.

Poetry

Prose

The short stories of the 15th and 16th centuries cannot be said to have high literary value. Many still look back to the world of the Heian court, but others introduce folk materials or else elements of the miraculous in the attempt to interest readers who lacked the education to appreciate the conventional literary manner. Even though many promising stories are ruined by absurdities before their course is run, for a few moments they often give unforgettable glimpses of a society torn by disorder. The stories are anonymous, but the authors seem to have been both courtiers and Buddhist priests.

Drama

Unquestionably the finest literary works of the 15th century are the Nō dramas, especially those by Zeami Motokiyo (see EAST ASIAN ARTS). They were written in magnificent poetry (often compared to "brocade" because of the many allusions to the poetry of the past) and were provided with a structure that is at once extremely economical and free. Many are concerned with the Buddhist sin of attachment: an inability to forget his life in this world prevents a dead man from gaining release but forces him to return again and again as a ghost to relive the violence or passion of his former existence. Only prayer and renunciation can bring about deliverance. Zeami's treatises on the art of Nō display extraordinary perceptivity. His stated aims were dramatic conviction and reality, but these ideals meant ultimates to him and not superficial realism. Some Nō plays, it is true, have little symbolic or supernatural content, but the central elements of a typical program of five Nō plays were found in the highly poetic and elusive masterpieces that suggest a world invisible to the eye but evokable by the actors through the beauty of movements and speech. Unhappiness over a world torn by disorder may have led writers to suggest in their works truths that lie too deep for words. This seems to have been the meaning of *yūgen* ("mystery and depth"), the ideal of the Nō plays. Parallel developments occurred in the tea ceremony, the landscape garden, and monochrome painting, all arts that suggest or symbolize rather than state.

Literature during the Tokugawa period (1603–1867). The restoration of peace and the unification of Japan were achieved in the early 17th century, and for approximately 250 years the Japanese enjoyed almost uninterrupted peace. During the first half of the Tokugawa period, the cities of Kyōto and Ōsaka dominated cultural activity; but from about 1770 Edo (the modern Tokyo) became paramount. From the mid-1630s to the early 1850s Japan was closed, by government decree, to contact with the outside world. Initially, this isolation encouraged the development of indigenous forms of literature; but, eventually, in the virtual absence of fertilizing influence from abroad, it resulted in provincial writing. The adoption of printing in the early 17th century made a popular literature possible. The Japanese had known the art of printing since at least the 8th century, but they had reserved it exclusively for reproducing Buddhist writings. The Japanese classics existed only in manuscript form. It is possible that the demand for copies of literary works was so small that it could be satisfied with manuscripts, costly though they were; or perhaps aesthetic considerations made the Japanese prefer manuscripts in beautiful calligraphy, sometimes embellished with illustrations. Whatever the case, not until 1591 was a nonreligious work printed. About the same time, Portuguese missionaries in Nagasaki were printing books in the Roman alphabet. In 1593, in the wake of the Japanese invasion of Korea, a printing press with movable type was sent as a present to the emperor Go-Yōzei. Printing soon developed into the hobby or extravagance of the rich, and many examples of Japanese literature began to appear in small editions. Commercial publication began in 1609; by the 1620s even works of slight literary value were being printed for a public eager for new books.

Role of printing in creating a popular literature

Poetry

Early Tokugawa period (1603–c. 1770). Poetry underwent many changes during the early part of the Tokugawa period. At first the court poets jealously maintained their monopoly over the tanka, but gradually other men, many of them *kokugakusha* ("scholars of national learning"), changed the course of tanka composition by attempting to restore to the form the simple strength of *Man'yōshū* poetry. The early 18th-century poet Kamo Mabuchi was the best of the neo-*Man'yōshū* school, but his tanka rarely rise above mere competence in the ancient language.

The chief development in poetry during the Tokugawa shogunate was the emergence of the haiku as an important genre. This exceedingly brief form (17 syllables arranged in lines of five, seven, and five syllables) had originated in the hokku, or opening verse of a renga sequence, which had to contain in its three lines mention of the season, the time of day, the dominant features of the landscape, etc., making it almost an independent poem. The hokku became known as the haiku late in the 19th century, when it was entirely divested of its original function of opening a sequence of verse; but today even the 17th-century hokku are usually called haiku.

As early as the 16th century haikai renga, or comic renga, had been composed by way of diversion after an evening of serious renga composition, reverting to the original social, rather than literary, purpose of making linked verse. As so often happened in Japan, however, a new art, born as a reaction to the stultifying practices of an older art, was "discovered," codified, and made respectable by practitioners of the older art, generally at the cost of its freshness and vitality. Matsunaga Teitoku, a conventional 17th-century poet of tanka and renga who revered the old traditions, became almost in spite of himself the mentor of the new movement in comic verse, largely as the result of pressure from his eager disciples. Teitoku brought dignity to the comic renga and made it a demanding medium, rather than the quip of a moment. His haikai were distinguishable from serious renga not by their comic conception but by the presence of a *haigon*—a word of Chinese or recent origins that was normally not tolerated in classical verse.

Inevitably, a reaction arose against Teitoku's formalism. The poets of the Danrin school, headed by Nishiyama Sōin and Ihara Saikaku, insisted that it was pointless to waste months if not years perfecting a sequence of 100 verses. Their ideal was rapid and impromptu composition; and their verses, generally colloquial in diction, were intended to amuse for a moment rather than to last for all time. Saikaku especially excelled at one-man composition of extended sequences; in 1684 he composed the incredible total of 23,500 verses in a single day and night, too fast for the scribes to do more than tally.

The haiku was perfected into a form capable of conveying poetry of the highest quality by Matsuo Bashō. After passing through an apprenticeship in both Teitoku and Danrin schools, Bashō founded a school of his own, insisting that a haiku must contain both a perception of some eternal truth and an element of contemporaneity, combining the characteristic features of the two earlier schools. Despite their brief compass, Bashō's haiku often suggest, by means of the few essential elements he presents, the whole world from which they have been extracted; the reader must participate in the creation of the poem. Bashō's best known works are travel accounts interspersed with his verses; of these, *Oku no hosomichi* (1694; *The Narrow Road Through the Deep North*) is perhaps the most popular and revered work of Tokugawa literature.

Prose

The general name for the prose composed between 1600 and 1682 is *kana-zōshi,* or "*kana* books," the name originally having been used to distinguish popular writings in the Japanese syllabary from more learned works in Chinese. The genre embraced not only fiction but also works of a near historical nature, pious tracts, books of practical information, guidebooks, evaluations of courtesans and actors, and miscellaneous essays. Only one writer of any distinction is associated with the *kana-zōshi*—Asai Ryōi, a samurai who became the first popular and professional writer in Japanese history. Thanks to the development of relatively cheap methods of printing and a marked increase in the reading public, Ryōi was able to make a living as a writer. Although some of his works are Buddhist, he wrote in a simple style, mainly in *kana.* His most famous novel, *Ukiyo monogatari* (*c.* 1661; "Tales of the Floating World"), is primitive both in technique and in plot; but under his mask of frivolity Ryōi attempted to treat the hardships of a society where the officially proclaimed Confucian philosophy concealed the gross inequalities in the lots of different men.

The first important novelist of the new era was Ihara Saikaku. Some Japanese critics rank him second only to Murasaki Shikibu in all Japanese literature, and his works have been edited with the care accorded only to great classics. Such attention would surely have surprised Saikaku, whose fiction was dashed off almost as rapidly as his legendary performances of comic renga, with little concern for the judgments of posterity.

Saikaku's first novel, *Kōshoku ichidai otoko* (1682; *The Life of an Amorous Man*), changed the course of Japanese fiction. The title itself had strong erotic overtones, and the plot describes the adventures of one man, from his precocious essays at lovemaking as a child of seven to his decision at the age of 60 to sail to an island populated only by women. The licensed quarters of prostitution established in various Japanese cities by the Tokugawa government (despite its professions of Confucian morality), in order to help control unruly samurai by dissipating their energies, became a centre of the new culture. Expertise in the customs of the brothels was judged the mark of the man of the world. The old term *ukiyo,* which had formerly meant the "sad world" of Buddhist stories, now came to designate its homonym, the "floating world" of pleasure; this was the chosen world of Saikaku's hero, Yonosuke, who became the emblematic figure of the era.

Saikaku's masterpiece, *Kōshoku gonin onna* (1686; *Five Women Who Loved Love*), described the loves of women of the merchant class, rather than prostitutes; this was the first time that women of this class were given such attention. In other works he described, sometimes with humour but sometimes with bitterness, the struggles of merchants to make fortunes. His combination of a glittering style and warm sympathy for the characters lifted his tales from the borders of pornography to high art.

Saikaku was a central figure in the renaissance of literature of the late 17th century. The name Genroku (an era name designating the period 1688–1703) is often used of the characteristic artistic products: the Ukiyo-e ("pictures of the floating world"); the *ukiyo-zōshi* ("tales of the floating world"); the Kabuki and *jōruri,* or puppet theatres; and haiku poetry. Unlike its antecedents, this culture prized modernity above conformity to the ancient traditions; to be abreast of the floating world was to be up-to-date, sharing in the latest fashions and slang, delighting in the moment rather than in the eternal truths of Nō plays of medieval poetry.

Another, darker side to Genroku culture is depicted in Saikaku's late works, with their descriptions of the desperate expedients to which men turned in order to pay their bills. Saikaku seldom showed much sympathy for the prostitutes he described; but the chief dramatist of the time, Chikamatsu Monzaemon, wrote his best plays about unhappy women, driven by poverty into their lives as prostitutes, whose only release from the sordid world in which they were condemned to dwell came when they joined their lovers in double suicides. In the world of merchants treated by Chikamatsu, a lack of money, rather than the cosmic griefs of the Nō plays, drove men to death with the prostitutes they loved but could not afford to buy.

Drama

Chikamatsu wrote most of his plays for the puppet theatre, which, in the 18th century, enjoyed even greater popularity than Kabuki. His plays fell into two main categories: those based, however loosely, on historical facts or legends, and those dealing with contemporary life. The domestic plays are rated much higher critically because they avoid the bombast and fantastic displays of heroism that mark the historical dramas; but the latter, adapted for the Kabuki theatre, are superb acting vehicles.

The mainstays of the puppet theatre were written not by Chikamatsu but by his successors; his plays, despite their literary superiority, failed to satisfy the audiences' craving for displays of puppet techniques and for extreme representations of loyalty, self-sacrifice, and other virtues of the society. The most popular puppet play (later adapted also for the Kabuki actors) was *Chūshingura* (1748; "The Treasury of Loyal Retainers"; Eng. trans., *Chūshingura*) by Takeda Izumo and his collaborators; the same men were responsible for half a dozen other perennial favourites of the Japanese stage. The last great 18th-century writer of puppet plays, Chikamatsu Hanji, was a master of highly dramatic, if implausible, plots.

Poetry

Late Tokugawa period (c. 1770–1867). The literature of the late Tokugawa period is generally inferior to earlier achievements, especially those of the Genroku masters. Authentic new voices, however, were heard in traditional poetic forms. Later neo-*Man'yōshū* poets such as Ryōkan, Ōkuma Kotomichi, and Tachibana Akemi proved that the tanka was not limited to descriptions of the sights of nature or disappointed love but could express joy over fish for dinner or wrath at political events. Some poets who felt that tanka did not provide ample scope for the display of such emotions turned, as in the past, to writing poetry in Chinese. The early 19th-century poet Rai Sanyō probably wrote verse in Chinese more skillfully than any previous Japanese.

Later Tokugawa poets also added distinctive notes of their own to the haiku. Yosa Buson, for example, introduced a romantic and narrative element, and Kobayashi Issa employed the accents of the common people.

Prose

A great variety of fiction was produced during the last century of the Tokugawa shogunate, but it is commonly lumped together under the somewhat derogatory heading of *gesaku* ("playful composition"). The word "playful" did not necessarily refer to the subject matter but to the professed attitude of the authors, educated men who disclaimed responsibility for their compositions. Ueda Akinari, the last master of fiction of the 18th century, won a high place in literary history mainly through his brilliant style, displayed to best advantage in *Ugetsu monogatari* (1776; *Tales of Moonlight and Rain*), a collection of supernatural tales. The *gesaku* writers, however, did not follow Akinari in his perfectionist attention to style and construction; instead, they produced books of almost formless gossip, substituting the raciness of daily speech for the elegance of the classical language, and relying heavily on the copious illustrations for success with the public.

The *gesaku* writers were professionals who made their living by sale of their books. They aimed at as wide a public as possible, and when a book was successful it was usually followed by as many sequels as the public would accept. The most popular of the comic variety of *gesaku* fiction was *Tōkai dōchū hizakurige* (1802–22; "Travels on Foot on the Tōkaidō"; Eng. trans., *Shank's Mare*), by Jippensha Ikku, an account of the travels and comic misfortunes of two irrepressible men from Edo along the Tōkaidō, the great highway between Kyōto and Edo. *Shunshoku umegoyomi* (1832–33; "Spring Colours: The Plum Calendar"), by Tamenaga Shunsui, is the story of Tanjirō, a peerlessly handsome but ineffectual young man for whose affections various women fight. The author at one point defended himself against charges of immorality: "Even though the women I portray may seem immoral, they are all imbued with deep sentiments of chastity and fidelity." It was the standard practice of *gesaku* writers, no matter how frivolous their compositions might be, to pretend that their intent was didactic.

The *yomihon* ("books for reading"—so called to distinguish them from works enjoyed mainly for their illustrations) were much more openly moralistic. Although they were considered to be *gesaku,* no less than the most trivial books of gossip, their plots were burdened with historical materials culled from Chinese and Japanese sources, and the authors frequently underlined their didactic purpose. But despite the serious intent of the *yomihon,* they were romances, rather than novels; and their characters, highly schematized, tended to be witches, fairy princesses, and impeccably noble gentlemen. Where they succeeded, as in a few works by Takizawa Bakin, they are absorbing as examples of storytelling rather than as embodiments of the principle of *kanzen chōaku* ("the encouragement of virtue and the chastisement of vice"), Bakin's professed aim in writing fiction.

Japanese literature in general was at one of its lowest levels at the end of the Tokugawa period. A few tanka poets and the Kabuki dramatist Kawatake Mokuami are the only writers of the period whose works are still read today. It was an exhausted literature that could be revived only by the introduction of fresh influences from abroad.

Modern literature. Even after the arrival of Commodore Perry's fleet in 1853 and the gradual opening of the country to the West and its influence, there was little noticeable effect on Japanese literature. The long closure of the country and the general sameness of Tokugawa society for decades at a time seems to have atrophied the imaginations of the *gesaku* writers. Even the presence of curiously garbed foreigners, which should have provoked some sort of reaction from authors searching for new materials, at first produced little effect. The *gesaku* writers were oblivious to the changes in Japanese society, and they continued to grind out minor variants on the same hackneyed themes of the preceding 200 years.

It was only after the removal in 1868 of the capital to Edo (renamed Tokyo) and the declaration of the emperor Meiji that he would seek knowledge from the entire world that the *gesaku* writers realized their days of influence were numbered. They soon fell under attack from their old enemies, the Confucian denouncers of immoral books, and also from advocates of the new Western learning. Although the *gesaku* writers responded with satirical pieces and traditional Japanese fiction deriding the new learning, they were helpless to resist the changes transforming the entire society.

Early translations and their influence

Introduction of Western literature. Translations from European languages of nonliterary works began to appear soon after the Meiji Restoration. The most famous example was the translation (1870) of Samuel Smiles's *Self-Help;* it became a kind of bible for ambitious young Japanese eager to emulate Western examples of success. The first important translation of a European novel was *Ernest Maltravers,* by the British novelist Lord Bulwer-Lytton, which appeared in 1879 under the title *Karyū shunwa* ("A Spring Tale of Blossoms and Willows"). The early translations were inaccurate, and the translators unceremoniously deleted any passages they could not understand readily or which they feared might be unintelligible to Japanese readers. They also felt obliged to reassure readers that, despite the foreign names of the characters, the emotions they felt were exactly the same as those of a Japanese.

It did not take long, however, for the translators to discover that European literature possessed qualities unknown in the Japanese writings of the past. The literary scholar Tsubouchi Shōyō was led by his readings in European fiction and criticism to reject didacticism as a legitimate purpose of fiction; he insisted instead on its artistic values. His critical essay *Shōsetsu shinzui* (1885; "The Essence of the Novel") greatly influenced the writing of subsequent fiction not only because of its emphasis on realism as opposed to didacticism but because Tsubouchi, a member of the samurai class, expressed the conviction that novels, hitherto despised by the intellectuals as mere entertainments for women and children, were worthy of even a scholar's attention.

Ukigumo (1887–89; "Floating Cloud"; Eng. trans., *Japan's First Modern Novel*), by Futabatei Shimei, was the first modern Japanese novel. The author was familiar with Russian literature and contemporary Western literary criticism. Futabatei wrote *Ukigumo* in the colloquial, apparently because his readings in Russian literature had convinced him that only the colloquial could suitably be used when describing the writer's own society. Despite Futabatei's success with this experiment, most Japanese writers continued to employ the literary language until the end of the century. This was due, no doubt, to their reluctance to give up the rich heritage of traditional expression in favour of the unadorned modern tongue.

Western influences on poetry. Translations of Western poetry led to the creation of new Japanese literary forms. The pioneer collection, *Shintaishi-shō* (1882; "Selection of Poems in the New Style"), contained not only translations from English but also five original poems by the translators in the poetic genres of the foreign examples. The translators declared that although European poetry had greater variety than Japanese poetry—some poems are rhymed, others unrhymed, some are extremely long, others abrupt—it was invariably written in the language of ordinary speech. The insistence on modern language and the discovery of the many forms available to the poet were not the only lessons learned from European poetry. The translators also made the Japanese public aware of how much of human experience had never been treated in the tanka or haiku forms.

Innumerable Western critics have sarcastically commented on the Japanese proclivity for imitating foreign literary models and on their alleged indifference to their own traditions. It is true that without Russian examples Futabatei could not have written *Ukigumo,* and without English examples such poets as Shimazaki Tōson could not have created modern Japanese poetry; but far from recklessly abandoning their literary heritage, most writers were at great pains to acquaint themselves with their traditional literature. The outstanding novelists of the 1890s—Ozaki Kōyō, Kōda Rohan, Higuchi Ichiyō, and Izumi Kyōka—all read Saikaku and were noticeably influenced by him. Ichiyō's short novel *Takekurabe* (1895; *Growing Up*) described the children of the Yoshiwara quarter of Edo in a realistic manner quite unlike that of the usual stories about prostitutes and their customers, but she used the language of Saikaku for her narration. Izumi Kyōka, though educated partly at a Western mission school, wrote superbly in the vein of late Tokugawa fiction; something of the distant Japanese literary past pervaded even his writings of the 1930s, the final years of his life.

Influence of Saikaku

In poetry, too, the first products of Western influence were comically inept experiments with rhyme and with such unpromising subjects as the principles of sociology. Shimazaki Tōson's "Akikaze no uta" (1896; "Song of the Autumn Wind"), however, is not merely a skillful echo of Shelley but a true picture of a Japanese landscape; and the irregular lines of his poem tend to fall into the traditional pattern of five and seven syllables.

A decade after the works of such English Romantic poets as Shelley and Wordsworth had influenced Japanese poetry, the translations made by Ueda Bin of the French Parnassian and Symbolist poets made an even more powerful impression. Ueda wrote, "The function of symbols is to help create in the reader an emotional state similar to that in the poet's mind; symbols do not necessarily communicate the same conception to everyone." This view was borrowed from the West, but it accorded perfectly with the qualities of tanka.

Because of the ambiguities of traditional Japanese poetic expression, it was natural for a given poem to produce different effects on different readers; the important thing, as in Symbolist poetry, was to communicate the poet's mood. If the Japanese poets of the early 1900s had been urged to avoid contamination by foreign ideas, they would have declared that this was contrary to the spirit of an enlightened age. But when informed that eminent foreign poets preferred ambiguity to clarity, the Japanese responded with double enthusiasm.

Revitalization of tanka and haiku. Even the traditional forms, tanka and haiku, though moribund in 1868, took on new life, thanks largely to the efforts of Masaoka Shiki, a distinguished late 19th-century poet in both forms but of even greater importance as a critic. Yosano Akiko, Ishikawa Takuboku, and Saitō Mokichi were probably the most successful practitioners of the new tanka. Yosano Akiko's collection *Midaregami* (1901; *Tangled Hair*) stirred female readers especially, not only because of its lyrical beauty but because Akiko herself seemed to be proclaiming a new age of romantic love. Takuboku emerged in the course of his short life (he died in 1912 at the age of 26) as perhaps the most popular tanka poet of all time. His verses are filled with strikingly individual expressions of his intransigent personality. Saitō Mokichi combined an absorption with *Man'yōshū* stylistics and a professional competence in psychiatry. Despite the austere nature of his poetry, he was recognized for many years as the leading tanka poet. In haiku, Takahama Kyoshi built up a following of poets strong enough to withstand the attacks of critics who declared that the form was inadequate to deal with the problems of modern life. Kyoshi himself eventually decided that the function of haiku was the traditional one of an intuitive apprehension of the beauties of nature; but other haiku poets employed the medium to express entirely unconventional themes.

Most tanka and haiku poets continued to use the classical language, probably because its relative concision permitted them to impart greater content to their verses than modern Japanese permits. Poets of the "new style," therefore, were readier to employ the colloquial. Hagiwara Sakutarō, generally considered the finest Japanese poet of the 20th century, brilliantly exploited the musical and expressive possibilities of the modern tongue. Other poets, such as Horiguchi Daigaku, devoted themselves mainly to translations of European poetry, achieving results so compelling in Japanese that these translations are considered to form an important part of the modern poetry of Japan.

"New-style" poetry

The novel between 1905 and 1941. The dominant stream in Japanese fiction since the publication of *Hakai* (1906; *The Broken Commandment*), by Shimazaki Tōson, and *Futon* (1907; *The Quilt*), by Tayama Katai, has been naturalism. Although the movement was originally inspired by the works of the 19th-century French novelist Émile Zola and other European naturalists, it quickly took on a distinctively Japanese colouring, rejecting (as a Confucian scholar might have rejected *gesaku* fiction) carefully developed plots or stylistic beauty in favour of absolute verisimilitude in the author's confessions or in his minute descriptions of the lives of unimportant people hemmed in by circumstances beyond their control.

By general consent, however, the two outstanding novelists of the early 20th century were men who stood outside the naturalist movement, Mori Ōgai and Natsume Sōseki. Ōgai began as a writer of autobiographical fiction with strong overtones of German Romantic writings. Midway in his career he shifted to historical novels that are virtually devoid of fictional elements but are given literary distinction by their concise and masculine style. Sōseki gained fame with humorous novels such as *Botchan* (1906; "The Young Master"; Eng. trans., *Botchan*), a fictionalized account of his experiences as a teacher in a provincial town. *Botchan* has enjoyed phenomenal popularity ever since it first appeared. It is the most approachable of Sōseki's novels, and the Japanese have found pleasure in identifying themselves with the impetuous, reckless, yet basically decent hero. The coloration of Sōseki's subsequent novels became progressively darker, but even the most gloomy have maintained their reputation among Japanese readers, who take it for granted that Sōseki is the greatest of the modern Japanese novelists and who find echoes in their own lives of the mental suffering he described. Sōseki wrote mainly about intellectuals living in a Japan that had been brutally thrust into the 20th century. His best known novel, *Kokoro* (1914; "The Heart"; Eng. trans., *Kokoro*), revolves around another familiar situation in his novels, two men in love with the same woman. His last novel, *Meian* (1916; *Light and Darkness*), though unfinished, has been acclaimed by some as his masterpiece.

An amazing burst of creative activity occurred in the decade following the end of the Russo-Japanese War in 1905. Probably never before in the history of Japanese literature were so many important writers working at once. Three novelists who first emerged into prominence at this time were Nagai Kafū, Tanizaki Jun'ichirō, and Akutagawa Ryūnosuke. Nagai Kafū was infatuated with French culture and described with contempt the meretricious surface of modern Japan. In later years, however, though still alienated from the Japanese present, he showed nostalgia for the Japan of his youth, and his most appealing works contain evocations of the traces of an old and genuine Japan that survived in the parody of Western culture that was Tokyo.

Tanizaki Jun'ichirō

Tanizaki's novels, notably *Tade kuu mushi* (1928–29; "Insects That Eat Knotweed"; Eng. trans., *Some Prefer Nettles*), often presented a conflict between traditional Japanese and Western-inspired ways. In his early works he also proclaimed a preference for the West. Tanizaki's views changed after he moved to the Kansai region in the wake of the Great Kantō Earthquake of 1923, and his subsequent writings traced his gradual accommodation with the old culture of Japan that he had previously rejected. Between 1939 and 1941 Tanizaki published the first of his three modern-language versions of *Genji monogatari.* He willingly sacrificed years of his career to this task because of his unbounded admiration for the supreme work of Japanese literature.

Tanizaki's longest novel, *Sasameyuki* (1943–48; "A Light Snowfall"; Eng. trans., *The Makioka Sisters*), evoked with evident nostalgia the Japan of the 1930s, when people were preoccupied not with the prosecution of a war but with marriage arrangements, visits to sites famous for their cherry blossoms, or the cultural differences between Tokyo and Ōsaka. Two postwar novels by Tanizaki enjoyed great popularity, *Kagi* (1956; *The Key*), the account of a professor's determination to have his fill of sex with his wife before impotence overtakes him; and *Fūten rōjin nikki* (1961–62; *Diary of a Mad Old Man*), a work in a comic vein that describes a very old man's infatuation with his daughter-in-law. No reader would turn to Tanizaki for wisdom as to how to lead his life, nor for a penetrating analysis of society, but his works not only provide the pleasures of well-told stories but also convey the special phenomenon of adulation and rejection of the West that is so prominent a part of the Japanese culture of the 20th century.

Akutagawa established his reputation as a brilliant storyteller who transformed materials found in old Japanese collections by infusing them with modern psychology. No writer enjoyed a greater following in his time, but Akutagawa found less and less satisfaction in his reworkings of existing tales and turned eventually to writing about himself in a sometimes harrowing manner. His suicide in 1927 shocked the entire Japanese literary world. The exact cause is unknown—he wrote of a "vague malaise"—but perhaps Akutagawa felt incapable either of sublimating his personal experiences into pure literature or else of giving them the accents of the proletarian literature movement, then at its height.

The proletarian literature movement in Japan, as in various other countries, attempted to use literature as a weapon to effect reform and even revolution in response to social injustices. Although the movement gained virtual control of the Japanese literary world in the late 1920s, governmental repression beginning in 1928 eventually destroyed it. The chief proletarian writer, Kobayashi Takiji, was tortured to death by the police in 1933. Few of the writings produced by the movement are of literary worth, but the concern for classes of people who had formerly been neglected by Japanese writers gave these works their special significance.

Other writers of the period, convinced that the essential function of literature was artistic and not propagandistic, formed schools such as the "Neo-sensualists" led by Yokomitsu Riichi and Kawabata Yasunari. Yokomitsu's politics eventually moved far to the right, and the promulgation of these views, rather than his efforts to achieve modernism, coloured his later writings; but Kawabata's works (for which he won the Nobel Prize for Literature in 1968) are still admired for their lyricism and intuitive construction. Though Kawabata began as a modernist and experimented with modernist techniques to the end of his career, he is better known for his portraits of women, whether the geisha of *Yukiguni* (1948; *Snow Country*) or the different women whose lives are concerned with the tea ceremony in *Sembazuru* (1952; *Thousand Cranes*).

The "I novel"

Japanese critics have divided the fiction of the prewar period into schools, each usually consisting of one leading writer and his disciples. Probably the most influential author was Shiga Naoya. His characteristic literary form was the "I novel" (*watakushi shōsetsu*), a work that treats autobiographical materials with stylistic beauty and great intelligence but is not remarkable for invention. Shiga's commanding presence caused the I novel to be more respected by most critics than outright works of fiction; but the writings of his disciples are sometimes hardly more than pages torn from a diary, of interest only if the reader is already devoted to the author.

The postwar novel. The aggressive wars waged by the Japanese militarists in the 1930s inhibited literary production. Censorship became increasingly stringent, and writers were expected to promote the war effort. During the Pacific War of 1941–45 little worthwhile literature appeared. Tanizaki began serial publication of *The Makioka Sisters* in 1943, but publication was halted by official order, and the completed work appeared only after the war. The immediate postwar years signalled an extraordinary period of activity, both by the older generation and by new writers. The period is vividly described in the writings of Dazai Osamu, notably *Shayō* (1947; *The Setting Sun*). Other writers described the horrors of the war years; perhaps the most powerful was *Nobi* (1951; *Fires on the Plain*) by Ōoka Shōhei, which described defeated Japanese soldiers in the Philippine jungles. The atomic bombs also inspired much poetry and prose, though it was often too close to the events to achieve artistic integrity. A few works, especially *Kuroi ame* (1965; *Black Rain*) by Ibuse Masuji, succeeded in suggesting the ultimately indescribable horror of the disaster.

Mishima Yukio

The Japan of the immediate postwar period and the prosperous Japan of the 1950s and 1960s provided the background for most of the works of Mishima Yukio, an exceptionally brilliant and versatile novelist and playwright who became the first Japanese writer generally known abroad. Mishima's best known works include *Kinkaku-ji* (1956; *The Temple of the Golden Pavilion*), a psychological study, based on an actual incident, of a young monk who burned a famous architectural masterpiece; and *Hōjō no Umi* (1965–70; *The Sea of Fertility*), the tetralogy he completed on the day of his death. Abe Kōbō was notable among modern writers in that he managed, sometimes by resorting to avant-garde techniques, to transcend the particular condition of being a Japanese and to create myths of suffering humanity in such a work as *Suna no onna* (1962; *The Woman in the Dunes*). The special nature of traditional Japanese culture, which made it infertile ground for Christianity in the 16th century, was treated in several moving novels by Endō Shūsaku, notably *Chimmoku* (1966; *Silence*). The novels of Kita Morio were characterized by an attractive streak of humour that provided a welcome contrast to the prevailingly dark tonality of other contemporary Japanese novels. *Nire-ke no hitobito* (1963–64; *The House of Nire*), though based on the careers of his grandfather and father (the poet Saitō Mokichi) was saved by its humour from becoming no more than an I novel. For almost 20 years Ōe Kenzaburō, a novelist of exceptional power, was treated as the youngest writer of importance, but in the 1970s a new generation at last began to appear with a promise of a renewal of the modern Japanese novel.

The modern drama. The modern Japanese theatre also began with translations and adaptations of Western plays. This new theatre originated at the end of the 19th century, when the public was still too much under the influence of Kabuki to appreciate plays without music or dance. Even in the 20th century, a distinguished dramatist such as Kishida Kunio rarely had the opportunity to see his works performed. The development of modern drama no doubt was hampered by the introduction, at about the same time, of motion pictures, which had a much greater appeal for the public. The most successful playwrights of the 1920s and 1930s, such as Mayama Seika, wrote works that, although the products of modern minds, exploited the special talents of Kabuki actors by treating historical themes and by preserving the traditional stage language. Various distinguished writers were attracted from time to time to the theatre, but they were forced to devote their major efforts to writing fiction, if only because they were so badly remunerated for their plays. It was not until after World War II that modern dramas worthy of an international audience were written and staged.

BIBLIOGRAPHY

General works: SHUICHI KATO, *A History of Japanese Literature*, 3 vol. (1979–83; originally published in Japanese, 2 vol., 1975–80), considers Japanese literature as a key to Japanese intellectual history. JIN'ICHI KONISHI, *A History of Japanese Literature*, trans. from the Japanese, ed. by EARL MINER (1984–), gives special attention to relations among the literatures of other countries of Asia. DONALD KEENE (ed.), *Anthology of Japanese Literature to the Nineteenth Century*, rev. ed. (1978), offers selections from earliest times, with introductions. A good but brief discussion of Japanese myths is E. DALE SAUNDERS, "Japanese Mythology," in SAMUEL N. KRAMER (ed.), *Mythologies of the Ancient World* (1961). GEOFFREY BOWNAS and ANTHONY THWAITE (trans.), *The Penguin Book of Japanese Verse* (1964), gives examples in every form. ROBERT H. BROWER and EARL MINER, *Japanese Court Poetry* (1961), is an excellent study extending from the earliest period to the 14th century. BURTON WATSON (trans.), *Japanese Literature in Chinese*, 2 vol. (1975–76), provides an excellent selection of poetry and some examples in prose, and his work with HIROAKI SATO (eds. and trans.), *From the Country of Eight Islands: An Anthology of Japanese Poetry* (1981), contains poetry of every period. EARL MINER (ed.), *Japanese Poetic Diaries* (1969, reprinted 1976), contains examples from the Heian period to the 19th century.

Early and Nara periods: Kojiki (1968), and *This Wine of Peace, This Wine of Laughter: A Complete Anthology of Japan's Earliest Songs* (1968), are both translations by DONALD L. PHILIPPI. *The Manyōshū: One Thousand Poems Selected and Translated from the Japanese* (1940, reissued 1965), contains an excellent selection from this great anthology; and IAN HIDEO LEVY, *The Ten Thousand Leaves* (1981–), is the first volume of a complete translation.

Heian period: IVAN MORRIS, *The World of the Shining Prince* (1964, reissued 1979), provides the social and historical background for the Heian literary masterpieces. KYOKO MOTOMOCHI NAKAMURA (trans.), *Miraculous Stories from the Japanese Buddhist Tradition: The Nihon Ryōiki of the Monk Kyōkai* (1973), also contains Buddhist tales from Indian and Chinese sources. *Kokinshū* is a complete translation by LAUREL RASPLICA RODD and MARY CATHERINE HENKENIUS (1984). *The Gossamer Years,* trans. by EDWARD SEIDENSTICKER (1964, reprinted 1973); MURASAKI SHIKIBU, *The Tale of Genji,* trans. by EDWARD SEIDENSTICKER (1976, reprinted 1981), and another translation by ARTHUR WALEY (1926–33); and *The Pillow Book of Sei Shōnagon,* 2 vol. (1967), and *As I Crossed a Bridge of Dreams* (1971), both trans. by IVAN MORRIS, are all poetic renderings of these classics. *The Izumi Shikibu Diary,* trans. by EDWIN A. CRANSTON (1969); *Murasaki Shikibu, Her Diary and Poetic Memoirs,* trans. by RICHARD BOWRING (1982); *Tales of Ise,* trans. by HELEN C. MCCULLOUGH (1968); and *Tales of Yamato,* trans. by MILDRED M. TAHARA (1980), are more literal versions with scholarly introductions. HELEN C. MCCULLOUGH (trans.), *Ōkagami, the Great Mirror* (1980), and, with WILLIAM H. MCCULLOUGH (trans.), *A Tale of Flowering Fortunes: Annals of Japanese Aristocratic Life in the Heian Period* (1980), are histories written with an admixture of poetry and fiction. THOMAS H. ROHLICH (trans.), *A Tale of Eleventh Century Japan: Hamamatsu Chūnagon Monogatari,* is an example of later Heian fiction. MARIAN URY (trans.), *Tales of Times Now Past* (1979), includes 62 stories from the *Konjaku monogatari.* WILLIAM R. LAFLEUR (trans.), *Mirror for the Moon* (1978), is a collection of free but poetic translations of waka by Saigyō.

Middle Ages: WILLIAM R. LAFLEUR, *The Karma of Words: Buddhism and the Literary Arts in Medieval Japan* (1983), provides a general background for the literature of the period. *The Tale of the Heike,* trans. by HIROSHI KITAGAWA and BRUCE T. TSUCHIDA (1975), is a rendering of the classic account of the warfare between the Taira and Minamoto clans. *The Taiheiki* (1959, reprinted 1976), and *Yoshitsune* (1966), both trans. by HELEN C. MCCULLOUGH, are accurate versions of war tales in rather old-fashioned language. DONALD KEENE (trans. and ed.), *Essays in Idleness* (1967), and *Twenty Plays of the Nō Theatre* (1970), are readable though fairly literal; ARTHUR WALEY (trans.), *The Nō Plays of Japan* (1922, reissued 1979), gives freer versions of the plays. D.E. MILLS (trans.), *A Collection of Tales from Uji* (1970), is a readable and scholarly study. KAREN BRAZELL (trans.), *The Confessions of Lady Nijō* (1973), is a diary of exceptional interest. EARL MINER, *Japanese Linked Poetry* (1979), is a study of renga and haikai poetry with translations.

Tokugawa period: DONALD KEENE, *World Within Walls* (1976, reprinted 1978), is a history of the literature of the period. The several excellent translations of works by Ihara Saikaku include *Five Women Who Loved Love,* trans. by W. THEODORE DE BARY (1956, reprinted 1973); *The Japanese Family Storehouse,* trans. by G.W. SARGENT (1959, reprinted 1969); *The Life of an Amorous Woman, and Other Writings,* trans. by IVAN MORRIS (1963, reissued 1969); and *Worldly Mental Calculations,* trans. by BEN BEFU (1976). HOWARD HIBBETT, *The Floating World in Japanese Fiction* (1959, reissued 1975), is a critical study with translations. MAKOTO UEDA, *Matsuo Bashō* (1970, reissued 1982), contains biographical and critical material on the great haiku poet. NOBUYUKI YUASA (trans.), *The Narrow Road to the Deep North, and Other Travel Sketches* (1966), contains all of Bashō's travel accounts. EARL MINER and HIROKO ODAGIRI (trans.), *The Monkey's Straw Raincoat and Other Poetry of the Bashō School* (1981), gives examples of the linked verse. CHIKAMATSU MONZAEMON, *Major Plays* (1961); and TAKEDA IZUMO, MIYOSHI SHŌRAKU, and NAMIKI SENRYŪ, *Chūshingura: The Treasury of Loyal Retainers* (1971), both trans. by DONALD KEENE; TAKEDA IZUMO *et al., Sugawara and the Secrets of Calligraphy,* trans. by STANLEIGH H. JONES, JR. (1985); and MOKUAMI KAWATAKE, *Love of Izayoi and Seishin,* trans. by FRANK T. MOTOFUJI (1966), are representative plays of the Tokugawa period. UEDA AKINARI, *Ugetsu Monogatari: Tales of Moonlight and Rain,* trans. by LEON ZOLBROD (1974), is a collection of stories of the supernatural. *Ryōkan* (1977), and *Grass Hill* (1983), both trans. by BURTON WATSON, contain poems by monks.

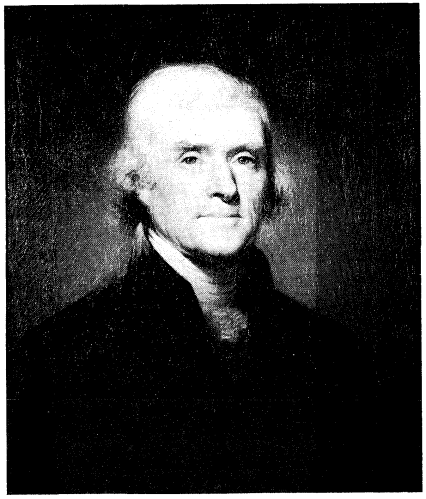
Modern period: DONALD KEENE, *Dawn to the West: Japanese Literature of the Modern Era,* 2 vol. (1984), is a history since 1868. MAKOTO UEDA, *Modern Japanese Writers and the Nature of Literature* (1976), and *Modern Japanese Poets and the Nature of Literature* (1983), are valuable studies. MASAO MIYOSHI, *Accomplices of Silence: The Modern Japanese Novel* (1974, reprinted 1982), is an absorbing study. J. THOMAS RIMER, *Modern Japanese Fiction and Its Tradition: An Introduction* (1978), traces the native elements in modern literature. DONALD KEENE (ed.), *Modern Japanese Literature* (1956); IVAN MORRIS (ed.), *Modern Japanese Stories* (1961, reissued 1970); YUKIO MISHIMA and GEOFFREY BOWNAS (eds.), *New Writing in Japan* (1972); and HOWARD HIBBETT (ed.), *Contemporary Japanese Literature* (1977), are anthologies of modern writing in different genres. TAKAMICHI NINOMIYA and D.J. ENRIGHT (eds.), *The Poetry of Living Japan* (1957, reprinted 1979); ICHIRA KŌNO and RIKUTARO FUKUDA (eds. and trans.), *An Anthology of Modern Japanese Poetry* (1957, reprinted 1971); KENNETH REXROTH and IKUKO ATSUMI (eds. and trans.), *The Burning Heart: Women Poets of Japan* (1977); and *Modern Japanese Poetry,* trans. by JAMES KIRKUP, ed. by A.R. DAVIS (1978), are representative collections. JANINE BEICHMAN, *Masoka Shiki* (1982); and MASAOKA SHIKI, *Peonies Kana: Haiku by the Upasaka Shiki,* trans. and ed. by HAROLD J. ISAACSON (1972), are devoted to the Meiji poet. AMY VLADECK HEINRICH, *Fragments of Rainbows* (1983), is a study of life and poetry of Saitō Mokichi. RICHARD BOWRING, *Mori Ōgai and the Modernization of Japanese Culture* (1979); and J. THOMAS RIMER, *Mori Ōgai* (1975), are studies of this important writer. EDWIN MCCLELLAN, *Two Japanese Novelists* (1969, reissued 1971), discusses major works by Natsume Sōseki and Shimazaki Tōson. EDWARD SEIDENSTICKER, *Kafū the Scribbler* (1965), is a biography of Nagai Kafū with translations from his writings. DENNIS KEENE, *Yokomitsu Riichi, Modernist* (1980), is a study of the chief figure in the Shinkankaku movement. *Modern Japanese Drama,* ed. and trans. by TED T. TAKAYA (1979), contains plays by five outstanding modern dramatists.

(D.Ke.)

Jefferson

Third president of the United States, principal author of the Declaration of Independence, and influential political philosopher, Thomas Jefferson was born on April 13, 1743, at Shadwell, in Albemarle County, Virginia, the son of Peter Jefferson, an early settler and leader in the county, and Jane Randolph Jefferson. Peter Jefferson was a surveyor and cartographer and was largely self-educated. Upon his death in 1757 he left his son considerable property, but the inheritance for which Thomas Jefferson expressed particular gratitude was his father's determination that he should have a sound classical education. After several years of study at local grammar and classical schools, Jefferson entered the College of William and Mary in 1760. In spite of his youth, he became a close friend of three leading residents of Williamsburg—William Small of the college faculty, George Wythe of the Virginia bar, and Francis Fauquier, lieutenant governor of the colony. These three older men gave Jefferson a taste for the pleasures of a society more urbane and sophisticated than that of rural Virginia, and Small and Wythe gave direction to his intellectual drive. Small introduced him to the natural sciences and to rational methods of inquiry; Wythe led him to see the study of law not as a narrow vocational preparation but as a means of understanding the history, culture, institutions, and morals of a people. After two years at the college, Jefferson studied law for five years under Wythe's direction and was admitted to the bar in 1767. In 1769 he entered the lower house of the colonial legislature, thus beginning a long career in politics that ended 40 years later with his retirement as president of the United States.

By courtesy of the White House Collection, Washington, D.C.



Jefferson, oil painting by Rembrandt Peale, 1800. In the White House Collection, Washington, D.C.

Author of the Declaration. When Jefferson entered the House of Burgesses, Virginia and the other colonies were already engaged in the long decade of opposition to British colonial policies that led eventually to revolution and independence. Jefferson joined with Patrick Henry and others who favoured strong resistance to George III and the British Parliament and soon became one of the leaders of this group. His political style was very different from that of Henry. He was assiduous in committee work, a skilled legal craftsman, a scholar who drew on his comprehensive knowledge of law and history to support the colonial case against Great Britain. He rarely made speeches, disliked oral dispute, whether in formal debate or informal conversation, and he recognized the necessity of consensus for effective political action; the pen was his natural means of expression, and he was a virtuoso in its use. His first major essay, "A Summary View of the Rights of British America" (1774), displayed an impressive array of learning and logic, demonstrated his capacity for intense passion and the ability to express it eloquently, and revealed an inclination to intellectual radicalism. The majority of his colleagues were not then prepared for his conclusion that the British Parliament had no authority at all to legislate for the colonies, but, as relations with Great Britain grew steadily worse, his arguments became increasingly acceptable and his language both persuasive and provocative. "The God who gave us life gave us liberty at the same time: the hand of force may destroy, but cannot disjoin them."

Delegate to the Second Continental Congress

In the spring of 1775 the Virginia legislature, sitting as a revolutionary convention in defiance of the royal governor, appointed Jefferson as a member of its delegation to the Second Continental Congress meeting in Philadelphia. There he joined with the more radical group in the Congress, and again his skills as a committeeman and stylist were recognized and used. In June of 1776, when the decision to break irrevocably with Great Britain seemed near, Jefferson was appointed to the committee assigned to draft a formal statement of the reasons for such a decision. Benjamin Franklin and John Adams were also on the committee, but they recognized the superior talent of the Virginian and gracefully bowed to it. Jefferson thus became the principal author of the Declaration of Independence. It was an official state paper, and in later life he stated that it was intended to be an expression of the American mind. That was no doubt true, but it is also true that his personal commitment to its principles was profound and intense. It was this commitment, not the mere fact of literal authorship, that rendered Jefferson uniquely symbolic of the ideals expressed in the Declaration.

Role in Virginia politics. Jefferson meant his revolutionary manifesto to be more than an eloquent justification of revolt against Great Britain. He intended to translate its principles into practice and to create in America a society in which the gap between aspiration and achievement would be narrowed. He had wanted to begin by taking part in framing the new constitution of Virginia, which was adopted in June of 1776, but his duties in Philadelphia made that impossible, and he did not enter the Virginia legislature until October. He then set in motion a plan for comprehensive reform of the laws and institutions of Virginia. Two parts of the plan show the thoroughness with which he had considered the nature of representative government and the conditions necessary to its successful operation. A third embodied his passionate commitment to intellectual freedom.

Jefferson sought and secured abolition of the laws of primogeniture and entail in Virginia in order to discourage concentration of property in the hands of a few great landowners. He believed that property was among the natural rights to which man was born and that it meant the right to a decent means of subsistence. After observing the economic conditions in France a few years later, he wrote:

> Whenever there is in any country, uncultivated lands and unemployed poor, it is clear that the laws of property have been so far extended as to violate natural right. The earth is given as a common stock for man to labour and live on. If for the encouragement of industry we allow it to be appropriated, we must take care that other employment be provided to those excluded from the appropriation. If we do not the fundamental right to labour the earth returns to the unemployed.

No society that denied this right could be just, nor was it likely to enjoy for long a republican government. Jefferson believed that the virtues required for that form of government could not flourish in conditions of extreme poverty or complete economic dependence.

The educational system proposed for Virginia was also a part of Jefferson's comprehensive plan for republican government. The lower schools would provide literacy for the entire population, which, combined with a free press, was necessary for an informed public opinion. The upper schools would develop a natural aristocracy to supply the leadership so essential to representative government, while scholarships awarded on the basis of merit would prevent identification of educational opportunity with economic privileges. Jefferson did not believe that an ignorant people could make rational and responsible decisions about public affairs, nor did he believe that men were equal in intelligence or that the operation of a government was a simple job easily mastered by the common man. He assumed that men of superior capabilities were those naturally suited for public office, and his scheme of education was intended to insure that such men, regardless of their economic circumstances, be given an opportunity to develop their talents. Jefferson's fellow Virginians were not prepared for so comprehensive a system of free public education, however, and the only part of it that he secured was the University of Virginia.

The statute of Virginia

The third and most famous reform, the statute of Virginia for religious freedom, met with bitter and persistent opposition and was not enacted until 1786, while Jefferson was in France. Although Americans had largely abandoned the gross forms of persecution common a few generations earlier, the toleration they practiced was limited and erratic. In some states, as in Virginia, a single church was established; others restricted public office to Protestants; some required belief in specific doctrines of the Christian religion, such as the divinity of Jesus, the Trinity, and immortality. The Virginia statute constituted a complete break with the traditional relationship between church and state. It prohibited support of any religion by public taxation and forbade all civil disabilities imposed on citizens because of religious belief or the lack of it. Jefferson regarded the statute as partial fulfillment of his celebrated vow: "I have sworn upon the altar of God eternal hostility against every form of tyranny over the mind of man."

After three years in the legislature, Jefferson was elected governor in 1779 and served for two years in a position characterized by much responsibility and very little power. When Virginia was invaded by British forces in the winter of 1780–81, Jefferson was unable to organize effective opposition and barely escaped capture when a detachment of troops raided Charlottesville and Monticello. His conduct during the emergency was criticized, and, although the legislature gave him a unanimous vote of confidence, he could not forget the slur cast upon his character as a public official. He refused to serve again either as governor or legislator and retired to Monticello determined to live out his life as a private citizen.

There was a reason other than wounded pride for this decision. He was worried about the health of his wife, Martha Wayles Skelton Jefferson. Since their marriage in 1772, she had borne him five children of whom only two survived, and in the fall of 1781 she was again pregnant. Jefferson's fears were justified, for she did not recover strength after the birth of the sixth child and died September 6, 1782. Jefferson's grief was incalculable.

"Notes on Virginia." After his retirement as governor and before he returned to public service in December of 1782, Jefferson wrote and revised the major portion of *Notes on Virginia,* his only book. It originated in a comprehensive but routine series of questions put to him by the secretary of the French legation in order to compile information about the new country. Jefferson's response was as revealing of himself as it was informative about the state of Virginia. In later years he learned to guard his pen carefully, especially after letters he considered to be purely private were printed in newspapers or elsewhere without his permission. The language of this book was frequently unrestrained. It was as if the *Notes,* written for the most part after his abrupt and unhappy withdrawal from Virginia politics and during the months of desperate fear for the life of his wife, provided a means for the release of otherwise restrained emotions.

Discussion of race and slavery in the *Notes*

The *Notes* include a discussion of slavery, its effects on both whites and blacks, and an attempt to delineate the racial characteristics of the latter. Although he was unalterably opposed to slavery and reiterated his reasons in this essay, he both expressed and reflected one of the principal obstacles to abolition—the belief that, because of inherent racial differences, blacks and whites could not live together in peace and harmony. Jefferson's summary of the supposed differences may now be seen as a classic example of the failure of an individual mind—and in this case one of exceptional independence and critical rigour—to transcend the cultural boundaries of its age. It is a curious blend of attempted objectivity flawed by the intrusion of unconscious prejudices and unexamined assumptions. He argued, among other points, that the blacks were inferior in physical beauty, that they might be lacking in foresight, that they were equal in memory but inferior in reason and imagination to the white race. He was aware of the influence of environment on behaviour and belief, accepted it as a general principle, and even cited it to explain the slave's alleged disposition to theft. Yet, he could not or did not apply it consistently and rigorously throughout his examination of the subject. It would appear that he clearly recognized the difficulties involved in applying the methods of scientific analysis to problems of racial characteristics, but they were difficulties beyond his power to resolve.

> The opinion that they are inferior in the faculties of reason and imagination, must be hazarded with great diffidence. To justify a general conclusion, requires many observations, even where the subject may be submitted to the anatomical knife, to optical glasses, to analysis by fire or by solvents. How much more then where it is a faculty, not a substance, we are examining; where it eludes the research of all the senses; where the conditions of its existence are various and variously combined; where the effects of those which are present or absent bid defiance to calculation; let me add too, as a circumstance of great tenderness, where our conclusion would degrade a whole race of men from the rank in the scale of beings which their Creator may perhaps have given them. To our reproach it must be said, that though for a century and half we have had under our eyes the races of black and of red men, they have never yet been viewed by us as subjects of natural history. I advance it, therefore, as a suspicion only, that the blacks, whether originally a distinct race, or made distinct by time and circumstances, are inferior to the whites in the endowments both of body and mind.

The *Notes* are otherwise interesting because they reveal the mind of a revolutionist in the midst of a revolution he regarded as unfinished. With some equanimity, he attributed the "very capital errors" in the Virginia constitution of 1776 to inexperience; it was with passionate outrage that he criticized proposals made twice in the Virginia legislature to follow Roman precedent and establish a temporary dictator in time of emergency:

> The very thought alone was treason against the people; was treason against mankind in general; as riveting forever the chains which bow down their necks, by giving to their oppressors a proof which they would have trumpeted through the universe, of the imbecility of republican government, in times of pressing danger, to shield them from harm.

He urged revision of the constitution and enactment of his plans for universal education and full freedom of religion because he believed that the public virtue then prevalent among both the people and their leaders was impermanent, in part a function of the revolutionary situation, and destined to diminish. Rulers would become corrupt and abuse their power, and the people "will forget themselves, but in the sole faculty of making money, and will never think of uniting to effect a due respect for their rights." Jefferson's belief in republican government did not rest on naïve and unqualified faith in the people. Republican government would operate successfully only under certain conditions: a wide distribution of property or the availability of a substitute that provided men with a decent subsistence honestly earned; an educated and informed population; laws and institutions designed to compensate for the diminution of public virtue that Jefferson thought was sure to come when the crises of the revolutions were over.

Return to politics. In December 1782 he returned to public service and was for several months a member of the Virginia delegation to the Continental Congress. During this time Virginia ceded to the national government the area northwest of the Ohio River, which it claimed under grants made during the colonial period. In an ordinance drafted for the governance of this land, Jefferson set forth the principle that it should not be held by the original 13 states as colonial territory but should be divided into areas that, upon reaching a designated condition of population and organization, should enter the Union as states equal to the original 13. He also included a prohibition that would have forbidden slavery after 1800 in this territory and any others of which the United States might become possessed. The provision was defeated by one vote; a similar one had been incorporated in the Northwest Ordinance of 1787, but it applied only to that territory. Had Jefferson's original proposal been adopted, and had it remained in force, then slavery would have been outlawed in the whole area of the Louisiana Purchase. As he himself later commented,

> Thus we see the fate of millions of unborn hanging on the tongue of one man, and heaven was silent in that awful moment.

Visit to France

In 1784 Jefferson went to France to join Benjamin Franklin and John Adams in negotiating treaties with European powers. After a few months he succeeded Franklin as resident U.S. minister to the French government. His diplomatic duties were not onerous, and Paris offered him the intellectual and artistic society he had first glimpsed as a student at the College of William and Mary. There he could attend the theatre and opera, visit museums, keep up with science and inventions, associate freely with European scientists and intellectuals, share the *politesse* of French society, and indulge his passion for books. He loved France and the French, but not uncritically. His observations of economic and social conditions strengthened his aversion to absolute monarchy, and the contrast he saw between French and U.S. domestic morality led to a series of letters condemning the former and warning against the dangers of corruption should young men of his own country be sent to France for their education. (He did not want his daughters to marry abroad and so took them back to Virginia when the older was 17.) As author of the Declaration of Independence and of the Statute for Religious Freedom of Virginia, he had considerable influence with such moderate political leaders as the Marquis de Lafayette, and during the early stages of the French Revolution he was optimistic about the future of their efforts to effect gradual changes in the monarchy and its attendant laws and institutions. It was the greatest intellectual error of his life: France had almost none of the ingredients that had contributed to the success of the United States War of Independence, a fact Jefferson would surely have realized had he not allowed himself to indulge in wishful thinking. Jefferson observed only the opening stages of the Revolution, for he returned to the United States at the end of 1789.

Controversy with Hamilton. In the meantime, the Articles of Confederation had been replaced by the Constitution drafted in Philadelphia in 1787 and ratified the following year. Jefferson approved of most of that document but was critical of its lack of a bill of rights and its failure to impose limitations on the length of tenure for the presidency. Upon his return to Virginia in the fall of 1789, he was requested by George Washington to become secretary of state in the new government. With considerable reluctance, he accepted. Soon after he assumed the new office he became involved in controversy with Alexander Hamilton, who was secretary of the treasury. He opposed Hamilton's financial policies on the grounds that they exceeded the powers delegated to the central government by the Constitution, that they were contrary to the interests of the majority of the people, and that they represented a threat to republican institutions. Jefferson and Hamilton also disagreed on questions of foreign policy, with Jefferson at first leaning toward France and Hamilton toward Great Britain.

The issues between the two men were not purely personal; they extended to the country at large and led to the formation of national political parties based on policy and principle as well as personality. Thus was established the precedent and pattern of a national two-party system. Both Jefferson and Hamilton retired from the Cabinet near the end of Washington's first term, but each continued to be the symbol of the new parties, Jefferson of the Democratic-Republican, Hamilton of the Federalist. Both sides developed organizational skills among the electorate, the Congress, and the state legislatures, and both made effective use of the press. James Madison was, as usual, Jefferson's able collaborator and supplied active leadership of the party until the latter returned to the centre of national politics as vice president under John Adams in 1797. In 1798, when the United States was close to war with France, the Federalist-controlled Congress enacted the Alien and Sedition Acts. The latter, particularly as applied by Federalist judges, was used to stifle Democratic-Republican criticism of the government. Jefferson and Madison believed it to be contrary to the first amendment and therefore unconstitutional, a position they argued in the Virginia and Kentucky Resolutions of 1798–99.

Role in the development of the two-party system

The decade ended with the defeat of the Federalists in the election of 1800. It was a critical period in the development of the new nation; politics were sharply divisive, conducted with extreme animosity, and permeated with fundamental cleavages over political principles. Jefferson regarded Hamilton as an enemy of republican government; Hamilton regarded Jefferson as a demagogic radical. Hamilton had a dream of national grandeur to which he was prepared to subordinate the immediate interests of the individual. Jefferson saw the purpose of government as the protection of the individual's right to life, liberty, and the pursuit of happiness. Jefferson's attitudes and behaviour during this period were revealing. He did not exercise an Olympian calm; his letters sometimes displayed anger and passion toward the policies of his opponents and toward some of them personally. At the same time, he sensed and feared the divisive and destructive effects of unrestrained ideological conflict. Not only could the latter disrupt the social harmony that Jefferson valued so highly, but it could also conceivably rip the fabric of republican government altogether. A desire to forestall in America what had so frequently been the fate of such governments in the past seemed to influence Jefferson's conduct of the presidency during his first term.

Presidency. The Federalist candidates clearly lost the presidential election of 1800, but under the electoral system then prevailing neither of the Republican candidates, Jefferson and Aaron Burr, could claim victory. The Constitution had provided no means for electors to distinguish between their choices for president and vice president, and both candidates had received the same number of votes. The choice between them was therefore made in the House of Representatives. Partly because of the influence of Hamilton, who distrusted Burr even more than he disliked Jefferson, the latter was chosen president and inaugurated March 4, 1801.

The spirit and content of Jefferson's inaugural address were conciliatory, and so, to a considerable extent, were the policies of his first administration. There was no attempt at wholesale reversal of Federalist policies of the preceding 12 years, and in at least two instances—the Louisiana Purchase and the Embargo Act—he was said to be even more Federalist than the Federalists themselves. There was, however, an effort to nullify the Federalist attempt to fill the federal judiciary with partisan appointees holding office for life, and there was sufficient turnover in other federal offices to give some substance to the accusation that Jefferson introduced the spoils system. But, in spite of the very bitter controversy of the preceding years, Jefferson's inauguration ushered in no drastic or radical changes. Had Jefferson been more doctrinaire or less aware of the danger of unrestrained political passion and of the delicate situation created by the first party change of administration in the new government's history, the future of U.S. politics might have been characterized by less stability than has been the case. The precedent he deliberately set must rank with the Louisiana Purchase as one of the major achievements of his presidency.

The Louisiana Purchase

The acquisition of the Louisiana Territory in 1803 was of incalculable importance, nearly doubling the size of the United States. Jefferson's original plan was to purchase merely a small area at the mouth of the Mississippi River. When Napoleon offered to sell the entire territory, Jefferson saw his chance and took it, even though, as he frankly admitted, he had no constitutional authority to do so. He believed that the purchase would contribute to the security of the United States by removing from the continent a major foreign power and that it would ensure the stability of republican government for generations to come by providing a vast reservoir of land for settlers.

Jefferson was re-elected in 1804; George Clinton replaced Burr as vice president. Jefferson's second administration was notable for his unsuccessful efforts to convict his former vice president, Burr, of treasonable acts in the southwestern territories, and for his efforts to pursue a policy of neutrality during the Napoleonic Wars and maintain the rights of neutrals on the high seas. His overwhelming desire to avoid war with either side led to charges of timidity and vacillation, and his Embargo Act (1807) was criticized as inconsistent with the principles of individual freedom and his former opposition to a strong national government. The act was securely based on the power given to the Congress to regulate commerce with foreign nations—a power of which Jefferson approved long before he became president—but the enforcement provisions of the act and its amendments can rightly be questioned as contravening the Fourth Amendment's prohibition of unreasonable search and seizure.

During Jefferson's presidency the power and prestige of the Supreme Court grew under the leadership of Chief Justice John Marshall. In the case *Marbury* v. *Madison* (1803), the court explicitly asserted the right and power of judicial review. Jefferson opposed the power of the court as the ultimate and exclusive interpreter of the Constitution and argued that such a power lodged in one department of the government whose members held office for life was irresponsible and therefore contrary to the principles of republican government.

Establishment of the University of Virginia

Jefferson might have been elected president for a third term but chose to follow Washington's example of withdrawing after two terms. On March 4, 1809, he turned the office over to his successor, James Madison, and went home to Monticello. There was one more official act he sought to accomplish, the establishment of the University of Virginia, to which he referred as "the last of my mortal cares, and the last service I can render my country." He designed the buildings and supervised their construction to the most minute detail; he gathered the faculty, planned the curriculum, and even selected the reading for some of the courses. He had never been able to persuade his fellow Virginians to support public education for elementary and secondary pupils, but the university was an appropriate conclusion to a political career remarkable for its creativity as well as for its duration and success.

Jefferson's political career was undoubtedly impressive, but it was far from absorbing all of the energy, time, and talent of the man himself. He probably enjoyed politics more than he was willing to admit; it is also true that his often-expressed longing to retire to private life and pursue his other interests was very real. These interests were numerous and varied.

Personal and intellectual interests. He was an extraordinarily learned man, and the range of his knowledge and inquiry is scarcely credible in the modern age of specialization. He knew Latin, Greek, French, Spanish, Italian, and Anglo-Saxon and concerned himself with such questions as the difference between the ancient and modern pronunciation of Greek. At the age of 71 he tackled Plato's *Republic* in the original and found its author greatly overrated. He attempted an analysis of the New Testament in order to discover what Jesus really said as distinguished from what he was reported to have said. He enjoyed the study of mathematics and found its precision and certitude a welcome relief from the untidiness of politics and government. He was an ardent student of the natural sciences, carried on an extensive correspondence with such men as Joseph Priestley, and sometimes contributed time and money to progress in these fields. The discovery of fossil remains in various parts of the country fascinated him, and he tried to collect and classify as many as he could. He was much interested in the experiments with balloons and submarines then being made, and, while he was abroad, he sent back to his friends at home various mechanical and scientific gadgets produced in Europe, including a polygraph and phosphorus matches. His travel notes record impressions ranging from nearly ecstatic admiration of architectural monuments to sober economic analysis of the reasons for the differences in prosperity between regions producing white and red wine.

Interest in the natural sciences

He was an enthusiastic practitioner of scientific farming, conducted numerous experiments at Monticello, was always on the lookout for some new plant or seed that might contribute to the prosperity of the United States (once going so far as to smuggle a particular variety of rice across the Italian border); kept meticulous meteorological records; and, as a keen linguist, instigated the first systematic collection of American Indian dialect. His interest in architecture was intense and enduring, and his influence on the Neoclassical style in the United States was great.

The pursuit of these various interests concurrently with his political activities and the management of his estates (which included several thousand acres and at one time about 150 slaves) is remarkable. To this record of industry must be added the voluminous correspondence Jefferson maintained with extraordinary conscientiousness until very near his death. He could have accomplished so much only through rigorous self-discipline and an efficient organization of his time and activities. Yet, he was one of the most generous and approachable of men. Friends and strangers alike wrote to him for advice or came to Monticello when he was in residence. Relatives and guests filled Monticello to capacity—sometimes beds were made for as many as 50 people—and devoured his food as well as his time. For privacy he retreated several times a year to Poplar Forest, a second home built as a refuge in Bedford County.

Personal appearance and personality

Jefferson was 6 feet 2 inches in height, large boned, slim, erect, and sinewy. He had angular features, a ruddy complexion, sandy hair, and hazel-flecked gray eyes. His carriage was relaxed and somewhat awkward, and by 18th-century standards he seems to have been regarded as pleasant rather than handsome in appearance. He was sensitive and perceptive in personal relations, gracious and charming in manner (though sometimes cold upon first meeting strangers), and almost invariably even tempered. As a matter of both principle and inclination, he attempted to prevent political differences from creating personal ill will, and though he was subjected to malicious abuse during the political controversies in which he was involved, he endured it with relative equanimity and felt genuine animosity toward only a very few of his opponents and critics.

Because he was so central a figure, so widely known, so articulate, and so meticulous in preserving his letters and papers, it is possible to reconstruct a remarkably complete account of his career and his work. Yet, the man himself—the private man—remains elusive. There was a reserve of privacy that he kept inviolate. For example, no letters exchanged between him and his wife exist. Their marriage was, by contemporary accounts, an extraordinarily happy one, and it would therefore appear that Jefferson destroyed whatever letters once existed in order to keep their relationship forever private. Jefferson was, as his modern editor has suggested, ultimately a lonely man.

Ten days before his death, Jefferson replied to an invitation to join the residents of Washington, D.C., in celebrating the 50th anniversary of the proclamation of the Declaration of Independence. He could not attend because of illness, but he sent his best wishes, and, of the Declaration that was to be celebrated, he wrote:

> May it be to the world, what I believe it will be, (to some parts sooner, to others later, but finally to all,) the signal of arousing men to burst the chains under which monkish ignorance and superstition had persuaded them to bind themselves, and to assume the blessings and security of self government.

While Jefferson grew steadily weaker at Monticello, his

Deaths of Jefferson and Adams

old friend John Adams was nearing death in Massachusetts. It seems certain from the accounts of friends and relatives of both that each man wanted badly to live until the 50th anniversary of the day that symbolized the central endeavour and achievement of their lives. They succeeded. Jefferson died shortly before one o'clock on the afternoon of July 4, 1826; Adams died a few hours later, his last words said to have been, "Jefferson still survives." Jefferson was buried at Monticello. The epitaph that he had chosen was inscribed on his tombstone: "Here was buried Thomas Jefferson, author of the Declaration of American Independence, of the statute of Virginia for religious freedom, and father of the University of Virginia." (C.M.K.)

BIBLIOGRAPHY. JULIAN P. BOYD (ed.), *The Papers of Thomas Jefferson*, 19 vol. (1950–74), covering the period up to March 10, 1791, is the definitive edition of Jefferson's papers. It includes extensive notes on the background, context, and significance of the documents printed, among which are papers written to Jefferson as well as those written by him. This edition, when completed, will comprise 60 volumes. Two other collections of Jefferson's writings can be used for the period not yet reached by the Boyd edition: ANDREW A. LIPSCOMB and ALBERT ELLERY BERGH (eds.), *The Writings of Thomas Jefferson*, 20 vol. in 10 (1905); and PAUL LEICESTER FORD (ed.), *The Works of Thomas Jefferson*, 12 vol. (1904–05). A selection is presented in MERRILL D. PETERSON (ed.), *The Portable Thomas Jefferson* (1975, reissued 1977). The correspondence between Jefferson and John and Abigail Adams is reproduced in LESTER J. CAPPON (ed.), *The Adams–Jefferson Letters*, 2 vol. (1959, reprinted 1971). The letters are notable for their warmth and, especially those between the two men after 1812, for their discussions of the intellectual and political developments of the times. Examinations of their friendship are offered in JOHN M. ALLISON, *Adams and Jefferson: The Story of a Friendship* (1966); and MERRILL D. PETERSON, *Adams and Jefferson: A Revolutionary Dialogue* (1976, reprinted 1978). Information about Jefferson's intellectual life may be found in E. MILLICENT SOWERBY (comp.), *Catalogue of the Library of Thomas Jefferson*, 5 vol. (1952–59, reprinted 1983), which contains a list of the books sold by Jefferson to the Congress in 1815, replacing the library burned by the British and forming the nucleus of the present Library of Congress. A convenient single-volume anthology of Jefferson's letters and papers is ADRIENNE KOCH and WILLIAM PEDEN (eds.), *The Life and Selected Writings of Jefferson* (1944, reissued 1982). EDWIN M. BETTS and JAMES A. DEAR (eds.), *The Family Letters of Thomas Jefferson* (1966), includes some 570 letters to and from his children and grandchildren. The variety of Jefferson's interests is revealed in VIRGINIA. UNIVERSITY. LIBRARY, *The Jefferson Papers of the University of Virginia* (1973), comprising more than 3,000 items; EUGENE L. HUDDLESTON, *Thomas Jefferson: A Reference Guide* (1982), is an annotated bibliography.

The definitive biography is that by DUMAS MALONE, *Jefferson and His Time*, 6 vol. (1948–81). A good popular biography is NATHAN SCHACHNER, *Thomas Jefferson: A Biography*, 2 vol. (1951, reissued in 1 vol., 1960). A comprehensive single-volume biography, unfortunately published without footnotes, is MERRILL D. PETERSON, *Thomas Jefferson and the New Nation* (1970). The 19th-century biography by HENRY S. RANDALL, *The Life of Thomas Jefferson*, 3 vol. (1858, reprinted 1972), is valuable because of Randall's extensive consultation with people then living who had known Jefferson personally. The influence of Jefferson in America is treated in MERRILL D. PETERSON, *The Jefferson Image in the American Mind* (1960). Other biographical works include JOSEPH C. FARBER and WENDELL D. GARRETT, *Thomas Jefferson Redivivus* (1971), a book of photographs of places Jefferson knew; THOMAS FLEMING, *The Man from Monticello: An Intimate Life of Thomas Jefferson* (1969); GENE GURNEY and CLARE GURNEY, *Monticello* (1966), an illustrated history of Jefferson's home; and THOMAS JEFFERSON, *Thomas Jefferson: A Biography in His Own Words*, by the eds. of *Newsweek Books* (1974).

Jefferson as an artist emerges in several works: HELEN CRIPE, *Thomas Jefferson and Music* (1974); DESMOND GUINNESS and JULIUS T. SADLER, *Mr. Jefferson, Architect* (1973); WILLIAM H. ADAMS (ed.), *Jefferson and the Arts: An Extended View* (1976); PAGE SMITH, *Jefferson: A Revealing Biography* (1976); LALLY WEYMOUTH (ed.), *Thomas Jefferson: The Man, His World, His Influence* (1973); and HOWARD C. RICE, *Thomas Jefferson's Paris* (1976).

For analyses of Jefferson's attitudes toward race and slavery, see DAVID B. DAVIS, *The Problem of Slavery in the Age of Revolution, 1770–1823* (1975); FAWN M. BRODIE, *Thomas Jefferson: An Intimate History* (1974), a biography developing the thesis that Jefferson's slave Sally Hemings was his mistress; BARBARA CHASE-RIBOUD, *Sally Hemings* (1979), a historical novel; VIRGINIUS DABNEY, *The Jefferson Scandals: A Rebuttal* (1981), which argues against Brodie and Chase-Riboud; JOHN C. MILLER, *The Wolf by the Ears: Thomas Jefferson and Slavery* (1977, reprinted 1980), an examination of contradictions in Jefferson's approach to slavery that disputes Brodie's biography; and ERIK H. ERIKSON, *Dimensions of a New Identity* (1974), an exploration of Jefferson's opposition to slavery as a characteristic of American identity. Jefferson's views on American Indians are covered in FREDERICK M. BINDER, *The Color Problem in Early National America as Viewed by John Adams, Jefferson and Jackson* (1968); and BERNARD W. SHEEHAN, *Seeds of Extinction: Jeffersonian Philanthropy and the American Indian* (1973, reprinted 1974).

Interpretive studies include GARRY WILLS, *Inventing America: Jefferson's Declaration of Independence* (1978); EDMUND S. MORGAN, *The Meaning of Independence: John Adams, George Washington, Thomas Jefferson* (1976, reprinted 1978); JONATHAN DANIELS, *Ordeal of Ambition: Jefferson, Hamilton, Burr* (1970); LANCE BANNING, *The Jeffersonian Persuasion: Evolution of a Party Ideology* (1978, reprinted 1980), an exploration of the roots of American opposition. Several monographs examine Jefferson's administration: FORREST MCDONALD, *The Presidency of Thomas Jefferson* (1976); ROBERT M. JOHNSTONE, *Jefferson and the Presidency: Leadership in the Young Republic* (1978); RICHARD E. ELLIS, *The Jeffersonian Crisis: Courts and Politics in the Young Republic* (1971, reprinted 1974); NOBLE E. CUNNINGHAM, *The Process of Government Under Jefferson* (1978); HENRY STEELE COMMAGER, *Jefferson, Nationalism, and the Enlightenment* (1975). Foreign policy and American expansion are discussed in BURTON SPIVAK, *Jefferson's English Crisis: Commerce, Embargo, and the Republican Revolution* (1979); LAWRENCE S. KAPLAN, *Jefferson and France: An Essay on Politics and Political Ideas* (1967, reprinted 1980); GEORGE DARGO, *Jefferson's Louisiana: Politics and the Clash of Legal Traditions* (1975); and DONALD JACKSON, *Thomas Jefferson & the Stony Mountains: Exploring the West from Monticello* (1981).

Jerusalem

Jerusalem (Hebrew Yerushalayim; Arabic Bayt al-Muqaddas, or al-Quds), one of the world's oldest and holiest cities, was in December 1949 proclaimed by the State of Israel to be its capital. The city plays a central role in the spiritual and emotional perspective of the three major monotheistic religions. For Jews throughout the world, Jerusalem is the focus of age-old yearnings, a living proof of ancient grandeur and independence and a centre of national renaissance; for Christians, it is the scene of their Saviour's agony and triumph; for Muslims, it is the goal of the Prophet Muḥammad's mystic night journey and the site of one of Islām's most sacred shrines. For all three faiths it is a centre of pilgrimage—the Holy City, the earthly prototype of the heavenly Jerusalem.

From 1948 until 1967, Jerusalem was divided into Israeli (West Jerusalem) and Jordanian (East Jerusalem) sectors, with the Israeli sector of the city becoming the capital of Israel. During the Six-Day War of June 1967, however, Israel occupied the former Jordanian sector, over which it proclaimed jurisdiction as an integral part of the unified city. Its standing as capital of the nation was reaffirmed by a special Israeli law passed in 1980. Since 1975 the unified Jerusalem has been Israel's largest city.

An outstanding characteristic of Jerusalem, which covers an area of 42 square miles (109 square kilometres), is the variety of its people and culture. The Old City has Jewish, Christian, Armenian, and Muslim quarters. The Jewish quarter suffered during the 1947–48 fighting but since has been completely rebuilt. Its historical synagogues have been restored, and the new residential quarters, though modern, preserve some of their old Oriental atmosphere. The old Jewish neighbourhoods outside the Old City, on the other hand, reflect much of the atmosphere brought from Jewish habitats elsewhere in the Orient, as well as in eastern Europe. Similarly, many of the Christian institutions made direct copies of the architecture common to their native lands.

Arabs in traditional and modern dress; Christians, Western and Oriental, in their infinite variety of secular and monastic vestments; Jews in fashionable and Orthodox dress; and hosts of tourists combine in colourful, kaleidoscopic patterns. Synagogues, churches, mosques, and dwellings in various styles make up the city's unique architectural mosaic. The scent of Oriental cooking and spices, the peal of church bells, the calls of muezzins from minarets, and the chanting of Jewish prayers at the Western (Wailing) Wall all add a particular tinge to the life of the city. These impressions, however, are in a large measure limited to the Old City. Outside the walls Jerusalem is in every sense a modern city with its network of streets and transportation, high-rise buildings, supermarkets, businesses, schools, and restaurants and coffeehouses. It is the persistent mingling of Hebrew, Arabic, and English in the streets that brings to mind the multicultural and political complexities of life in this revered city.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. On the east, Jerusalem looks down on the Dead Sea and across the Jordan River to the arid mountains of Moab; on the west, it faces the coastal plain and the Mediterranean Sea, about 35 miles (58 kilometres) away. The main north–south road bisects the city in its course along the watershed between the coastal plain and the Great Rift Valley of the Jordan River and links Nābulus (ancient Shechem) to the north with Bethlehem, Hebron, and Beersheba to the south. A major road links Jerusalem with Jericho, about 36 road miles to the east, and hence along the Jordan to the Sea of Galilee in the north. The (Yigal) Allon Road cuts across the Judaean Desert, linking with the new Israeli settlements in Samaria. The west–east road from Tel Aviv–Yafo, 58 miles to the west, crosses the Jordan north of the Dead Sea and runs to Amman, Jordan (biblical Rabbah), about 60 miles to the east; the Tel Aviv–Jerusalem sector has been developed into a modern four-lane highway. A newer transversal road leads westward from Jerusalem, eventually converging on the Ben-Gurion Airport at Lod.

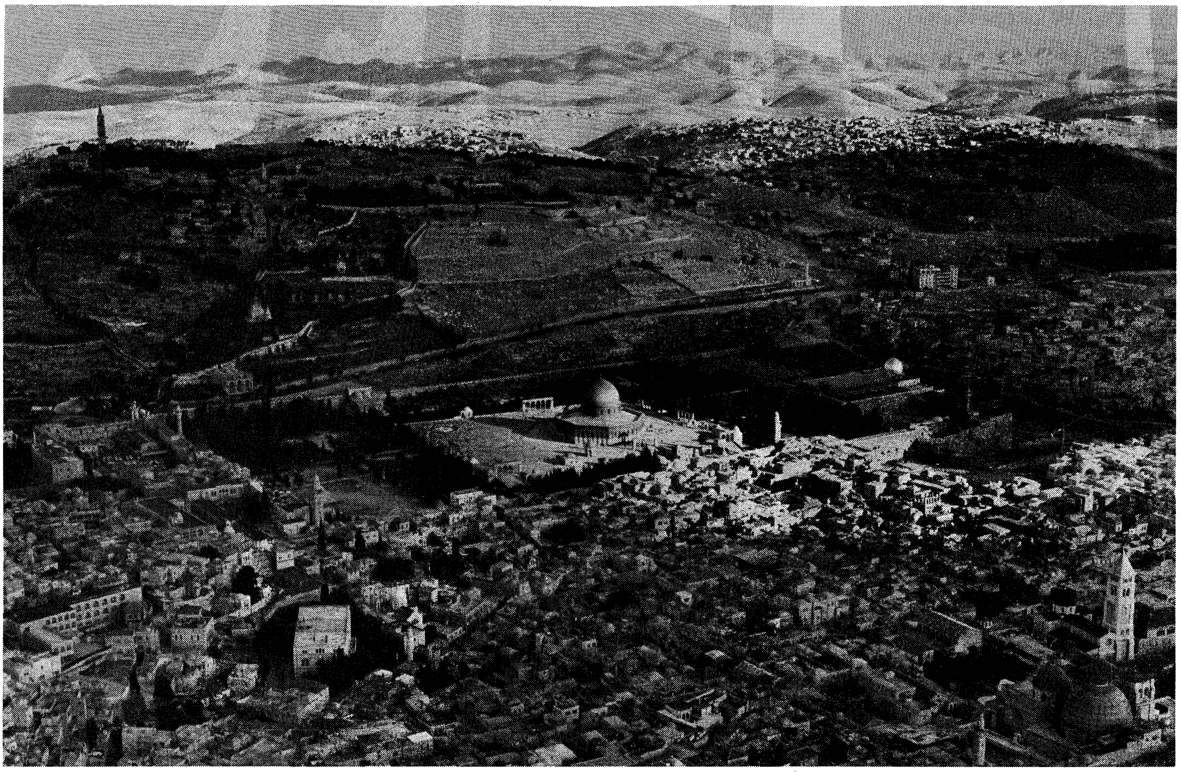
Climate. Jerusalem has a mixed subtropical, semiarid climate with warm, dry summers and cool, rainy winters. The average annual rainfall is about 20 inches (500 millimetres), and snow falls every two or three years. Average temperatures range from about 75° F (24° C) in August to about 50° F (10° C) in January. The hot desert wind, called *sharav* (*khamsīn*), is fairly common in autumn and spring. Average daily humidity is about 62 percent in the daytime but may drop 30–40 percent under *sharav* conditions. Summer exposure to the Sun's rays in Jerusalem is among the most intense on the globe, attributable partly to the lack of clouds or humidity and partly to the angle of the Sun (80°) over the horizon at that season.

Jerusalem has no serious air pollution. Its altitude ensures the free mixing of surface air, and pollutant sources are few, for there is little heavy industry.

Plant and animal life. Lying on the watershed between the relatively rainy Har Yehuda ("Hills of Judaea") and the dry Judaean Desert, Jerusalem has both Mediterranean and Irano-Turanian vegetation. The various red and brown Mediterranean soils, formed by the different types of limestone chalk covering the hills, support as many as 1,000 plant species.

There is a great variety of birds, including 70 resident species and about 150 winter visitors. Those most commonly seen are the hooded crow, jay, swift (which nests in old walls and buildings), and bulbul. The only venomous snake is the Palestine viper; the smooth lizard and common chameleon frequent gardens.

The city layout. The municipal boundaries, defined in 1967, stretch from the Jerusalem Airport in the north to



The Old City of Jerusalem. In the centre is the Dome of the Rock; the Dead Sea is in the background.
Werner Braun

a point almost reaching Bethlehem in the south and from the ridge of Mt. Scopus and the Mount of Olives in the east to Mt. Herzl, ʿEn Kerem, and the Hadassah Medical Centre of Hebrew University in the west.

The Old City

The Old City, which is believed to have been continuously inhabited for almost 5,000 years, forms a walled quadrilateral about 3,000 feet (900 metres) long on each side. It is dominated by the raised platform of the Temple Mount (Hebrew Har ha-Bayt), the site of the First and Second Temples, known to Islām as al-Ḥarām ash-Sharīf ("The Noble Holy Place"). The rest of the area within the walls—divided by the ancient street layout into Muslim, Christian, Jewish, and Armenian quarters—is a typical Oriental city, with its mosques and its medieval vaulted triple bazaar in the centre and a labyrinth of smaller suqs, or bazaars, along David Street, which leads from Jaffa Gate and the Citadel toward the Temple Mount. The Old City is distinguished by its many churches and by the ancient synagogues and study houses of the Jewish Quarter.

Neighbourhoods outside the walls

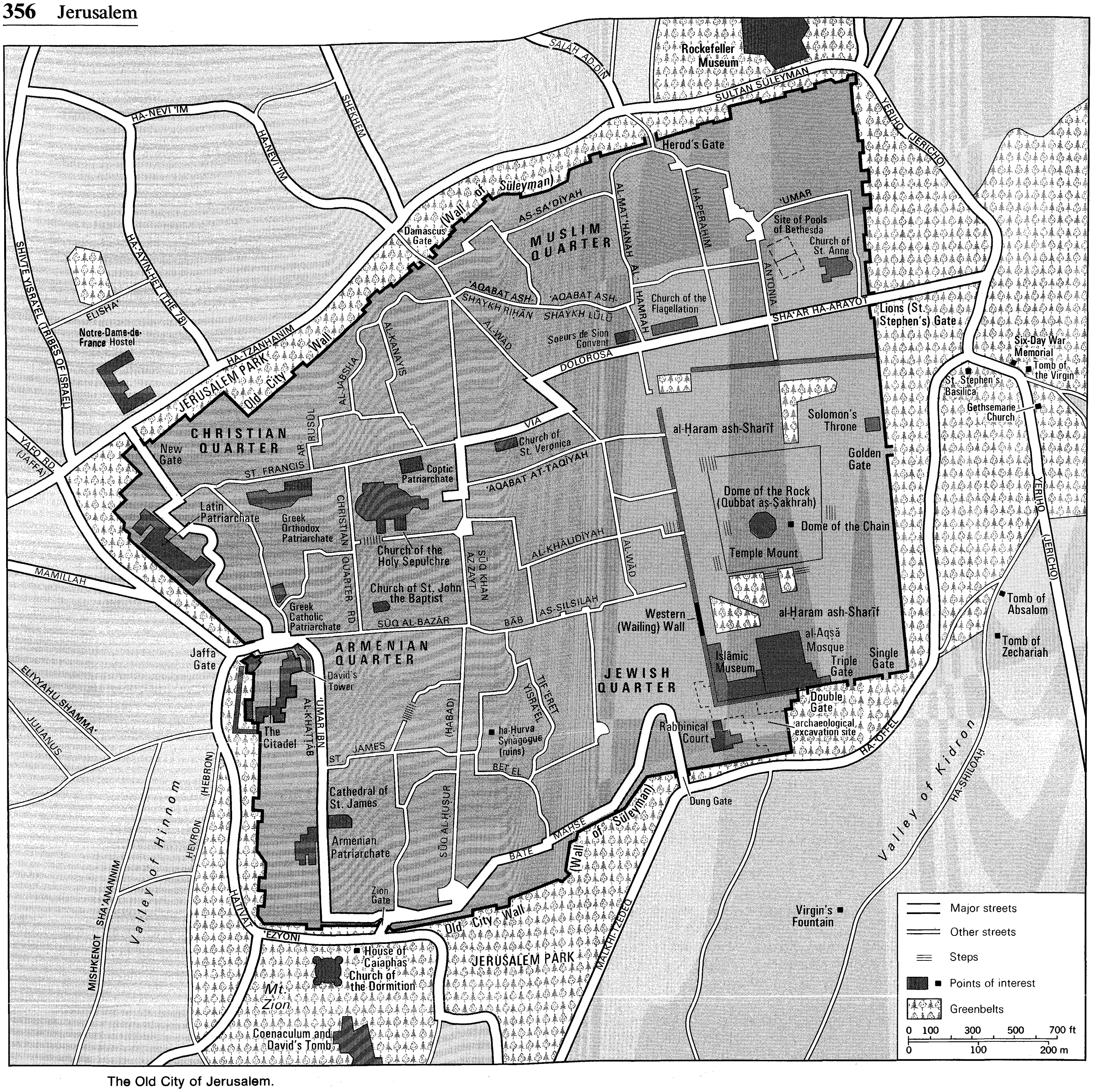
The first neighbourhoods outside the Old City walls, built from the 1860s onward, were scattered chiefly along the main roads leading into the city. The earliest of the Jewish communities were paralleled by non-Jewish expansion prompted by Christian religious or nationalistic motivation and included establishment of the Russian Compound, the German Colony, and the American Colony. Some early communities, such as Mishkenot Sha'anannim and Yemin Moshe, with its famous windmill landmark, have been reconstructed and partially settled or turned into cultural centres. Others include the Bukharan Quarter; Me'a Sheʿarim, founded by Orthodox Jews from eastern and central Europe, with its scores of small synagogues and Talmudic study houses; and Maḥane Yehuda, with its fruit and vegetable market, inhabited mainly by Jews of Oriental origin. Residential quarters established between World Wars I and II include Reḥavya in the centre, Talpiyyot in the south, and Qiryat Moshe and Bet ha-Kerem in the west. The old campus of the Hebrew University at Mt. Scopus, which formed for 20 years (1948–67) an Israeli enclave in the Jordanian-dominated sector, was entirely rebuilt after the Six-Day War. Some Arab districts, such as Talbieh (modern Qomemiyyut) and Katamon (Gonen), abandoned during the fighting of 1947–48, became Jewish neighbourhoods; and thousands of houses were built for new immigrants in districts to the west, newly incorporated into the city. Arab neighbourhoods outside the Old City include the American Colony, ash-Shaykh Jarrāḥ, Wādī al-Jōz, and Bayt Ḥanīnā in the north and villages such as Silwān and Bayt Ṣafāfā in the south. Other important communities include Gillo, Newe Yaʿaqov, and Ramot Allon.

Housing. There is a great variety of housing in the city. In the Old City are antiquated buildings constructed of ancient stones; 19th-century Jewish neighbourhoods, some of which have declined into slums; modern quarters with tree-lined streets; and government-built housing projects, mainly for new immigrants. The most common basic dwelling unit in the Old City consists of a complex of structures, often on different levels, built around an inner court that is entered through a narrow corridor. Steps have been taken to facilitate slum clearance.

Architecture. The outstanding characteristic of the architecture of Jerusalem is the coexistence of old and new, sacred and secular, in a variety of styles. The most conspicuous feature is the Old City Wall, erected 1538–40 by the Ottoman sultan Süleyman the Magnificent, largely on the foundations of earlier walls going back chiefly to the period of the Crusades but in some places dating to Byzantine, Herodian, and even Hasmonean times. On three sides of the Temple Mount, parts of the original supporting walls still stand. During the centuries when Jews were excluded from the Temple Mount, its Western Wall became Jewry's holiest shrine. Since 1967 the wall has been further exposed, and plans have been made to landscape the area once excavations are completed. The main buildings on the platform are the gold-capped Dome of the Rock, completed in 691, and the silver-domed al-Aqṣā Mosque, built in the early 8th century.

The Wall of Süleyman

The Citadel (with David's Tower) beside the Jaffa Gate, which acquired its present form in the 16th century, was created over ruins from the Hasmonean and Herodian periods, integrating large parts of crusader structures and some Mamlūk additions. The large number of churches mainly represent two great periods of Christian architecture, the Byzantine and crusader periods. The predominant characteristic of the former are monumental, two-

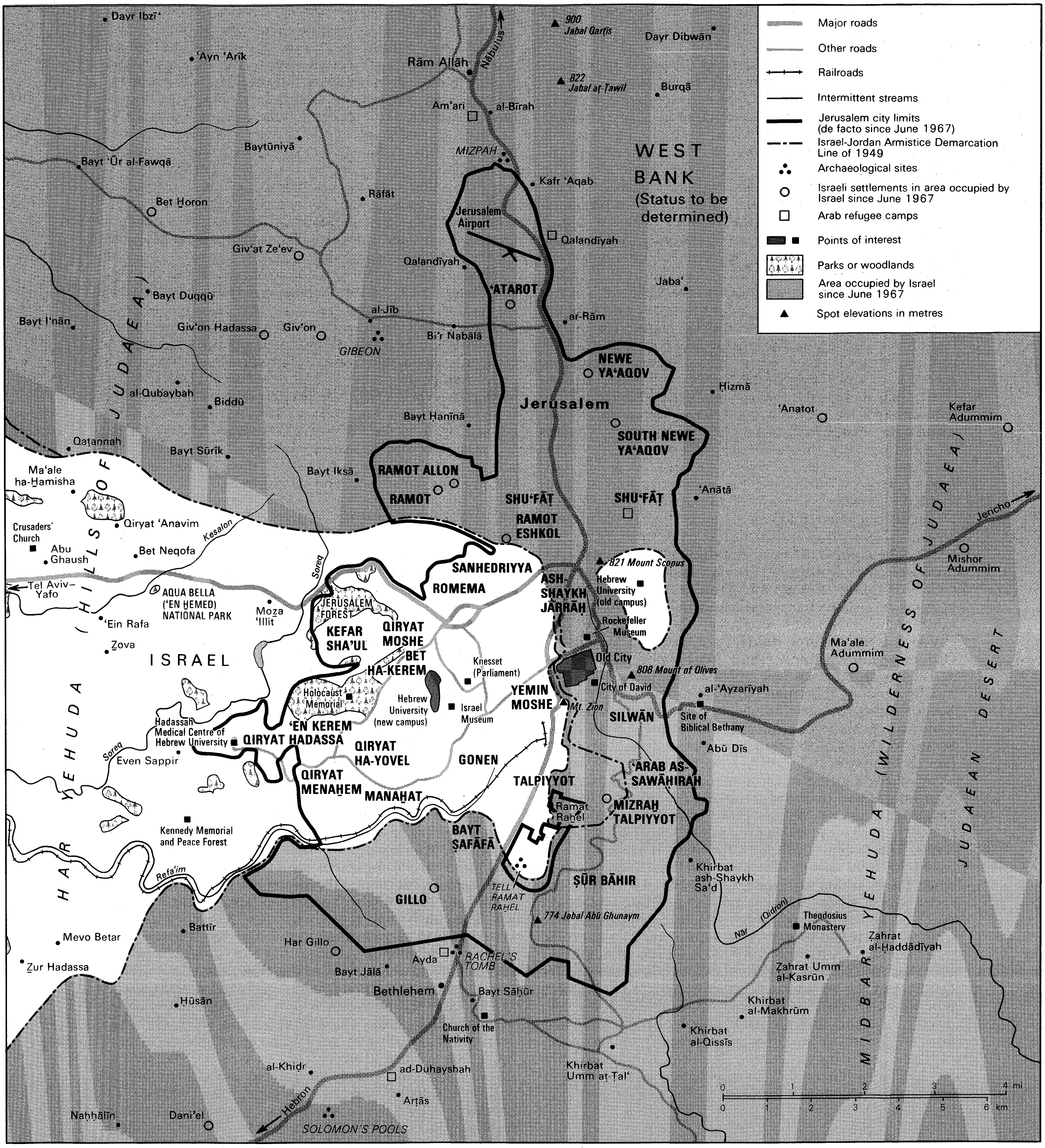


The Old City of Jerusalem.

or three-tiered ornamental or carved basketlike capitals, the latter reflecting Romanesque styling, which features pointed arches and ribbed vaults. The Church of the Holy Sepulchre incorporates elements of both styles, but its facade and layout are architecturally Romanesque. The best example of the mixed style is the Church of St. Anne (its substructure is Byzantine); others are the Armenian Cathedral of St. James, which combines Romanesque with Oriental elements, and the Tomb of the Virgin, which is Romanesque in its upper part but Byzantine in its lower. The central part of the triple bazaar, as well as its link with the Cardo (a restored Roman-Byzantine mall), is of crusader origin. Mamlūk constructions of the 13th to the 15th century, as well as coats of arms of Mamlūk rulers, are found along David Street and near the Gate of the Chain at the Western Wall. The constructions are characterized by "stalactite" or "honeycomb" ornamentation and the use of multicoloured slabs of stone. Ottoman architecture of the beginning of the 16th century continued the Mamlūk style and is represented in some structures of the Temple Mount. The rock-cut tombs east and north of the Old City exemplify architecture of the first half of the 1st millennium BC (Tomb of Pharaoh's Daughter) and the Second Temple period (Tombs of the Kings, Tomb of Absalom, and Tomb of Zechariah). The restored Monastery of the Cross, in the heart of modern Jerusalem, dates from the 5th century.

Modern architecture

As Jerusalem spread outside the walls, the architecture came to be characterized chiefly by iron beams and red-tiled roofs. From 1930 there was a radical change, and flat roofs and reinforced concrete faced with naturally dressed stone predominated. Whereas residential buildings are seldom taller than four stories and office buildings seldom taller than eight, there is a growing tendency, despite opposition, for high-rise construction. This is the case with a number of modern hotels at the western entrance to the city, and the construction of office buildings in the city centre is following the trend. All building, however,



Jerusalem.

must follow a city ordinance requiring construction of stone. Outstanding modern architecture is reflected by the buildings on the university campuses on Mt. Scopus and Givʻat Ram, the Knesset (Parliament), the Israel Museum, the Jerusalem Theatre, and the Hebrew Union College. More modern trends are represented by the Bank of Israel, the Jerusalem Great Synagogue, and the President's Home. An earlier generation is represented by the Government House (UN Headquarters), the King David Hotel, the Rockefeller (Archaeological) Museum, and the Young Men's Christian Association.

THE PEOPLE

Composition of the population

Because Jerusalem is a holy city, uniquely revered by three major religions, its people can, perhaps, be best described according to religious affiliation. Most of the city's residents are Jews. The Muslims are the most homogeneous of the communities, the Christians the most diversified.

Responsibility for the city's holy places and religious communities is vested in the Ministry of Religious Affairs, which has special desks for the individual denominations. The administration, protection, and care of holy places are in the hands of the respective religious authorities. Penalties of up to seven years' imprisonment may be inflicted for desecration of these places. The rites and observances of all the faiths are publicly celebrated.

Jews. Among the Jews, the main divisions are between Ashkenazim and Sephardim, names denoting, not very accurately, places of origin. Of more importance is the division between the orthodox and the more secular-minded segments of the population, whose attitudes over religious as well as political matters often conflict. Jerusalem is the centre of Jewish religious reverence and aspiration. The most sacred spot is the Temple Mount, on which Orthodox Jews refrain from setting foot for fear of profaning its sanctity. In addition to the Western Wall—the most important centre of prayer and pilgrimage—other holy places are Mt. Zion, with the reputed tomb of King David, the Mount of Olives, with its ancient Jewish cemetery, and the tombs of priestly families in the Valley of Kidron. Ancient synagogues and study houses in the Old City are being restored; particularly worthy of mention is the interconnected group of four synagogues begun in the 16th century by Jews exiled from Spain. There are scores of Jewish houses of prayer in the New City. Notable modern institutions include the synagogue at Hekhal Shelomo, the seat of the Chief Rabbinate of Israel. Jerusalem is the world's foremost centre of rabbinic learning and contains scores of yeshivas, the Talmudic academies.

Muslims. Jerusalem is revered by Muslims as the third holiest place on earth, and the pilgrimage to Jerusalem (*taqdīs*) completes the main pilgrimage (*ḥajj*). A Council of Waqf (religious endowment) and Muslim Affairs was created in 1967, with jurisdiction over Sharīʿah courts and waqfs. The council assumed the responsibility for the administration of Muslim affairs that had previously rested with the Council of Waqf and Muslim Affairs in Amman.

Christians. Christians constitute the smallest but the most highly diversified section of the population. The city is the seat of three resident patriarchs of the Eastern Orthodox churches and many archbishops and bishops, and it has an ecclesiastical embassy for almost every sect in Christendom. The main groups are Eastern Orthodox, Monophysite, Roman Catholic, and Protestant. Major denominations share control over the Church of the Holy Sepulchre. Most of the church bodies in Jerusalem maintain scholarly research institutes with fine libraries.

The Greek Orthodox Church maintains a patriarchate with jurisdiction over the entire Holy Land. The Russian Orthodox churches (one governed from Moscow and one in exile) have considerable properties dating back to tsarist times. The Roman Catholic Church in Jerusalem, established in 1099 during the First Crusade, was dissolved when the Muslims won the city in 1244. The Franciscan order, which since 1334 has been the "Custodian of the Holy Land," is charged with the safekeeping of Roman Catholic rights and properties in Jerusalem. The Latin Patriarchate was reestablished in 1847. Of the Monophysite churches, the Armenian is the largest (others include the Coptic and the Abyssinian), its patriarchate having been established in the 6th century. Most of the Armenians dwell in a compound around the seat of the patriarchate at the Cathedral of St. James, which constitutes the largest monastic centre in the country. The Protestant community is small but influential. The jurisdiction of the Anglican archbishop extends over the entire Middle East.

Armenian Church

THE ECONOMY

The main source of livelihood in Jerusalem is government and public service employment (including the academic and clerical professions). Since 1967, business activity and investment in the city have been stimulated by the housing boom and the ever-increasing influx of pilgrims and tourists. Personal income has risen steadily, and unemployment is marginal, although the city still deals with a large number of social welfare cases.

Industry and trade. The establishment of heavy industries has not been encouraged, in the interest of preserving the traditional character of the city. Combined with transport and marketing difficulties, this has limited the city to a number of small industries. They include diamond cutting and polishing and the manufacture of home appliances, furniture, shoes, pencils, plastics, textiles, clothing, and pharmaceuticals and chemicals. There are also printing and publishing houses, as well as workshops producing jewelry, giftware, religious articles, curios, and printed fabrics. More recent additions include the modern science-based industries and the development of industrial quarters in the outskirts of the city and in some of its easily accessible satellite settlements. Nevertheless, the percentage of the work force engaged in industry remains quite small, whereas about two-thirds is engaged in services. The tourist boom has stimulated the construction of first-rate hotels in the city, which receives the highest number of tourists in the country. The heaviest influx is linked with the Jewish high holidays, Christmas, Passover, Easter, and the Muslim pilgrimage.

Transportation. Despite a considerable increase in roads and streets since the mid-1970s, the traffic problem remains one of the most acute in city planning because of the ever-increasing number of private vehicles. Public transportation is provided by bus companies, which operate in eastern and western Jerusalem and also make interurban connections. The latter are also offered by taxi services (*sherut*), which connect Jerusalem with all major Israeli towns, including Elat and Tiberias. The old railway connecting Jerusalem with Tel Aviv–Yafo and Haifa on the coast and with Beersheeba inland is of secondary importance; the Jerusalem Airport at the northern edge of the city serves mainly inland tourist traffic.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Jerusalem is the seat of the president and the Knesset (Parliament) of Israel. In 1947 the United Nations recommended that it be made an international city, but the proposal was opposed by both Israel and Jordan. Although a large number of countries do not recognize Jerusalem as the capital of Israel, a large proportion of the resident foreign embassies and legations in Israel were located in Jerusalem until the 1980 law officially proclaimed the unified city the nation's capital. Most of the foreign delegations then moved to Tel Aviv–Yafo to reaffirm nonrecognition of Jerusalem as the Israeli capital. France and the United States each maintain consulates in the eastern and western parts of the city. Diplomats living in the Tel Aviv area go to Jerusalem to present their credentials to the president and transact business at the Foreign Ministry. The ministries are concentrated in the Qiryat Ben-Gurion, the government complex, which is flanked by the Knesset on one side and the Bank of Israel on the other. The Ministry of Defense is still located in Tel Aviv–Yafo, and several ministries are in temporary housing in Jerusalem. In addition to the Supreme Court and the Chief Rabbinate, the city also houses the head offices of many world Jewish bodies, such as the Jewish Agency and the World Zionist Organization, as well as the Martyrs' and Heroes' Remembrance Authority (Yad va-Shem), which commemorates the victims of the Holocaust.

The Municipal Council is composed of 31 members who are elected every four years by proportional representation. The council is headed by the mayor, who, since 1975, has been elected by direct vote. Permanent residents, even if not Israeli nationals, are entitled to vote. Of the administrative staff, more than 20 percent are Arabs. Official correspondence is issued in both Hebrew and Arabic.

Voting

Services. Jerusalem has always depended on human ingenuity to bring its water from afar. The underground aqueduct built by King Hezekiah in the 8th century BC is still extant, and many reservoirs and rainwater cisterns date from ancient times. Since the 1950s the New City has enjoyed an unlimited supply from the Israeli national water grid; East Jerusalem was reconnected to the West Jerusalem system in 1967.

The city has a modern sewerage system. The six miles of ancient piping that run through the Old City still present serious engineering problems. Drainage repairs in the Christian Quarter have uncovered Byzantine pavements, which have now been restored. Additionally, parts of the Via Dolorosa, said to follow the path along which Jesus carried the cross to Golgotha, have been repaved to facilitate the Christian pilgrimage.

Electricity is supplied by the national grid of an Israeli government corporation, as well as by a small diesel plant in East Jerusalem.

Health. The Hadassah Medical Centre at ʿEn Kerem is one of the most advanced institutions of its kind in the world. It treats patients from throughout the country, as well as from the West Bank, the Gaza Strip, and Jordan. Other hospitals are the Hadassah Hospital on Mt. Scopus; Shaʿare Tzedeq, which pays special attention to the requirements of Orthodox Jews; Biqur Ḥolim; St. John's Ophthalmic Hospital; Ezrat Nashim for mental patients; Alyn for crippled children; an Arab-Muslim hospital, al-Maqāṣid al-Khayrīyah, at et-Tur; and an Arab-Christian hospital, al-Muṭallaʿ (Augusta Victoria), on the Mount of Olives, which is run by Lutheran organizations that mainly care for the Arab population. A modern medical centre that also serves the Arab population was opened in 1982 at ash-Shaykh Jarrāḥ in northeast Jerusalem. Also important are the Austrian Hospice inside the old town, the French Hospital, St. Louis (accepting terminal cases), and the Sisters of Charity (for the crippled and handicapped). After unification of East and West Jerusalem the Kupat Ḥolim, the medical insurance arm of the General Federation of Labour, established several clinics in the eastern part of the city. Supplementing the regular medical facilities are the Magen David Adom and the Red Crescent (counterparts of the Red Cross), which provide additional emergency services.

Most families belong to one of the medical insurance funds run by the labour federations and other nongovernmental bodies. Medical insurance is by law obligatory for all citizens. The municipal social welfare department takes care of social cases that are not covered by medical insurance. Municipal clinics have been established for mothers and children. Health supervision, including dental inspection and treatment, is provided in all of the city's schools. All health services are subsidized by the government. Rapidly developing and expanding social services provide adult education, senior citizen clubs, and youth clubs among a variety of programs in both parts of the city. Community centres are the focal points of educational and recreation programs in the neighbourhoods.

Religious and secular education

Education. Because of the high birth rate and the strong religious convictions of many among the population, education has always involved complex problems. The language of instruction is Hebrew in Jewish schools and Arabic in Arab schools. Hebrew and Arabic are alternately first or second languages in all schools. In structure and curriculum, as a rule, the Arab schools follow the Jordanian system. While the majority of school-age children attend government schools, there are numerous private institutions maintained by Jewish, Muslim, and Christian religious organizations. In the latter, the language of instruction is sometimes French or English. State kindergartens were introduced in East Jerusalem in 1967. Education is the single most important item in the city's budget, and the municipality is responsible for the maintenance of classrooms from kindergarten through high school.

The Hebrew University of Jerusalem (opened 1925) is the country's foremost institution of higher learning. It has two main campuses—at Mt. Scopus in the east and at Givʿat Ram in the west, in addition to the medical school at ʿEn Kerem and the Faculty of Agriculture in Reḥovot. The old buildings on Mt. Scopus have been renovated and supplemented by a new complex of buildings. Other institutes of higher learning are the Bezalel Academy of Arts and Design, the Rubin Academy of Music, the Hebrew Union College, several teachers'-training colleges, and an Armenian seminary.

Libraries

The Jewish National and University Library, with about 2,500,000 volumes is the largest in the country. It has the foremost collection of books, incunabula, and periodicals of Judaica in the world, in addition to an excellent library on archaeology and Oriental studies, including the history of Palestine. In addition there are the Library of the Knesset and the State Archives (each of which receives a copy of every book printed in Israel) and the Municipal Library and its branches. Numerous other libraries serve a variety of needs.

CULTURAL LIFE

An important cultural institution is the Israel Museum, which, in addition to its general art collection, houses a comprehensive Middle Eastern archaeological collection, several important Dead Sea Scrolls and other relics, a notable collection of Jewish ritual art, Middle Eastern ethnological exhibits, a sculpture garden, and a youth wing. The Rockefeller Museum concentrates on the archaeology of the Holy Land, and there is an Islāmic Museum near the al-Aqṣā Mosque, as well as the L.A. Mayer Memorial Institute for Islāmic Art in West Jerusalem.

Special mention should be made of the École Biblique et Archéologique Française, the Studium Biblicum Franciscanum, the Pontifical Biblical Institute, the British School of Archaeology, the William Foxwell Albright Institute of Archaeological Research, the Swedish Theological Institute, and the Ben Zvi Institute. All of these have libraries dealing with theology and the ancient and modern history of Israel and the Middle East; some have collections of antiquities and valuable manuscripts.

Art exhibitions are held in the Israel Museum, the Artists' House, Hutzot ha-Yotzer (the craftsmen's centre), the International Cultural Centre for Youth, and private galleries. Concerts and theatre performances are given at Binyane ha-ʾUma (the Convention Centre), the Khan (housed in a restored 18th-century building), and the Wise Auditorium of the Hebrew University. The beginnings of an Arab theatre have been established.

The only English-language daily in Israel, the *Jerusalem Post,* is printed in Jerusalem, as are all of the Arabic-language dailies. Most Hebrew scholarly periodicals are also printed in the city. The Government Press Office is located at Beit Agron, the headquarters of the Jerusalem Journalists' Association. The headquarters of the Israel Broadcasting Authority (television and radio) are also in Jerusalem. Radio broadcasts are mainly in Hebrew and Arabic, though some programs are also broadcast in a number of languages, including English, French, Ladino, Romanian, Russian, and Yiddish.

The Jerusalem Foundation, established in 1966, recruits funds for the preservation of the city's multi-religious heritage and the embellishment of its barren areas. This foundation is responsible for Jerusalem's many parks, gardens, woodlands, and forests. The largest is Jerusalem Park, designed as a greenbelt to encircle the Old City walls. There are also small gardens, playgrounds, and recreational areas dotting the city. The Biblical Zoo houses specimens of all the animals that are mentioned in the Bible, and the Natural History Museum focuses on the country's fauna.

The municipality, the Young Men's Christian Association (YMCA), and local clubs run comprehensive sports programs. The YMCA soccer field can accommodate 10,000 spectators, and the ha-Poʿel (Workers' Sports Club) field 7,500. There are a number of open-air swimming pools. Community centres in the suburbs also provide sports facilities.

Archaeological excavations

Since 1968 extensive excavations have been carried out in the Old City on behalf of the Hebrew University Institute of Archaeology, the Israel Department of Antiquities and Museums, and the Israel Exploration Society. The digs around the southern and western walls of the Temple Mount, going down to the Herodian pavements, have revealed the steps leading to the Temple, the priests' underground entrance to the Temple, and many religious objects. There are also impressive remains of public buildings alongside a main street. Remains found within the precincts of the First Wall in the Jewish Quarter bore the imprint of burning and destruction during the sack of the city by the Romans in AD 70. For the first time were found walls of structures dating to the 8th and 7th centuries BC. One of these has been identified as the "Broad Wall" described by Nehemiah. A crucified body, from Roman times, with a nail still lodged in the ankle, was discovered in a Jewish tomb at Givʿat ha-Mivtar. Extensive excavations in the Citadel uncovered structures of the Hasmonean, Herodian, crusader, and Mamlūk periods.

Below the Temple Mount outside the walls, impressive remains of an Umayyad palace have been found. The excavations since 1978 in the Mt. Ophel and City of David area have revealed Canaanite and early Hebrew settlements, the latter with a wealth of seals, bulls, epigraphical material, and everyday utensils. A most significant discovery was that of the Roman and Byzantine Cardo, running from the vicinity of the Zion Gate through the restored Jewish Quarter to its crusader part and crossing with the Old City bazaars. The street was reconstructed using the ancient pavement, columns, and capitals. The discovery of the Crusaders' Church, hospice, and hospital of the crusader Teutonic Order (12th century AD) in the Jewish quarter and the huge expanse of wall and towers (from the crusader and Ayyūbid periods of the 12th and 13th centuries AD) between the Dung Gate and the Zion Gate is a major contribution to the history of the city. (Jo.Pr./Ed.)

History

THE EARLY PERIOD

Ancient origins. The earliest traces of human settlement in the city area, found on a hill to the southeast, are from the late Chalcolithic Period and Early Bronze Age (*c.* 3000 BC). Excavations have shown that a settlement existed on the site south of the Temple Mount, and a massive town wall was found just above the Gihon Spring, which determined the location of the ancient settlement. The name, known in its earliest form as Urusalim, is probably of western Semitic origin and apparently means "Foundation of Shalem" ("Foundation of God"). The city and its earliest rulers, the Egyptians, are mentioned in the Egyptian Execration Texts (*c.* 1900–1800 BC) and again in the 14th-century Tell el-Amarna correspondence, which contains a message from the city's ruler, Abdi-Kheba (Abdu-Ḥeba), requiring his sovereign's help against the invading Hapiru (Habiru, ʿApiru). A biblical narrative mentions the meeting of Canaanite Melchizedek, said to be king of Salem (Jerusalem), with the Hebrew patriarch Abraham, and in a later episode it mentions another king, Adonizedek, who headed an Amorite coalition and was vanquished by Joshua.

Kings of Judah

About the year 1000 BC Jerusalem, on the frontier of Benjamin and Judah, inhabited by a mixed population described as Jebusites, was captured by David, founder of the joint kingdom of Israel and Judah, and the city became the Jewish kingdom's capital. His successor, King Solomon, extended the city and built his Temple on the threshing floor of Araunah (Ornan) the Jebusite. Thus Jerusalem became the place of the royal palace and the sacred site of a monotheistic religion.

On Solomon's death the northern tribes seceded. In 922 BC the Egyptian pharaoh Sheshonk I sacked the city, to be followed by the Philistines and Arabians in 850 and Joash of Israel in 786. After Hezekiah became king of Judah, he built new fortifications and an underground tunnel, which brought water from Gihon Spring to the Pool of Siloam inside the city, but he succumbed to the might of Sennacherib of Assyria, who in 701 forced payment of a heavy tribute. In 612 Assyria yielded its primacy to Babylon. Eight years later Jerusalem was despoiled, and its king was deported to Babylon. In 586 BC the city and Temple were completely destroyed by Nebuchadrezzar, and the captivity began. It ended in 538 BC when Cyrus II the Great of Persia, who had overcome Babylon, permitted the Jews, led by Zerubbabel, of the Davidic house, to return to Jerusalem. The Temple was restored (515 BC) despite Samaritan opposition, and the city became the centre of the new statehood and its position strengthened when Nehemiah (*c.* 444) restored its fortifications.

Hellenistic and Hasmonean periods. With the coming of Alexander the Great and his victory at Issus in 333 BC, Jerusalem was drawn for the first time into the orbit of Western power politics.

The growth of Greek influence

After Alexander's death, Palestine fell to the share of his marshal, Ptolemy I Soter, son of Lagus, who had occupied Egypt and had made Alexandria his capital. In the year 198 BC Jerusalem was acquired by the northern dynasty, descended from Seleucus I Nicator, another of Alexander's marshals, which ruled from Antioch (contemporary Antakya, Tur.). The growth of Greek, or pagan, influence affronted the orthodox, whose hostility burst into armed rebellion in 167 BC after the Seleucid Antiochus IV Epiphanes had deliberately desecrated the Temple. The revolt was led by a pious countryman called Mattathias, son of Hasmoneus (Hasmon), and was carried on by his son Judas, known as the Maccabee. The Hasmoneans succeeded in expelling the Seleucids, and Jerusalem regained its position as the capital of an independent state ruled by the priestly Hasmonean dynasty.

The kingdom of Herod

Roman rule. Rome had for some time been expanding its authority in Asia, and in 63 BC Pompey captured Jerusalem. A clash with Jewish nationalism was averted for some time by the political skill of a remarkable family, whose most illustrious member was Herod the Great. Herod was of Edomite descent, though of Jewish faith, and was allied through his mother with the nobility of Nabataean Petra, the rich Arab state that lay to the east of the Jordan. In 40 BC Herod, who had distinguished himself as governor of Galilee, was appointed "client" king of Judaea by the Roman Senate. He was the friend of Mark Antony and, after the defeat of Antony by Octavian at Actium in 31 BC, of Octavian himself.

Herod was king for the next 36 years, during which period Jerusalem reached its peak of greatness, growing in richness and expanding even beyond the new double line of walls. The Temple Mount esplanade was artificially enlarged with supporting walls (including the Western Wall) to house Herod's greatest creation, the splendid new Temple, which took more than a generation to build. The new royal palace was strengthened by immense towers that were integrated in the older Hasmonean walls, whereas the Temple was defended by a new citadel. An amphitheatre added to the Hellenistic character of the city. Centre of religion, goal of obligatory pilgrimage, and the seat of the ruler and of the autonomous court of the Sanhedrin (Jewish Council of Elders), Jerusalem became a great metropolis of the Hellenistic world. Herod died in 4 BC and was succeeded by his son Archelaus, who was subsequently deposed by the Romans in AD 6 and replaced by the first of a series of Roman procurators. It was under the fifth procurator, Pontius Pilate, that Jesus of Nazareth was put to death.

From AD 41 to 44 the kingdom of Herod was reconstituted for his grandson Herod Agrippa I, upon whose premature death the procurators returned. In 66 the Jews rebelled against Rome, and in 70 the city was besieged and almost wholly destroyed by the Roman forces under Titus. The Temple, Herod's greatest creation, was reduced to ashes. By 130 the city had been partially repopulated, and the Jews again revolted unsuccessfully against Rome from 132 to 135. Hadrian decided to plant a Roman city, Aelia Capitolina, on the site. The general layout of Hadrian's town has lasted into the 20th century.

Christian pilgrims early found their way to Jerusalem. It was, however, the conversion to Christianity of Constantine the Great and the famous pilgrimage (326) of his mother, Empress Helena, who found "the True Cross," that made possible the building of the famous shrines in Jerusalem, including the Church of the Holy Sepulchre, and inaugurated one of the city's most splendid and prosperous epochs. The Christian glorification was carried on into the 6th century when, under the emperor Justinian, the Church of Resurrection was rebuilt and many other churches, as well as monasteries and hospices, were established. In 614 this golden age was brought to an end by the Persian invasion, in which the inhabitants of Jerusalem were massacred and the churches destroyed.

The Islāmic and crusader periods. In 638 the Muslim caliph ʿUmar I entered Jerusalem, and in 688–691 the 10th caliph, ʿAbd al-Malik ibn Marwān, built the Dome of the Rock. The city, however, lost some of its earlier importance, despite being proclaimed a goal of Muslim pilgrimage, when the caliphate was moved from Damascus to Baghdad. It shrank in size, and the new line of walls (11th century) did not include the City of David and Mt. Zion. Both the Umayyads and their successors, the ʿAbbāsids, pursued a liberal policy toward Christians and Jews. In 969 control of Jerusalem passed to the Shīʿite Fāṭimid caliphs of Egypt, and in 1010 the caliph al-Ḥākim ordered the destruction of Christian shrines. In

1071 the Seljuq Turks defeated the Byzantines, displaced the Egyptians as masters of the Holy Land, and cut the pilgrim routes, thus stimulating the Crusades.

The city was recaptured by the Egyptians (1098) a year before the hosts of the First Crusade besieged the city. The crusader state took its name from the city, as the Kingdom of Jerusalem. The city regained its position as capital of the kingdom, which (with its northern principalities) stretched from the confines of modern Turkey to the Red Sea. The great Muslim sanctuaries became Christian churches, and in 1149 the Church of the Holy Sepulchre as it exists today was consecrated. Muslims and Jews were barred from living in the city. The kingdom in Jerusalem lasted from 1099 to 1187, when it was overthrown by Saladin, whose Ayūbbid successors ruled from Damascus and Cairo. Jerusalem was again in Christian hands from 1229 to 1239 and from 1240 to 1244, when it was sacked by the Khwārezmian Turks. In 1247 the Holy City fell once more to Egypt and remained subject to the Mamlūks. The great sanctuaries became Muslim again, and the only Christians who remained were the Greek Orthodox and other Oriental groups. In the 14th century the Franciscans began to represent the Roman Catholic interests. The Jews, who had been barred by the crusaders, returned and from the mid-13th century inhabited their own quarter. The layout of the quarters was fixed in that period. The Mamlūks dotted the Temple Mount and the city with mosques, madrasahs, and ornamental tombs.

In 1517 the Ottoman sultan Selim I took the city and inaugurated a Turkish regime that lasted 400 years. The 16th century was a period of great urban development. In addition to the new walls, which still encompass the Old City, and the repaired water supply, new madrasahs, waqfs, and charity institutions multiplied. But by the end of the century the city began to decline, a process that lasted for the next 300 years.

MODERN JERUSALEM

Several factors determined the fate of the city in the 19th century. In 1831 Ibrāhīm Pasha, son of the Egyptian ruler Muḥammad ʿAlī, captured Jerusalem and introduced a series of far-reaching reforms, which were retained when the Turks regained the city (1840). A municipality was established in 1887, and by the middle of the century all of the great European powers had established consulates in the city, which had a salutary influence on the position of the non-Muslim population. Finally, Jewish immigration, mainly from eastern Europe, changed the city's demographic structure and the relative importance of the Old City and the new quarters outside the walls. Christian and Muslim quarters followed suit. By the mid-19th century half of the city's population was Jewish, and it was expanding beyond the walls.

In 1917 British troops under Sir Edmund Allenby entered Jerusalem after the retreat of the Turks. This opened a new era lasting until 1948, during which Jerusalem again became a capital, now ruled under the British Mandate and headed by a high commissioner. About half of its population of some 80,000 was Jewish, with the rest divided between Muslims and Christians. The city developed quickly, expanding its economy and population despite bloody confrontations between Arabs and Jews in 1920, 1929, and 1936 and with skirmishes continuing to the eve of World War II. In 1947 the hostilities renewed; the British pulled out in 1948, and in the ensuing fighting Transjordan captured the Old City, and Jerusalem was divided between Transjordan and the Israelis, the latter proclaiming it the capital of the State of Israel. During the Six-Day War of June 1967 the Israelis stormed the Old City and claimed Jerusalem to be unified once again. Despite a UN resolution that disapproved of the action, the city continued to develop as a unified entity under Israeli administration. The declaration by the Knesset in 1980 that officially made unified Jerusalem the capital of Israel stirred considerable international controversy, and recognition of the capital was withheld by numerous countries. In the following years the city's status remained a point of international contention as restoration work and new construction progressed throughout the old and new sections.

BIBLIOGRAPHY. A good general account of the city is given by TEDDY KOLLEK and MOSHE PEARLMAN in *Jerusalem, Sacred City of Mankind* (1968; U.S. title, *Jerusalem: A History of Forty Centuries*). The best guides to Christian monuments and holy places are EUGENE HOADE, *Guide to the Holy Land,* 11th ed. (1981), and *Jerusalem and Its Environs,* 5th ed. (1964). Noteworthy archaeological works include KATHLEEN M. KENYON, *Jerusalem: Excavating 3000 Years of History* (1967), and *Digging Up Jerusalem* (1974); and the *Encyclopedia of Archaeological Excavations in the Holy Land,* rev. trans. from the Hebrew, vol. 2 (1976), Eng. ed. by MICHAEL AVI-YONAH. For recent excavations and discoveries, see NAHMAN AVIGAD, *Discovering Jerusalem* (1983; originally published in Hebrew, 1980); YIGAEL YADIN (ed.), *Jerusalem Revealed: Archaeology in the Holy City, 1968–1974,* trans. by R. GRAFMAN (1975, reissued 1976); BENJAMIN MAZAR, *The Mountain of the Lord,* trans. from the Hebrew (1975); and MEIR BEN-DOV, *In the Shadow of the Temple* (1985).

For ancient history the Bible is the basic source, best consulted in the modern edition known as *The Jerusalem Bible* (1966), prepared in Jerusalem itself by scholars long conversant with the Holy City and its monuments. It may be supplemented by a good commentary such as JAMES HASTINGS, *Dictionary of the Bible,* rev. ed. by FREDERICK C. GRANT and H.H. ROWLEY (1963). Next to the Bible, the main original source for ancient Jerusalem is FLAVIUS JOSEPHUS, the Jewish historian who wrote under Roman patronage at the end of the 1st century AD; the *Loeb Classical Library* edition, *Josephus,* 9 vol. (1926–65, reprinted 1966–69), is recommended. Of modern works, see EMIL SCHURER, *The History of the Jewish People in the Age of Jesus Christ (175 B.C.–A.D. 135),* 2 vol. rev. ed. by GEZA VERMES, FERGUS MILLAR, and MATTHEW BLACK (1973–79; originally published in German, 2nd ed., 1886–90). GEORGE ADAM SMITH, *Jerusalem: The Topography, Economics and History from the Earliest Times to A.D. 70,* 2 vol. (1907–08, reprinted 1974 in 1 vol.), is a comprehensive survey by a great scholar. The works of the great Dominican scholars HUGHES VINCENT and F.M. ABEL, especially *Jérusalem: recherches de topographie, d'archéologie et d'histoire,* 2 vol. in 4 (1912–26), should also be read. For medieval history, STEVEN RUNCIMAN, *A History of the Crusades,* 3 vol. (1951–54, reprinted 1975), provides essential background, and each volume contains an exhaustive bibliography. See also JOSHUA PRAWER, *The Crusaders' Kingdom: European Colonialism in the Middle Ages* (1972; U.K. title, *The Latin Kingdom of Jerusalem,* 1973), and *Crusader Institutions* (1980). GUY LE STRANGE (trans.), *Palestine Under the Moslems: A Description of Syria and the Holy Land from A.D. 650 to 1500* (1890, reprinted 1975), is an exhaustive collection of medieval Arabic sources. T.S.R. BOASE, *Castles and Churches of the Crusading Kingdom* (1967), is a finely illustrated work by an outstanding scholar. AMNON COHEN and BERNARD LEWIS, *Population and Revenue in Towns of Palestine in the Sixteenth Century* (1978); and YEHOSHUA BEN-ARIEH, *Jerusalem in the 19th Century: The Old City* (1984; originally published in Hebrew, 1977), treat the Ottoman period. The beginning of modern scholarship was heralded by three books: EDWARD ROBINSON and ELI SMITH, *Biblical Researches in Palestine, Mount Sinai and Arabia Petraea,* 3 vol. (1841, reprinted 1977), by the founder of scientific biblical geography; JAMES FINN, *Stirring Times; or, Records from Jerusalem Consular Chronicles of 1853 to 1856,* compiled by his widow (1878); and CHARLES W. WILSON *et al., The Recovery of Jerusalem: A Narrative of Exploration and Discovery in the City and the Holy Land* (1871, reprinted 1872), the record of the first underground survey of the ancient city. Since then, the output of books on every aspect of Jerusalem has been unceasing. Current publications in all aspects and languages can be followed in the quarterly *Kiryat Sefer.* For the period of the British mandate, ALBERT M. HYAMSON, *Palestine Under the Mandate, 1920–1948* (1950, reprinted 1976), is the standard work. STEWART PEROWNE, *The One Remains* (1954, reissued 1955), is an eyewitness account of the succeeding period. Of particular importance to understanding modern history are MERON BENVENISTI, *Jerusalem, the Torn City* (1976; originally published in Hebrew, 1973); DAVID H.K. AMIRAN, ARIE SHACHAR, and ISRAEL KIMHI (eds.), *Atlas to Jerusalem* (1973), and *Urban Geography of Jerusalem* (1973), a companion volume; and JOEL L. KRAEMER (ed.), *Jerusalem: Problems and Prospects* (1980). Two general histories are JOHN GRAY, *A History of Jerusalem* (1969), an authoritative survey by a biblical scholar; and A.L. TIBAWI, *Jerusalem: Its Place in Islam and Arab History* (1969). The latest archaeological excavations are reported in the *Israel Exploration Journal* (quarterly). Studies of current problems are published by the Jerusalem Institute for Israel Studies, including ORA AHIMEIR (ed.), *Jerusalem: Aspects of Law,* 2nd rev. ed. (1983); and DAVID KROYANKER, *Jerusalem Planning and Development, 1979–1982,* trans. from the Hebrew (1982).

(S.H.P./Jo.Pr.)

Jesus: The Christ and Christology

Jesus of Nazareth, the founder of Christianity, a religion that claims more than a third of the world's population in the 20th century, was born in Judaea about 6 BC and died by crucifixion about AD 30. Because of the theological motifs and presuppositions in the faith of the early church in respect to Jesus, it is difficult to write with certainty an authentic life of Jesus.

This article is divided into the following sections:

The gospel tradition

SOURCES

The history of the life, work, and death of Jesus of Nazareth reveals nothing of the worldwide movement to which he gave rise. He lived and taught in a remote area on the periphery of the Roman Empire. His life was of short duration, and knowledge of it remained hidden from most of his contemporary world. None of the sources of his life and work can be traced to Jesus himself; he did not leave a single known written word. Also, there are no contemporary accounts written of his life and death. What can be established about the historical Jesus depends almost without exception on Christian traditions, especially on the material used in the composition of the Gospels of Mark, Matthew, and Luke, which reflect the outlook of the later church and its faith in Jesus.

References in Roman historical and administrative accounts

Non-Christian sources. Non-Christian sources are meagre and contribute nothing to the history of Jesus that is not already known from the Christian tradition. The mention of Jesus' execution in the *Annals* of the Roman historian Tacitus (XV, 44), written about AD 110, is, nevertheless, worthy of note. In his account of the persecution of Christians under the emperor Nero, which was occasioned by the burning of Rome (AD 64), the Emperor, in order to rid himself of suspicion, blamed the fire on the so-called Christians, who were already hated among the people. Tacitus writes in explanation: "The name is derived from Christ, whom the procurator Pontius Pilate had executed in the reign of Tiberius." The "temporarily suppressed pernicious superstition" to which Jesus had given rise in Judaea soon afterward had spread as far as Rome. Tacitus does not speak of Jesus but, rather, of Christ (originally the religious title "Messiah," but used very early among Christians outside Palestine as a proper name for Jesus). The passage only affords proof of the ignominious end (crucifixion) of Jesus as the founder of a religious movement and illustrates the common opinion of that movement in Rome. An enquiry of the governor of Asia Minor, Pliny the Younger, in his letter to the emperor Trajan (*c.* AD 111) about how he should act in regard to the Christians (*Epistle* 10, 96ff.) comes from the same period. Christians are again described as adherents of a crude superstition, who sang hymns to Christ "as to a god." Nothing is said of his earthly life, and the factual information in the letter undoubtedly stems from Christians.

Another Roman historian, Suetonius, remarked in his life of the emperor Claudius (*Vita Claudii* 25:4; after AD 100): "He [Claudius] expelled the Jews, who had on the instigation of Chrestus continually been causing disturbances, from Rome." This may refer to turmoils occasioned among the Jews of Rome by the intrusion of Christianity into their midst. But the information must have reached the author in a completely garbled form or was understood by him quite wrongly to mean that this "Chrestus" had at that time appeared in Rome as a Jewish agitator. Claudius' edict of expulsion (AD 49) is also mentioned in Acts 18:2.

References in Jewish sources

Josephus, the Jewish historian at the court of Domitian who has depicted the history of his people and the events of the Jewish–Roman war (66–70), only incidentally remarks about the stoning in AD 62 of "James, the brother of Jesus, who was called Christ . . ." (*Antiquities* XX, 200). He understandably uses the proper name "Jesus" first (for as a Jew he knows that "Christ" is a translation of "Messiah"), but he adds, though qualified by a derogatory "so-called," the second name that was familiar in Rome. (Some scholars have suggested, however, that this reference was a later Christian insertion.) Scholars also have questioned the authenticity of a second passage in the same work, known as the "Testimony of Flavius" (XVIII, 63ff.), which is generally thought to contain at least some statements, apparently later insertions, that summarize Christian teaching about Jesus.

In the Talmud, a compendium of Jewish law, lore, and commentary, only a few statements of the rabbis (Jewish religious teachers) of the 1st and 2nd centuries come into consideration. Containing mostly polemics or Jewish apologetics, they reveal an acquaintance with the Christian tradition but include several divergent legendary motifs as well. The picture of Jesus offered in these writings may be summarized as follows: born the (according to some interpretations, illegitimate) son of a man called Panther, Jesus (Hebrew: Yeshu) worked magic, ridiculed the wise, seduced and stirred up the people, gathered five disciples about him, and was hanged (crucified) on the eve of the Passover. The *Toledot Yeshu* ("Life of Jesus"), an embellished collection of such assertions, circulated among Jews during the Middle Ages in several versions.

These independent accounts prove that in ancient times even the opponents of Christianity never doubted the historicity of Jesus, which was disputed for the first time and on inadequate grounds at the end of the 18th, during the 19th, and at the beginning of the 20th centuries.

Christian sources. Christian testimonies about Jesus were collected in the New Testament. Though they certainly represent only a selection from a much broader stream of tradition (Luke 1:1–4), these testimonies are a very valuable and representative selection. They are, however, of very different kinds. From many of them next to nothing can be learned about the historical Jesus.

The Pauline Letters. The oldest New Testament writings, the genuine letters of Paul (written in the 50s of the 1st century), contain little information about the life of Jesus. Paul, the Apostle, who had not known Jesus personally (II Cor. 5:16), shows no interest in Jesus' biography. At the centre of Paul's thought and proclamation there stands only the theologically important significance of the death, Resurrection, exaltation, and Second Coming of Jesus Christ, contained in numerous short doctrinal and creedal formulas. These formulas the Apostle himself occasionally characterizes as being the tradition that he has received and handed on (I Cor. 11:23ff.; 15:3ff.) or they are in other ways indicated as a given tradition (Rom. 1:3ff.; Phil. 2:6–11).

The importance of the Synoptic Gospels

The Gospels. The most important sources for the life of Jesus are the Synoptic (parallel view of sources) Gospels: Mark, Matthew, and Luke. The Gospel According to John, the Fourth Gospel, assumes a special position. Though it offers some parallels to the other three, and though the independent traditions in it may in individual cases have historical kernels, the tradition in John shows that the gospel has reached an advanced theological state. Because a theological conception has been incorporated in the account to such an extent, this Gospel cannot be directly used as a historical source. It is also the latest of the Gospels, written about AD 100.

That the gospel literature was capable of developing in very different directions is also shown by the extracanonical tradition about Jesus, which is preserved in fragmentary form in quotations by the early Church Fathers and in other sources and which is marked by legendary features and tendencies. The Coptic *Gospel of Thomas* (written in the 2nd century by Gnostic Christians; *i.e.,* heretical believers in esoteric, dualistic doctrines), which was found in 1945 in Naj' Ḥammādī (Egypt), is an example of such extracanonical literature. It contains 114 sayings of Jesus loosely strung together, which have some points of contact with the sayings of Jesus in the canonical Gospels. But this Gospel has no earthly, historical contours in its account of Jesus (*e.g.,* no accounts of the Passion and Easter). As a bearer of heavenly revelation in this Gospel, Jesus instructs the esoteric circle of his disciples about the foreign world of matter that they must renounce in order to participate in the imperishable, transcendent world of light from which they originate. The *Gospel of Thomas,* thus, is of no use as a source for the historical Jesus.

The Synoptic Gospels were originally anonymous. According to questionable 2nd-century tradition, they were written by the immediate disciples of Jesus or companions of the oldest Apostles. Most probably the Gospels were composed between AD 70 and 100. That they were written at such a relatively late time does not detract from their historical significance, however, because an older, oral tradition is collected in them and has left its traces everywhere. The character and structure of the individual traditions are incorporated into the Gospels, which definitely do not have a historical or biographical interest in facts, circumstances, and the course of events. They do not reproduce the story of Jesus as such but, instead, recount history interpreted from the viewpoint of the Christian faith. What Jesus says, does, and suffers is interpreted as the fulfillment of the Old Testament promises, and his story is slanted toward his end (the Passion and the Resurrection), his significance as the divine Saviour, and his Second Coming. In other words, the Gospel texts do not intend to describe the Jesus of the past but rather to proclaim who he is for all ages of time. These perspectives of the post-Easter church to which the writers belong and for which their reports are intended must continually be taken into consideration.

A comparison of the first three canonical Gospels reveals a strange blending of agreements and differences. Mark, Matthew, and Luke contain, by and large, the same traditional material. Some parts, however, are to be found only in Matthew and Luke, and a considerable amount of material is peculiar only to Matthew or only to Luke (and a small amount to Mark, as well). According to almost all critical biblical scholars, Mark, the shortest Gospel, is viewed as the oldest—not Matthew, as was earlier assumed—and served as the main literary source for the other two. They also believe that the material common to Matthew and Luke comes from a second source (called Q, from the German *Quelle,* "source"). This second source (Q) consisted almost exclusively of sayings (logia) of Jesus and contained no Passion or Easter tradition and is therefore known among scholars as the logia, or sayings, source.

Investigation of the Gospels by German biblical scholars such as Karl Ludwig Schmidt, Martin Dibelius, and Rudolf Bultmann—who developed what is known as form criticism, the study of the origin and development of the traditions in the Gospels—has shown that the basic stock of the tradition consisted of numerous small, self-contained units (single sayings, parables, debates, anecdotes, and miracle stories), originally without any relation to each other, and mostly without any interest in dates, places, or historical circumstances. It was the Gospel writers (or some earlier collectors) who first joined these individual pieces together editorially, forming a kind of "discourse" out of sayings and groups of sayings and, through linking individual scenes, creating the impression of a connected chain of events. They used a very modest set of tools for this; *e.g.,* short introductory and connecting phrases, stereotyped, generalizing indications of time ("next," "a few days later"), and frequently repeated, indefinite indications of place (mountain, field, road, house, lake). These editorial turns of phrase are, as a rule, easy to sever from their context and are employed very differently by the separate Gospel writers.

In methodically distinguishing and separating traditional and editorial features, form criticism of the Gospels has apparently dissolved the presuppositions for a historically sound, connected life of Jesus, which scholars have again and again attempted to write in the course of the last 200 years. But such an analysis was only a first step of research into the older material itself. Popular oral tradition, to which the Synoptic material belongs, makes use of fixed forms appropriate in each case to the contents, so as to be easily fixed in the memory. The tradition about Jesus offers many examples of this: prophetic sayings, the Beatitudes, pronouncements of woe, wisdom sayings similar to proverbs, legal sayings, church rules, dialogues, and others. In a corresponding way, many miracles of Jesus are narrated by means of motifs and other features also known from reports of other miracle workers. From this one perceives that this tradition is interested not so much in what was historically unique as in what was typical. Thus, with regard to the Gospels, it has to be considered that their tradition was formed and collected from the point of view of the faith of the post-Easter church, under the influence of its ideas and ways of thought and in close connection with its vital interests and the ways in which its life found expression. When interpreting the texts, scholars must therefore be concerned with the question of their setting in life (*Sitz im Leben*) in the church as well.

This critical survey of the sources shows that there are limits set on a portrayal of the historical Jesus. Many questions are still under debate or have to remain open.

TIMES AND ENVIRONMENT

Political conditions. Politically, the small Jewish nation in Jesus' time was rent and impotent. Always situated in an area of tension between the great empires of the ancient world (*e.g.,* Egypt, Assyria, Babylonia, Persia, and Syria) as they struggled with each other and succeeded one another, it had already lost its political independence since the time of the Babylonian Exile (586–538 BC) and had come under changing foreign domination: in the Hellenistic period, first under the Egyptian Ptolemaic dynasty (3rd century BC) and then under the Syrian Seleucid dynasty (2nd century BC), and, finally, until its ultimate overthrow (AD 70), under the Romans, who continued to rule the area. Only for a short interim was there a Jewish kingdom. The Maccabees, a priestly family, reigned after their revolt (168–165 BC) against Antiochus IV (Epiphanes), the Syrian king. Their rule, however, came to an end as a result of internal disintegration and violent struggles for the throne.

Judaea under Roman domination

Initially courted by the rival parties, the Roman general Pompey marched into Palestine, capturing Jerusalem in 63 BC, and reduced the Jewish territory to Judaea, without the coastal cities and the confederacy of towns of the Decapolis (central Transjordan). Several other smaller regions—*e.g.,* inland Galilee of the northern province around Lake Gennesaret and Peraea, east of the Dead Sea—were left to the Jews. By exploiting the threat to the Roman Empire from the Parthians and by adapting himself skillfully to the changing power situations after the murder of Julius Caesar (44 BC), the clever and adroit Herod I (reigned 37–4 BC) managed with the help of the Romans to become "king of the Jews" and to extend the Jewish state over almost all of Palestine again. His regime was decidedly progressive. He promoted Hellenization (*i.e.,* emphasizing Greek culture) by modern building projects, the founding of towns, and in other ways. But he also attempted to win the favour of the Jews, above all by rebuilding Solomon's Temple in ostentatious form and on an enormous scale. It was begun in 20 BC but was not finally completed until AD 64, a few years before its destruction in AD 70 by Titus, who became emperor of Rome nine years later.

Though the Jews demanded of the Romans the abolition of Herodian rule after his death, the Romans divided the land up among the sons of Herod the Great. The most important and largest part, Judaea, with Jerusalem, Samaria, southern Judaea, and Idumaea, was granted to Archelaus, who was deposed by AD 6. His area was integrated into the Roman administration under a governor (procurator), who controlled military, taxation, and judicial affairs. As was their custom, the Romans allowed the Jews to practice their religion and to exercise restricted powers of administration and jurisdiction. Some of the procurators, however, did not hold themselves strictly to these principles. Pontius Pilate, who is designated in an inscription found in 1961 as *praefectus Judaeae,* ruled (AD 26–36) ruthlessly and with bursts of cruelty. He was dismissed for this reason. The reigns of Herod's other two sons were of rather longer duration: Philip (4 BC–AD 34) ruled as tetrarch of the non-Jewish region northeast of Lake Gennesaret, and Herod Antipas (4 BC–AD 39) served as ruler of Galilee and the remote Peraea.

As far as Jesus' story is concerned, the conditions in Galilee, the land of his origin and his ministry, are of paramount importance. Thoroughly changed in character by the settlement of foreign colonists, although again in the process of being re-Judaized, Galilee was held in contempt by the Judaeans. Though the land's culture and civilization were in large measure Hellenistic, especially at the court of Herod Antipas, in individual towns and among the owners of large estates, the Jewish population, which spoke Aramaic, lived with its own, largely unaffected religious traditions. At the time of Jesus, Galilee was known as a seat of Jewish resistance to Rome.

According to Josephus (*Antiquities* XVIII, 18 ff.), Herod Antipas—whom Jesus called a "fox" (Luke 13:32)—held John the Baptist, the prophet who preached repentance, to be politically dangerous, had him put in prison, and had him executed for this reason. The Synoptic tradition, however, gives the Baptist's harsh criticism of Herod's unlawful second marriage as the reason (Mark 6:17–29).

Various sectarian Jewish groups at the time of Jesus

Information about political conditions in Palestine at the time of Jesus is found mainly in non-biblical sources, especially in Josephus. Only a few details are mentioned in the Gospels. Such information is nevertheless significant as background for the story of Jesus, even if it does not contribute much to an understanding of his teaching. The attitude of the Jewish people to the foreign rule of the Romans was not uniform. There were conformists, especially among the priestly aristocracy in Jerusalem, and there were those who exhibited concealed and open resistance.

Religious conditions. Judaism in the time of Jesus presents a disunited, fragmented picture, composed of widely different groups.

Significance of form criticism

The Pharisees. In the reports of the Synoptic Gospels, the Pharisees serve almost entirely to exemplify his opponents. They are incensed by his preaching and behaviour, spy on him, press from the very beginning to have him done away with, and are, conversely, themselves attacked by him most fiercely as being self-righteous hypocrites. Debates with the Pharisees without doubt played an important role in Jesus' life. From the Gospels there has developed a crude popular view that "Pharisee" is synonymous with "self-righteous hypocrite." The New Testament sources, however, are to be used with discretion in this respect for the following reasons: (1) the later narrators were in large measure no longer conversant with the historical circumstances, especially because they were themselves outside the region of Palestine. As a rule, the Pharisees are introduced as a collective quantity in the Gospels but, in reality, were not a unified group. There are also sporadic references in the Synoptic tradition to the fact that Jesus maintained table fellowship with Pharisees (Luke 7:36; 11:37; 14:1). It is also worthy of note that they play no part in the Passion tradition (with the exception of the later legend in Matt. 27:62ff. about the Pharisees' requesting a guard at Jesus' tomb). (2) A Synoptic comparison reveals the tendency to give Jesus' opponents more concrete form, but in a schematic way. In the later texts, the Pharisees are frequently introduced as the constant foil for Jesus, whereas the older tradition speaks of Jesus' opponents in an indefinite way. (3) Matthew, especially, reflects the sharpened conflicts between Jews and Christians in the period after the destruction of Jerusalem (AD 70), when a theologically narrower brand of Pharisaism was finally asserting itself in the course of the religious reconstitution of Judaism. This later picture dominates the Talmudic tradition, but it may not be projected back into the time of Jesus.

Originating in the time of the Maccabees (or earlier, according to some scholars), the movement of the Pharisees (*i.e.,* the "separated ones") formed itself into a religious association composed chiefly of laymen from varied classes and callings. Its aim was strict adherence to the Torah (Law) in even the most remote areas of daily life, in order to realize the true Israel of God. This especially included the scrupulous observance of the individual ritual commandments for the practice of prayer and fasting, cultic purity, and the avoidance of all contact with the cultically unclean, whether that be lawless persons, sinners, corpses, animals, or unclean utensils. In the Pharisees' piety there was also to be found an eager longing for the future world of God, a doctrine of the resurrection of the dead, and a hope in the promised Davidic Messiah, who would establish his rule in Jerusalem and destroy the power of the heathen.

Jesus' relation to the Pharisees

In view of this religious situation, it is difficult to arrive at a uniform judgment on Jesus' relation to the Pharisees. Points of contact in matters of teaching definitely are present; *e.g.,* in the expectation of the resurrection of the dead, which they hold in common (Mark 12:25–27). Again, there are critical statements about formalized and hypocritical piety in Jewish Talmudic tradition, and not just in sayings of Jesus. It would therefore be unjust to judge all Pharisees to be alike. Obviously, many sayings of Jesus have parallels in Rabbinic tradition. Nevertheless, there is no question that Jesus rejected their claim to righteousness and their ideal of representing the true Israel, that he characterized their "tradition of the elders" as human tradition in contrast to the commandment of God, and that, through his attitude to tax collectors and sinners, he must have given them offense. Because of such opinions of Jesus, they probably would have influenced the people against Jesus. That certainly need not mean, however, that the Pharisees, who were politically anything but dominant, aimed at Jesus' crucifixion from the start (contrary to what is said; Mark 3:6).

The Sadducees. A party of quite another kind was that of the Sadducees, who belonged to the Jerusalem priestly caste. They carried much less authority among the people than the Pharisees. As a theologically conservative school, they differed from the latter also in their rejection of the additional "traditions" and the new doctrine of the resurrection of the dead. Because of the Sadducees' hierarchical tradition and their readiness to adapt themselves to the current political conditions, their influence in Jesus' time, before the destruction of the Temple, is not to be underestimated. Besides the Pharisees and the elders of the people, they had a decisive voice in the supreme religious and judicial authority, the Sanhedrin. A close relation probably existed between them and the Roman rulers. They did not, however, survive the catastrophic outcome of the war and the end of the Temple (AD 70).

The scribes. The scribes are frequently mentioned in the Gospels. In later Judaism, which, since the time of Ezra (5th century BC), was committed to the Mosaic Law, they formed a most respected class of the teachers. Corresponding to the normative significance of the Law for all religious, moral, social, and legal questions of Jewish life, the scribe was a combination of theologian and lawyer. Social origin and membership of a particular party played no role in this group. In Jesus' time, there were, apart from Pharisees and priests, also Sadducees and Zealots among the scribes. They were not paid as a professional class but, instead, had to find their own living. As scribes, they had to expound the Torah and give directives for daily life. Those who had undergone the long and careful training in their schools were accorded the status of scribe, wore the long robe of the scholar (Mark 12:38), were respectfully addressed as "rabbi" (Matt. 23:7), and were allowed to sit in a place of honour in the synagogue. Jesus, like the scribes, sat to teach (Matt. 5:1; Luke 4:20), engaged in debate, gave his opinion on the diverging doctrinal propositions of particular schools (Matt. 19:3ff.), and gathered disciples about him. The stereotyped way in which, particularly in Matthew, Pharisees and scribes are grouped together reflects the conditions obtaining at the time of the Gospel-writers, in which it was the Pharisees who controlled the instruction in the synagogues exclusively. But earlier, in the time of Jesus, the scribes were a more motley group. Also, it is not allowable to conclude from the fact that Jesus is frequently addressed as "rabbi" and "teacher" that he himself was a member of this profession.

Theocratic ideals and zeal for the Law among the Zealots

The Zealots. The involvement of the religiopolitical movement of the Zealots, a revolutionary group, in the historical development of Palestine was disastrous to the nation. No longer contented with the passive resistance of the Pharisees, out of whose ranks they certainly gained many adherents, the Zealots took the ideal of a theocracy and zeal for the Law extremely seriously. The first outbreak of their activities occurred in AD 6, when the Syrian legate Quirinius ordered the population in Judaea to register. This aroused indignation and was the signal for an insurrectionist movement, which confined itself initially to scattered individual acts of revolt but soon expanded, took military form, and finally instigated the First Jewish Revolt (AD 66–70). Biblical and nonbiblical sources name Judas, a Galilean scribe from Gamala, as founder of the Zealots. Like him, other fanatical messianic prophets also found significant followings. In Jesus' time, the conflict had not yet reached its zenith. The Zealots carried out sudden raids on the Roman occupation forces and conducted a guerrilla war from their hiding places in the wilderness. The Romans correspondingly held the land under strict control, reinforced their troops in Jerusalem at the times of the Jewish festivals, when great crowds of pilgrims gathered in the city, and took drastic and ruthless action if they anticipated sedition. This situation illuminates the events leading to Jesus' death. The Zealots' goals were political and, primarily, religious: the realization of a Jewish theocracy, the rule of the promised Messiah, and the destruction of the heathen regime.

The thesis that Jesus belonged to the Zealots or founded a related movement was first advanced in the 18th century and has repeatedly been supported in recent times. The most important point in its favour is Jesus' execution on the cross, a punishment that only the Roman authorities could inflict and did frequently against rebels. There were two others executed in the same manner with Jesus, and they, like Barabbas, who was granted amnesty in Jesus' place (Mark 15:15), are referred to as "robbers" (Mark 15:27), a customary term for rebels at this time. This could indicate that, at that Passover time, when many Jews were in the city, a Zealot revolt had been planned and was bloodily suppressed but also that Jesus had actually been willing to play a leading part in it.

Jesus' messianic entry into Jerusalem and the cleansing of the Temple (Mark 11) are also interpreted along these lines, the latter being understood as an attack on the dominant priestly class that sympathized with the Romans. Some also see a connection with the fact that one of the disciples was carrying a weapon when Jesus was arrested in Gethsemane (Mark 14:47). The later Christian tradition has, it is claimed, for apologetic and theological reasons, altered the true historical state of affairs until it has become unrecognizable. But isolated hints have nonetheless been preserved in it; *e.g.,* Jesus' critical sayings about that "fox" Herod (Luke 13:32) and the violent earthly rulers (Luke 22:25); similarly, the way he attracted Zealots, documented by the fact that among his disciples at least one, called Simon (Luke 6:15; Acts 1:13), was a Zealot.

There are, however, no sufficient reasons to support the hypothesis of Jesus belonging to the Zealots. The undeniable fact that he was crucified by the Romans as a political messianic pretender only proves that he was held to be a Zealot and was probably denounced as an enemy of the state, but not that he really was. The most important and decisive argument against the Zealotism assumption is found in Jesus' message of the dawning of the Kingdom of God, which belongs to the best established items in the tradition. It lacks any politico-nationalistic features and expressly says that God alone, and not any human activity, establishes his Kingdom (Mark 4:26–29) and offers his salvation to all without exception. If Jesus were directly or indirectly to be counted among the Zealots, this would mean at the same time that he must have fought to have the Law rigorously carried into effect and must have strictly avoided associating with sinners, especially with the tax collectors, who stood in the service of Rome. In the dialogue on paying tribute to Caesar (Mark 12:13–17), Jesus even expressly rejected rebellion against the Roman emperor, without thereby glorifying his regime.

The Essenes. Far removed from the above-mentioned religious groups were the sectarian, separatist Essenes, most probably identical with the sect of Qumrān (near the northwest bank of the Dead Sea). The sensational discovery in 1947 of many of their original writings (the Dead Sea Scrolls) and the later excavations of their settlement have extended knowledge of the Jewry of those times to an extraordinary degree and have occasioned the suggestion that both John the Baptist and Jesus came from this sect or were, at the least, heavily dependent on their teaching. Important arguments, however, speak against this assumption. This sect had arisen, like the Pharisees, in the 2nd century BC out of a conflict with the official priesthood in Jerusalem but had nevertheless preserved the priestly traditions and, at the same time, developed a strongly ritualistic practice of the Law. Characteristics of the Qumrān community are: its monastic seclusion from the outside world, including the rest of the Jews; the way it termed itself the "children of light" in contrast to the "children of darkness"; its rigid organization and discipline; and its apocalyptic expectations—centring on the intervention of God in history, along with dramatic and cataclysmic events—and other special features of its theology. Although the new texts found at the Dead Sea show numerous individual parallels to the Jesus tradition of the Gospels, there are already fundamental differences between the Qumrān sect and John the Baptist. His eschatological (last times) message of repentance addressed to the nation as a whole fits in with the sect as little as does his unique kind of Baptism—which one underwent once and for all—with the Essenes' regular ritual washings. Nor does John's temporal and geographic proximity to the strictly esoteric Qumrān sect justify asserting close relations between them. There also are diametrical differences between the views of the sect and the range of Jesus' ministry, his message of salvation, his understanding of God's will in a way free of all casuistry, and, especially, the radical character of his commandment of love and his fellowship with sinners and social outcasts.

Characteristics of the Qumrān community

THE LIFE AND MINISTRY OF JESUS

The birth and family. *The birth of Jesus.* The course of Jesus' life and the geographical setting of his ministry can only be given in rough outline. The details are surrounded by many uncertainties. The period within which his ministry and death occurred may, however, be narrowed down with considerable accuracy on the basis of a synchronistic dating of the appearance of John the Baptist in the 15th year of Tiberius (Luke 3:1)—*i.e.,* AD 28/29—which is confirmed by nonbiblical sources. But the year and place of Jesus' birth are uncertain. Mark and John say nothing about them. The only sources for them are the widely divergent birth and childhood legends in Matthew 1 and 2, where Jesus' birth and early lot are set in the time of Herod I and the change of regime (4 BC), and the narrative of Luke 2, which links Jesus' birth with the first registration in Judaea under the emperor Augustus (AD 6). There is also historical evidence of a census carried out around 8 BC. With all of this in mind, many sources estimate the year of birth as 7–6 BC. (The use of BC [before Christ] and AD [Anno Domini, or "in the year of the Lord"] was not common until the Middle Ages.)

Davidic descent of Jesus

The tradition of Bethlehem as the place of Jesus' birth has its source in all probability in the Old Testament conception of the Messiah as a descendant of David. Early Christianity took this view from the beginning. "Son of David" is found in many texts (*e.g.,* Mark 10:48) alongside other titles of Jesus. Its original political and national sense was abandoned, even though it is still recognizable in the acclamation of the people (Mark 11:10). The theological motif of Jesus' Davidic descent, however, did not necessarily involve the idea that he was born in Bethlehem, David's hometown. That is the case only in Matthew 2 and Luke 2. The accounts differ in that, in Matthew, Bethlehem is thought of as the parents' original place of residence, which they soon change to Nazareth because of the dangers threatening their child (*e.g.,* the flight to Egypt), whereas, in the Lucan story, Jesus' parents really live in Nazareth but stay in Bethlehem temporarily because they are obliged to register at the Davidic family's place of origin. Both traditions are to be judged as legendary variations of the theological theme of Jesus' messiahship, even though each in its own way assigns to his birth a place in history. The extent to which these texts are marked by theological motifs, above all by the thought that Jesus as Messiah fulfills the promises of the Old Testament and the hope of Israel and the world, is shown by the numerous quotations woven into the stories. (G.Bor./Ed.)

The widely differing genealogies in Matthew 1 and Luke 3 also belong in the context of the doctrine of the Davidic descent of the Messiah (Christ). They are the only New Testament evidences for genealogical reflection about Jesus' messiahship. The two texts, however, cannot be harmonized. They show that originally a unified tradition about Jesus' ancestors did not exist and that attempts to portray his messiahship genealogically were first undertaken in Jewish Christian circles with use of the Septuagint (Greek translation) text of the Old Testament. Both texts have to be eliminated as historical sources. They are nevertheless important for the development of Christology (doctrines on the nature of Christ), because they reveal the difficulty of reconciling the genealogical proof of Jesus' Davidic descent with the relatively late idea of his virgin birth.

This last tradition, too, is recorded in only two stories—in Luke 1 and Matthew 1—and was originally quite unconnected with the frequently found motif of Jesus' divine Sonship. Paul, John, and the rest of the New Testament writers are not acquainted with the idea. Also, it has left no traces in the rest of the Synoptic tradition, not even in the story of Jesus' birth (Luke 2:1–10), where Joseph and Mary appear as his natural parents. In Matthew 1, Jesus' miraculous birth is presupposed, and, in Luke 1, it is explained more closely. This tradition is not to be traced back directly to the idea, widely held in classical antiquity, of heroes and great men who derived from the union of a deity with a human woman. In other words, Jesus is not characterized as a demigod here. What underlies this tradition is, rather, the concept of the creative power of God and his Spirit, which is known from Hellenistic Judaism. This theological, not biological, motif has been applied to Jesus and, with the greatest probability, only secondarily combined with the Greek version of the messianic promise of Isa. 7:14 (in the Septuagint, the Hebrew word *'alma—i.e.,* "young woman"—is translated as "virgin"), and in this way the Christian story came about. According to a very old, reliable tradition, the village of Nazareth, which lay in the Galilean hill country, had a Jewish population, and was untouched by the influence of the Hellenistic cities, was the hometown, and then certainly also the birthplace, of the "Nazarene" (Mark 1:24; 10:47; 14:67; 16:6).

The Jewish character of Jesus' parents, brothers, and sisters

The family of Jesus. Four of Jesus' brothers and several sisters are mentioned in Mark 6. (There is no basis in the text for making them into half brothers and half sisters or cousins, and to do so betrays a dogmatic motive.) All his relatives' names testify to the purely Jewish character of the family: his mother's name was Mary (Miriam), his father's, Joseph, and his brothers', James (Jacob), Joseph, Judas, and Simon (names of Old Testament patriarchs). The same is true of the name Jesus. In the Septuagint it is the customary Greek form for the common Hebrew name Joshua; *i.e.,* "Yahweh helps." It is also mentioned in Mark 6 that Jesus or his father (there are variant textual versions) was a carpenter. There are several not unimportant pieces of information preserved about the later history of the family. Of his father, who probably died early, little is mentioned. His mother, brothers, and sisters did not join his movement at first but, rather, disapproved of his behaviour (Mark 3:31–35). Mary is, however, mentioned as a member of the Christian Church after his death (Acts 1:14). The same is true of his brother James, whom Paul names among the witnesses of the Resurrection (I Cor. 15:7) and who was the leader of the Jerusalem Church after Peter (Galatians, Acts). The author of the Letter of James has taken a brother's name for himself, as did the author of the Letter of Jude in respect to another brother. According to a later nonbiblical account (in the *Ecclesiastical History* of Eusebius, a 4th-century historian of the church), grandchildren of Jude (who otherwise remains unknown), who were living in Galilee, were summoned by the emperor Domitian as "descendants of David," but then released as representing no political danger.

Jesus most likely grew up in the piety that was cultivated in the home and in the synagogue (including Bible study, obedience to the Law, prayer, and expectation of the final coming of the Messiah) and also took part in pilgrimages to Jerusalem. From these scattered reports it is possible to gain some information about Jesus' background and theological education. The latter also comes to light in his teaching and in the frequently attested honorific form of address "rabbi" (teacher), which, in the language of the time, was not yet confined to members of the trained and ordained profession of the scribes. Nothing is precisely known, however, about Jesus' youth and inner development. What is known is contained in the sole narrative in Luke 2:40–52 (the boy Jesus in the Temple) and the legendary apocryphal gospels, which, after the manner of legend, sought to illumine the obscurity of Jesus' childhood.

The Christian significance of John the Baptist

The ministry. *The role of John the Baptist.* The Gospel accounts of the appearance and activity of John the Baptist and of Jesus' Baptism at his hands first establish a historically safe basis for knowledge of Jesus' life and work. Significantly, the oldest Gospel writer calls these events "the beginning of the gospel of Jesus Christ" (Mark 1:1), indicating that his would be a message about Christ, not a description of the contemporary background for Jesus' life. The Baptist is, therefore, represented exclusively from the Christian point of view. His place in the Christian history of salvation is that of a forerunner or pioneer; or he is a witness to Jesus, as in the Gospel According to John. But the tradition has nevertheless preserved unchallengeable information about John, especially in Q. Josephus characterizes him as a mere moral teacher and his Baptism as merely ritual washing. In reality, however, he made his appearance in the desert as a prophet of the imminent Last Judgment, calling all without exception to repentance in the eleventh hour, and baptized those who were ready to repent, in order to prepare them for the baptism of fire of the mightier one coming from heaven and to preserve them from his annihilating wrath (Matt. 3:7ff. and Luke 3:7ff.). His dress and diet as an ascetic nomad and, above all, the location of his ministry (the Judaean desert and the Jordan steppes), far away from the institutions and places of traditional religion and secularity, illustrate the earnestness of his eschatological preaching and his attack on all conventional piety; but they also correspond to the old prophetic promise that God would encounter his people in the Last Days, as he did once before, in the desert. Historically, all these features may not be understood immediately in Christian perspective; *i.e.,* as pointing to Jesus as the Messiah. The tradition of the Gospels visibly and increasingly interpreted the history of the Baptist in retrospect, and not least for the reason that there still existed for a considerable time alongside the disciples of Jesus a rival body of disciples of the Baptist.

The significance of Jesus' Baptism

That Jesus was baptized by John, as all the Gospels record, indicates that in all probability Jesus initially belonged to John's movement. The account of Jesus' Baptism is styled in the Gospels as an "epiphany (or manifestation) story" and deals with Jesus' installation at this time as Messiah (Mark 1:9–11). The announcement of the Kingdom of God by John and his call to repentance retained decisive significance for Jesus. His high estimate of the Baptist emerges unambiguously from the fact that he placed John above the prophets and called him the greatest among men (Matt. 11:7–11). He saw the signs of the approaching Kingdom of God in the work of the Baptist as in his own work, and he recognized the authority given John as being from heaven (Mark 11:27–33). These words carry all the more weight historically, because the tendency of the context here is to proclaim Jesus as the Messiah and to place the Baptist, as the lesser, in Jesus' service. It is significant that John himself is nowhere attacked in the Synoptic texts, nor is he designated as a follower of Jesus. Wherever polemic can be recognized in the Gospels (especially in John), it is always directed against the false belief, doubtlessly held by the (later) Baptist disciples, that John was the promised Messiah. The extent to which the close connection between Jesus and John occupied the theological reflection, apologetics, and imagination of the Christian Church is shown by several passages and, above all, by the cycle of legends in the introduction to Luke (chapter 1). Regardless of the close relationship between Jesus and John the Baptist, especially in their prophetic announcement of the approaching Kingdom of God and their call to repentance (*cf.* Matt. 3:2; 4:17), there are also radical differences.

The beginning of the ministry. At the latest, after the Baptist's imprisonment (as the Synoptics state), possibly even earlier (according to John), Jesus began as a grown man (Luke 3:23) an independent public ministry, but in the villages of his Galilean homeland and—sporadically—in the neighbouring countryside, rather than in the wilderness, as did John. The real area of his ministry was the district on the northwest bank of the Lake of Gennesaret (or Sea of Galilee; the towns of Beth-saida, Chorazin, and Capernaum). The change of scene is significant in itself. Jesus did not call the people into the desert. He sought men in their settlements and took part in their ordinary life, and not as an ascetic, like John the Baptist (Matt. 11:18). He worked among them as a wandering preacher (Matt. 8:20) and charismatic miracle worker, without, however, baptizing like John. But the image he presents is nonetheless highly peculiar. He taught not only in the synagogues but likewise in the open air, on the shore of the lake, and on the road. There also were strange people in the group surrounding him: women, children, and many who were viewed as godless or unclean. Further, the manner of his teaching is surprising. He did not derive it from the Holy Scriptures, although he was familiar with them, esteemed them, and appealed to them here and there. Instead, he constantly presented the reality of God and the validity of his will in an immediate way and made them comprehensible to his hearers without using the established structure of sacred texts and traditions and without presupposing a conventional, religious point of view. His metaphors, parables, and proverb-like utterances were not used to explain traditional teachings of biblical theology but, instead, appealed directly to the everyday experience and the understanding of his hearers, and they are therefore characterized by a unique self-evidence and a disarming simplicity.

Jesus' relationships to various people

This corresponds to the manner of his behaviour in his meetings with other people. The Gospels portray this in a large number of separate scenes. These persons vary considerably: pious and impious, rich and poor, respected and outcast, healthy and ill. In every encounter, Jesus' amazing sovereignty with which—free of all prejudices—he mastered the situation is made visible. He saw through his opponents' attempts to corner him in debate, disarmed their objections, saw the needs of the possessed and the sick who crowded around him, and associated with those who were avoided by others. Some of the scenes may only have been added or filled out in later popular tradition, but they clearly demonstrate the power with which Jesus helped people by word and deed, whether he grew passionately angry over the power of disease or over the pride and lovelessness of the "righteous" or whether he commanded the demons or blessed children and laid hands on the sick.

The calling of the disciples. According to the unanimous witness of the Synoptic Gospels, Jesus gave rise to a movement in Galilee and found numerous followers, although not without provoking rejection as well. This movement cannot yet be called a "church." (This concept first appears in the later tradition.) To spread his message and movement, he called on his disciples, for the sake of the approaching Kingdom of God, to resolutely surrender all ties of family and work (Mark 8:34ff.; Matt. 10:37ff.; Luke 14:26ff.) and to follow him and to become "fishers of men" (Mark 1:17; Luke 5:10). Many of his words are of extreme sharpness and do not conceal how difficult the disciples' road will be (Luke 14:25–33). But, at the same time, the patent immediacy of Jesus' sovereign power comes to light in these texts. In the scenes mentioned, it is Jesus who makes the decision. He calls, appoints, and selects particular men, without regard to their origin and previous training. There are fishermen (Andrew, Peter, James, and John), a tax collector (Matthew), and Zealots (Simon and, perhaps, Judas Iscariot) among them, perhaps also a few craftsmen and peasants. Whether it was a circle of 12 disciples from the start is questionable and under debate. It is clear, however, that he commissioned and authorized his disciples to preach and to drive out demons (Mark 3:14). Some of these disciples are well noted in the Synoptic tradition (*e.g.,* Peter and Judas Iscariot). In the Gospel According to John, others come into the foreground, including some from among the followers of the Baptist. Of others, only their names are known (*e.g.,* Thaddaeus). A characteristic of these companions of Jesus is that their discipleship is not, as with the rabbis, a transitional stage that ends with their "training." None of them moves up after sufficient study to the status of "master" (Matt. 23:8). Even if accounts of the calling of disciples have, in general, been styled in the later tradition as examples of what it means to be a Christian and individual scenes have been added to the original stock of stories, the recollection of incidents that occurred during Jesus' ministry in Galilee is doubtlessly preserved in the texts.

Origins of the disciples

The Galilean period. The loose and often differing order of the individual scenes only entitles scholars to speak of a rather ambiguous Galilean period of Jesus' activity: they cannot say with certainty how long it lasted. Because the Synoptic Gospels mention only one trip of Jesus to Judaea and Jerusalem, with the Passion following it, the impression is created that the period lasted no longer than one year. Editorial and theological considerations have, without question, also played a part in this presentation (*e.g.,* Jesus' activity in Galilee and his sufferings in Jerusalem). Scholars offer several good reasons, however, to support the assumption that the Synoptic outline still deserves to be preferred to the widely differing one in John. In the latter, Jesus is in Jerusalem for three celebrations of the Passover (John 2:13–23; 6:4; 11:55), as well as for one Sukkot (Feast of Tabernacles; John 7:2) and one Ḥanukka (Feast of Dedication; John 10:22). This involves a period of more than two full years. It is doubtful, however, that

John is based on an independent tradition, because the indications of time referred to serve the Evangelist as a means of changing the scene of Jesus' ministry between Jerusalem and Galilee. (The centre here is Jerusalem.)

THE MESSAGE OF JESUS

The central and dominant theme of Jesus' teaching

The Kingdom of God. Jesus announced the approaching Kingdom of God and therefore called people to repentance. The first two Gospels have set this at the beginning in a programmatic saying as a summary of his preaching and have thus characterized the central and dominant theme of his mission as a whole (Mark 1:15; Matt. 4:17). Thus, the Kingdom of God, or Kingdom of Heaven (a Jewish circumlocution for God preferred by Matthew), does not just denote a final chapter of his "system of doctrine" (a concept that cannot be applied to Jesus, in any case). The underlying Jewish word (*malkhuta*) means God's kingship, and not primarily his domain. This meaning prevails in the New Testament texts. But Kingdom of God or Heaven is also used in a spatial sense ("Enter . . ."). The burning expectation of the Kingdom of God was widely spread in contemporary Judaism in manifold form, based on the Old Testament faith in the God of the fathers, the Creator and Lord of the world, who had chosen Israel to be his people. But with this faith there had united itself the contradictory experience that the present condition of the world was ungodly, that Satanic powers reigned in it, and that God's kingship would only manifest itself in the future. In wide circles, this expectation had the form of a national, political hope in the Davidic Messiah, though it had expanded this hope in apocalyptic speculation to a universal expectation. In each case it was directed toward the Last Days. Likewise, in Jesus' message, the expression Kingdom of God has a purely eschatological—*i.e.,* future—sense and means an event suddenly breaking into this world from the outside, through which the time of this present world is ended and overcome.

Eschatological motifs

These traditional motifs of the end of the world, the Last Judgment, and the new world of God are not lacking in the sayings of Jesus preserved in the Gospel tradition. Thus, Jesus has not by any means changed the Kingdom of Heaven into a purely religious experience of the individual human soul or given the Jewish eschatological expectation the sense of an evolutionary process immanent in the world or of a goal attainable by human effort. Some of his parables have given rise to such misunderstanding (*e.g.,* the stories of the seed and harvest, the leaven, and the mustard seed). In such cases, the modern thought of an organic process has been wrongly introduced into the texts. People of classical and biblical times, however, heard in them connotations of the surprising and the miraculous. The Kingdom of God, thus, is not yet here. Hence the prayer, "Thy kingdom come!" (Matt. 6:10; Luke 11:2), and the tenses, for example, in Jesus' Beatitudes and predictions of woe (Luke 6:21–26). The poor, the hungry, and the weeping are not yet in heaven. The petitions of the Lord's Prayer presuppose the deeply distressing circumstance that God's name and will are abused, that his Kingdom is not yet come, and that men are threatened by the temptation to fall away.

In regard to Jesus' preaching, one cannot, therefore, speak of a realized eschatology—*i.e.,* the Last Times are now here (according to the view of C.H. Dodd, a British biblical scholar)—but of an eschatology "in process of realizing itself" (according to the view of Joachim Jeremias, a German biblical scholar); for God's Kingdom is very close. It is on the threshold, already casts its light into the present world, and is seen in Jesus' own ministry through word and deed. In this, his message differs from the eschatology of his time and breaks through all of its conceptions. He neither shared nor encouraged the hope in a national messiah from the family of David, let alone proclaimed himself as such a messiah, nor did he support the efforts of the Zealots to accelerate the coming of the Kingdom of God. He also did not tolerate turning the Kingdom of God into the preserve of the pious adherents of the Law (Pharisees; Qumrān sect), and he did not participate in the fantastic attempts of the apocalyptic visionaries of his time to calculate and thus depict in detail the end of the present world and the dawn of the new "aeon," or age (Luke 12:56). Nor did he undertake a direct continuation of the Baptist's preaching.

All the ideas and images in Jesus' preaching converge with united force in the one thought, namely, that God himself as Lord is at hand and already making his appearance, in order to establish his rule. Jesus did not want to introduce a new idea of God and develop a new theory about the end of the world. It would therefore be incorrect to understand his preaching in the Jewish apocalyptic sense of immediate expectancy, coming, as it were, to a boiling point. The proximity of the Kingdom of God actually means that God himself is at hand in a liberating attack upon the world and in a saving approach to those in bondage in the world; he is coming and yet is already present in the midst of the still-existing world. In Jesus' message, God is no longer the prisoner of his own majesty in a sacral sphere into which pious tradition had exiled him. He breaks forth in sovereign power as Father, Helper, and Liberator and is already now at work, as is indicated by Jesus' proclaiming of his nearness and by Jesus' actions in entering the field of battle himself, to erect the signs of God's victory over Satan: "But if it is by the finger of God that I cast out demons, then the kingdom of God has come upon you" (Luke 11:20). For this reason, Jesus called out: the shift in the aeons is here; now is the hour of which the prophets' promises told (Matt. 11:5; Isa. 35:5). This "here and now" carries all the weight in Jesus' message: "Blessed are the eyes which see what you see! For I tell you that many prophets and kings desired to see what you see, and did not see it, and to hear what you hear, and did not hear it" (Luke 10:23–24). In answer to the Pharisees' question about when the Kingdom of God is coming, Jesus therefore said, "The Kingdom of God does not come in an observable way, nor will they say, 'Look, here it is!' or 'There!' For look, the Kingdom of God is within your reach" (Luke 17:20–21; another translation: "in the midst of you").

The dominant feature of Jesus' preaching is the Heavenly Father's turning in mercy and love to the suffering, guilty, outcast, and to those who, according to the prejudices of the "pious," have no right to receive a share in the final salvation. Numerous parables described how God behaves toward them and shows himself as Lord and King (*e.g.,* Luke 15; Matt. 18:23ff.; 20:1ff.). They all speak of God's action in images drawn from daily life, so that everyone can understand. They belong to the uncontestedly oldest stock of the Jesus tradition. But Jesus did not only teach this, he practiced and illustrated it himself by his own behaviour and thereby offended the pious, who claimed the Kingdom of Heaven for themselves.

The message of repentance

In this message of the approaching Kingdom of God, Jesus' call to repentance is grounded. He called on all not to miss the hour of salvation (Luke 14:16ff.; 13:6ff.), to sacrifice everything for the Kingdom of God (Matt. 13:44ff.), and to receive it like a child (Mark 10:15), without the presumptuous and desperate conceit that one might win it and realize it by one's own works (Mark 4:26ff.; Matt. 13:24ff.). Jesus' summons to be wise, to be on the watch (Luke 16:1ff.; 12:35ff.; Mark 13:33ff.; Matt. 24:45ff.), and to surrender the fiction of one's own righteousness (Luke 18:10ff.) belongs here, too. In Jesus' preaching, repentance does not mean a prerequisite or precondition or even a penitent contemplation of oneself but, rather, a consequence of the proximity of the Kingdom of God (Matt. 4:17) and an opening of oneself for his future, a movement not backward, but forward. Jesus in this way binds future and present insolubly together. The apocalyptic's question about how much time still has to elapse before the new world of God is here is thus rendered meaningless. He who asks this only proves that he understands neither the future nor the present properly; namely, God's future as the salvation that is already dawning and one's own present in the light of the coming Kingdom of God.

Jesus therefore rejected the demand that he produce "signs" as proof of the dawning of the time of salvation (Matt. 12:38ff.; Mark 8:11). He himself is to be viewed as the "sign," just as once Jonah, the prophet of repentance, was the only sign given to the people in Nineveh (Luke 11:29ff.). The sign is not identical with the thing signified, but it is a valid indication of it.

According to the Synoptics, Jesus never made his "messiahship" the subject of his teaching or used it as legitimation for his message. It is significant that the "I am" sayings of John, which bear the stamp of Christology throughout, are not found in the Synoptic tradition. That does not in any way affect the fact that Jesus in a decisive way included his own person as eschatological prophet and charismatic miracle worker in the event of the Kingdom of God: "And blessed is he who takes no offense at me" (Matt. 11:6).

The will of God. In Jesus' teaching, the nearness of God is itself viewed as a moving force. It creates, as it were, a field of force and challenges the whole person to obey the will of God unconditionally ("Let your loins be girded and your lamps burning"; Luke 12:35). As little as Jesus tolerated attempts at calculating the time when the Kingdom of God should come, so much the more did he demand that men reckon with its coming. The relation between eschatology and ethics in Jesus' teaching, however, needs to be further clarified. His commandments nowhere have the character of prophetic sayings, and their content is not given an eschatological basis even where Jesus linked them with the promise of heavenly reward and, correspondingly, with the threat of damnation in the Last Judgment (*e.g.,* Matt. 24:24ff.; Luke 19:11ff.). God's will is valid in itself, always and everywhere. For this reason, it is incorrect to characterize Jesus' demands as "interim ethics"; *i.e.,* as exceptional emergency laws in the situation of the world that lies in the blaze of the cosmic catastrophes accompanying the shift of the aeons and the speedy dawn of the Kingdom of God (as did Albert Schweitzer, a great Alsatian theologian, medical missionary, and Nobel laureate). Jesus did not draw arguments for his ethical demands from the perishing order but, rather, from the existing world, the Old Testament commandments, the creation, and experiences known to everyone. Thus, he did not aim at forming a "holy remnant," which would escape the rejection awaiting others in the Last Judgment, on the basis of some kind of select monastic rule.

Eschatology and ethics

The certainty of God's nearness is, nevertheless, the open or concealed point of reference for Jesus' exposition of the will of God and explains his attitude to the Old Testament Law. Corresponding to the character of the Old Testament legal tradition, he refers to the will of God in single sayings and in comments in relation to individual commandments, and, it should be noted, he did not develop these into coherent "moral teaching." Rather, he took up quite different kinds of commandments as concrete examples, above all from the Decalogue and related texts, concerning one's behaviour toward one's fellow human beings (on murder and anger, adultery and divorce, oaths, retaliation, love for others; see Matt. 5:21ff.) and also ceremonial commandments (concerning the Sabbath, prayer, fasting, and defilement) and other cultic duties. Jesus always went to the root of these commandments, and he did not content himself with the mere letter of the Law but disclosed within the Law—sometimes even against the letter of the Law (Mark 10:1ff.)—the genuine will of God. Though Jesus respected the Law, it was no longer for him the only source of the knowledge of God's will and no longer the absolute intermediate authority that exclusively mediates people's relation to God. From this basis are to be understood both Jesus' exposition of the Law and also his criticism of all formalistic casuistry, which is for him only "human tradition."

Jesus thus brings about a confrontation between the reality of God, which is no longer disguised by holy letter and tradition, and the similarly undisguised reality of man. People also can no longer delude themselves into believing that their pious works would represent them before God and thus keep on piling them up, as it were, like the Pharisee (Luke 18:11ff.). What God wants from humanity is not something but humanity itself, unconditionally and undividedly. The classic passages for these thoughts are the antitheses of the Sermon on the Mount (Matt. 5:21–48). They sharpen God's demands to the utmost extreme and leave no room for merely legalistic behaviour. Their leitmotiv is: "Not only, but even. . . ." Even anger, the lustful look, the "legal" divorce, retaliation that keeps within the limits prescribed, and love that excludes the enemy are against God's will.

These extreme demands are meant not so much to be paradoxically overdemanding as, rather, liberating. Firstly, they are formulated in a way that everyone can understand. They include numerous references to the natural, unperverted practices of people in their daily lives. Secondly, the demands do not describe an unattainable distant goal, which all human action must of necessity fail to meet. Rather, Jesus pointed again and again to what the heavenly Father has done, does, and will do with his children and to God's possibilities, which are unlimited, whereas a person might despair of his or her own limited possibilities and impotence (Mark 10:27). Jesus' sayings about faith (Mark 9:23ff.), prayer (Luke 11:1ff.; Matt. 6:1ff.), or worry (Matt. 6:25ff.) are examples of this. Wherever Jesus calls on people to decide for themselves for God, he bases the argument on the fact that God has already decided for humanity. The unlimited readiness to forgive that he calls for also has its motivation in the limitless mercy of God, which he demonstrates toward the guilty in unfathomable measure (Matt. 18:23ff.). Jesus draws his hearers into this relation to God and, therefore, does not engage in abstract reflections about whether his demands are capable of fulfillment. In this way, what a person loses is the characteristic of being able to attain meritorious achievements (Matt. 20:1ff.). On the other hand, Jesus certainly did not give up the thought of "reward." The reward, however, is not a material prize, although images of this kind are not lacking, but the confirmation and perfection of the relation to God (Matt. 25:14ff.). The idea that human beings could claim and charge payment from God is for Jesus completely excluded (Luke 17:10).

The nearness of God, the real God, also brings humanity, no longer graded and classified in traditional categories, into urgent and imperious proximity. How much Jesus was concerned with human beings is shown especially by his commandment of love, which he not only taught but also practiced to the point of offensiveness. In it is concentrated the "better righteousness" that he demands of his disciples (Matt. 5:20). Jesus has taken over the Old Testament dual commandment of love of God and one's neighbour (Deut. 6:5; Lev. 19:18), which is also in Judaism a summary of the whole Law. But it is characteristic of Jesus' preaching (1) that he consistently subordinated all other laws—*e.g.,* the Sabbath commandment—to this highest critical standard (*e.g.,* Mark 2:27; 3:4), and (2) that he extended and heightened love of one's neighbour to love of one's enemies (Luke 6:27ff.), and (3) that his commandment does not have the abstract ideal of a general philanthropy at its root. Rather, he directed his hearers into the situations—always eventful and concrete—where they encounter their enemy (Matt. 5:38ff.) and their fellows in need (Luke 10:25ff.). Behaviour toward one's fellow is so important for Jesus that it is all that is spoken of in many of his utterances, without any mention of the first commandments of the Decalogue concerning behaviour toward God (*e.g.,* Matt. 5:25ff.; 7:12; 19:16ff.).

Individual and social ethics

The distinction that modern moral philosophy makes between individual and social ethics has, in respect to Jesus' teaching, only limited application. To be sure, Jesus did not draw up a program for a new order for the world and the nations, he did not demand a more just distribution of property, did not fight against the differences existing between masters, slaves, and hired workers, and did not give any directives for a better administration of justice. The world he had before his eyes was the world as it was, within the horizon of Palestinian Jewish rural conditions, and not the world as it ought to be. His sayings, parables, and illustrations show how keenly he assessed everyday life and how clearly he described it in his graphic, vigorous way—not glorifying this world as an eternally valid, divinely willed order, and also not getting morally indignant about it. Rather, he calls on people to behave in this given world in conformity to the original will of God and his dawning Kingdom; *e.g.,* to renounce the reign of mam-mon (Matt. 6:24; Luke 16:9ff.). Jesus did not, however, require a complete surrender of property from everyone. His followers were not to avail themselves of the legally regulated facilities for asserting one's own rights and were not to conform to the ways of customary behaviour in the world. The assertion that the world cannot be governed with the Sermon on the Mount is thus not to be denied. Jesus' sayings about retaliation and his commandment of love are not juristically practicable as they stand, because they can only serve as a guide for the one who has been wronged by someone else or who is required to divide his possessions with another person. Legislators and judges have to decide exclusively about the rights of others and must restrain evil for the sake of the general social order. But the truism about the impracticability of the Sermon on the Mount conceals the fact that Jesus' teaching contains strong impulses toward social criticism.

Jesus unmasks as hollow conventions many ostensibly valid standards, explaining the Law according to the norm of the commandment of love and applying it to concrete situations. For this reason he also resists egocentricity, not only of individuals but of entire religiously and socially privileged groups, and joins with discriminated-against people (*e.g.,* heathens, Samaritans, tax collectors, and harlots). Thus, Jesus calls on people to live a life that corresponds to the proximity of God's Kingdom, although the validity and urgency of his commandments require no apocalyptic basis. The act of their proclamation, however, stands nonetheless close to Jesus' own mission (Luke 11:32ff.). Whether, and in what way, he expressed the fact of his mission by the use of Christological titles is not thereby decided. Jesus knew that he had been sent to the "lost sheep of the house of Israel" (Matt. 15:24; 9:36). His ministry, seen as a whole, was confined to the sphere of his own people. Only a few significant words and scenes point forward to the inclusion of non-Jews, in a new, eschatological people of God (Matt. 8:11ff.). Jesus, however, did not organize a mission to the heathen (Matt. 10:5ff.) nor a worldwide "church." The only saying of this kind, spoken to Peter (Matt. 16:17ff.), has been placed in the mouth of the earthly Jesus by the later church and clearly reflects its situation, doctrine, and discipline. But Jesus certainly did call into existence a movement in Galilee and allowed at least the narrower circle of his disciples, if not all of his followers, to share in his wanderings and ministry. Later tradition first identified the latter group alone with the Apostles (authorized emissaries), the circle of whom was, however, not originally restricted to that group (*cf.* I Cor. 15:5ff.). The number 12 symbolizes the 12 tribes of Israel. If Jesus appointed these disciples himself, he thereby demonstrated his eschatological claim on the whole of Israel. According to the saying in Matt. 19:28 and Luke 22:30, which was probably not formulated until later, he conferred on them the office of ruling and judging the perfected Israel of the new aeon.

THE SUFFERINGS AND DEATH OF JESUS IN JERUSALEM

Jesus' decision to go to Jerusalem

Jesus' decision to go to Jerusalem is the turning point in his story. The events it set in motion soon came to have decisive significance for the faith of his followers. It is not coincidental that the Gospels narrate this period of his life in disproportionate breadth. Despite the many points of agreement among the Gospels, there also are considerable discrepancies within the tradition of the Passion. Thus, one cannot expect the tradition of the Passion to provide historically accurate reports, for it has been formed from the viewpoint of the church and its faith in Christ. The most important theological motifs in the narratives include the intention of presenting Jesus' sufferings and death as the fulfillment of God's will, the decision, in conformity with the words of the Old Testament Prophets and Psalms, to proclaim him as Messiah and Son of God, despite his brutal end. Nevertheless, important historical facts may be inferred from the texts.

Jesus probably went to Jerusalem with his disciples for the Passover in order to call the people of Israel gathered there to a final decision in view of the dawning Kingdom of God. He must have been aware of the heavy conflicts with the Jewish rulers that lay ahead of him. The story of the cleansing of the Temple, in particular, shows that Jesus did not avoid these conflicts. The later tradition, stylizing the story, gives as Jesus' sole motive for going to Jerusalem his desire to die there and to rise again in accordance with the will of God (Mark 8:31; 9:31; 10:32ff.). The best clue for a reconstruction of the outward course of Jesus' Passion is given by his Crucifixion. It proves that he was condemned and executed under Roman law as a political rebel. All reports agree that he died on Friday (Mark 15:42; Matt. 27:62; Luke 23:54; John 19:31). They differ, however, in that, according to the Synoptics, this was the 15th of Nisan (March/April); *i.e.,* the first day of the Passover. But, according to John, it was the previous day; *i.e.,* the one on which the Passover lambs were slaughtered and on which the festival was begun in the evening (in accordance with the Jewish division of days) with a common meal. Thus, according to John, Jesus' last meal with the disciples was not itself a Passover meal but took place earlier. Each of these datings may be theologically motivated, whether it be that the Eucharist is to be represented as the Passover meal (Synoptics) or whether Jesus himself is to be shown as the true Passover lamb, who died at the hour when the lambs were slaughtered (John). Historically, the Johannine dating is to be preferred, and the 14th Nisan (April 7) is to be regarded as the day of Jesus' death. The question of the occasion for Jesus' execution and the role that the Jews played is thereby more difficult and more important.

The date of Jesus' death

The way the Gospels present the facts of the case, Jesus was actually condemned to death by the supreme Jewish tribunal (Mark 14:55ff.). Pilate, on the other hand, was convinced of Jesus' innocence and made vain attempts to release him but finally yielded to the Jews' pressure against his better judgment (Mark 15:22ff.). The historical reliability of this account has rightly been questioned. First, the Synoptic reports differ among themselves. According to Mark and Matthew, the Jewish supreme court had already gathered in the home of the High Priest after Jesus' arrest in the night of Holy Thursday to Friday and condemned him to death as a blasphemer at that point (Mark 14:64). Thereafter, they resolved to hand Jesus over to Pilate in a new session in the early morning (Mark 15:1). Luke knows of only one session and has the interrogation take place in the morning (Luke 22:66), but he says nothing about Jesus' condemnation (Luke 22:71). John deviates even more; here, only the high priests Annas and Caiaphas are involved in the interrogation of Jesus (John 18:13ff.). Secondly, with regard to all the Gospel accounts, the question arises, what earwitness can be supposed later to have given the disciples an exact report? Thirdly, the jurisdictional competency of the Jewish Sanhedrin is disputed. In the opinion of some scholars, the Jewish authorities were permitted to pronounce sentence of death and to carry it out by stoning in the case of serious religious offenses (blasphemy). In the opinion of others, though, this required the confirmation of the Roman procurator. Also, trials of this kind were not to be conducted during the period of the festival.

The strongest argument against the Synoptic presentation is, however, that it is styled throughout in a Christian, and not in a Jewish, way; *i.e.,* on the basis of scriptural proof and the Christian confession to the messiahship and divine Sonship of Jesus. The High Priest's question, "Are you the Christ, the Son of the Blessed?" (Mark 14:61), is unthinkable from the viewpoint of Jewish premises, because Son of God was not a Jewish title for the Messiah. Thus, the account reflects the controversies of the later church with the Judaism of its day.

There also is in the Gospels a tendency to exonerate Pilate at the Jews' expense. His behaviour, however, does not match the picture that nonbiblical sources have handed down about him. But everything speaks for Jesus' having been arrested as a troublemaker, informally interrogated, and handed over to Pilate as the leader of a political revolt by the pro-Roman priestly and Sadducean members of the Sanhedrin, who were dominant in Jerusalem society in those days. The cleansing of the Temple and a prophetic, apocalyptic saying of Jesus (John 2:19; *cf.* Mark 14:58; Acts 6:14) about the destruction of the Temple may thereby have played a role. It can hardly be assumed that each and all of the Pharisees, who were without political influence at that time, were involved in the plot. Nor are they mentioned as a separate group in the Passion narratives alongside the priests, elders, and scribes.

The other scenes in the Passion story do not need to be listed here separately. They relate more to the theological meaning of Jesus' Passion and are, to a large measure, formed in an edifying cultic manner, even though they refer to events that are certainly historical; *e.g.,* Judas' betrayal, Jesus' last meal with his disciples, and Peter's denial of Jesus. The traces of an eyewitness account are perhaps still recognizable at certain points (Mark 14:52; 15:21).

The last words of Jesus

The accounts differ in their presentation of Jesus' death, especially in their rendering of his last words. It is only in Mark and Matthew that Jesus dies crying out the prayer from Psalm 22: "My God, my God, why hast thou forsaken me?" The distinction between the repentant and the defiant thief is only found in Luke. Jesus' last words are given differently in Luke ("Father, into thy hands I commit my spirit!") and John ("It is finished"). Each of these accounts, as also the testimony of the Roman centurion ("Truly this man was the Son of God!"; Mark 15:39), gives expression to the significance of Jesus and his story.

THE STORY OF JESUS AND FAITH IN JESUS

Did Jesus' violent death render his mission and story meaningless? In other words, did he enter definitively into the past as a failure and thus in this sense become the "historical" Jesus? For Pilate and the Romans, as for Jesus' Jewish opponents, there was no longer any problem. The decision had been taken. But Jesus' disciples were faced with this pressing question all the more. Their hopes were bitterly disappointed (Luke 24:13ff.). According to the unanimous witness of the New Testament texts, they did not find the answer themselves but were given it soon after Jesus' death through the Easter (Resurrection) appearances of Jesus (I Cor. 15:3ff.; etc.) and the experience of his Presence in the Spirit. The faith of early Christianity, with all of its practical and theological manifestations, grew out of this. This faith was not the preserve of a few enthusiasts or the personal opinion of individual Apostles. Wherever there were early Christian witnesses and communities, they were all united in believing and acknowledging the risen Lord (I Cor. 15:11).

The forms and ideas in which this faith found expression were various. According to the oldest view, Jesus' Resurrection meant his exaltation to divine lordship and was not necessarily connected with the tradition of the finding of the empty tomb, as the Gospels variously relate it. The theory of the resurrected one's having walked on earth for 40 days and only subsequently ascending into heaven is found only in Acts (1:3). Thus, there exists an undeniable tension between the unequivocal nature of the Easter *message,* on the one hand, and the equivocal nature of the Easter *accounts* and the historical problems connected with them, on the other. But the phenomenon of the whole Gospel tradition, rightly understood, is an expression of the faith in the living Christ without which neither a single word or deed of Jesus nor his Passion would have been handed down at all. The New Testament tradition does not aim at preserving the memory of Jesus as a figure of the past and telling only who Jesus *was,* but it wants to proclaim who Jesus *is.*

It may seem surprising that the question of Jesus' awareness of being the Messiah has been scarcely discussed in this article, let alone been given a precise answer. Usually, decisive significance is assigned to this question. Many scholars believe that access to the historical Jesus is only to be gained through the fact that Jesus had such an awareness in association with particular titles, such as Son of God, Son of man, and Messiah. In just the same way, they believe that the rise of the early Christian faith can only be understood by the same means. In light of the fact that the Gospels portray Jesus as the Christ (Greek term for the Messiah) and that numerous other titles of a similar kind are also found in them, the importance of this question is not to be underestimated. But it must nevertheless be noted that the Gospels are interested in the fact that Jesus was and is the Messiah, but not in his "consciousness" and inner development in a modern sense. The stories of Jesus' Baptism, temptation, and Transfiguration, for example, are not reports of experiences. But the question of whether Jesus applied one or several of those titles to himself still needs to be examined carefully, for each of them implies thoughts and concepts that must be of considerable relevance for his preaching and ministry. On this question, the opinions of scholars diverge widely, but it is uncontested that Jesus related his mission and activity in a unique way to the dawn of the Kingdom of God. It is another—and doubtful—question, however, that he expressed this understanding of himself through any traditional title.

Jesus' use of messianic titles

Three observations are important for the discussion of these problems. First, in the incontestably authentic texts in the Synoptics, Jesus never makes his own status a special topic of his teaching or the recognition of his rank a condition of salvation. Second, it is not only presumed but—by means of a comparison of the parallel texts and their modification from one Gospel to the next—often capable of proof and, in other cases, requiring to be assumed that the faith of the later church has had a major influence on the formation of the Christological texts. A third observation is also not without its importance. Wherever in such texts Jesus talks about the Messiah and the Son of David, the Son of man, the Son of God, and the Lord, there is never any indication that he is using these titles in a completely new sense. The meaning they have, however, is no longer congruent with the ideas that Jesus' contemporary hearers must have connected with the titles, to the extent that they were not completely unknown to them. Because the historical Jesus indubitably wanted to be understood, the critical question necessarily arises about texts of this kind reflecting the views of the later church and its environment.

Views about traditional messianic titles

Some of the traditional titles could not possibly have been used by Jesus with reference to himself. In those days his hearers could only have understood "Messiah" or "Son of David" in the political or national sense, which conflicted with Jesus' intentions. Also, the exclusive title "Son of God" must have been incomprehensible to the Jews of Palestine, although not to the later hearers of the Christian missionary preaching in the non-Jewish, Hellenistic world. The same applies to the expression "the (divine) Lord," which for Jews was reserved for God alone. Some scholars believe that Jesus understood himself to be the suffering servant of God, of whom Isaiah 53 speaks. But in the Gospels reference is hardly made to this important chapter. The sole passage (Mark 10:45) does not, at least in the form handed down, reproduce an authentic saying of the Lord but contains an interpretation of Jesus' death that goes back to the Greek-speaking Jewish Christian Church.

Thus, the problem is narrowed down to the question about Jesus' calling himself the "Son of man." This concept, which is frequently found in his statements about himself, is a title of sovereignty. It stems from Jewish apocalypticism and means not a normal human being but, rather, the mythical figure of the Judge of all the world, who will come on the clouds of heaven at the end of the days (Dan. 7:13ff.; etc.). An early group of Jesus' sayings (Mark 13:26; 14:62; Matt. 24:27) speaks of the Son of man in this eschatological, future sense and always in the third person, yet in some texts in such a way that Jesus does not explicitly identify himself with this Son of man (Mark 8:38; Luke 9:26; 12:8ff.). Two other groups of sayings speak of him quite differently. One speaks exclusively of his suffering, dying, and rising again in accordance with the will of God (Mark 8:31; 9:31; 10:33ff.); the other, of his authoritative work and wanderings on earth (Mark 2:10–28; Matt. 8:20; 11:19), both without a view of the Last Judgment. No Jewish hearer of Jesus could have recognized the apocalyptic Son of man in these sayings. Loosed from the ideas traditionally linked with it, the concept has here been given completely new contents in a retrospective view of Jesus' ministry and end. Thus, both of these groups of sayings are only to be understood from the point of view of the later church. Therefore, only some sayings of Jesus of the first group probably come into question as authentic ones. If Jesus spoke of the coming Son of man, those sayings prove he was speaking in the apocalyptic language and concepts of his day in order to express the promise that his disciples' loyalty to him will be recognized and confirmed in the Last Judgment. The relation of his earthly person to the figure of the coming Judge is not thereby made the subject of reflection. The Jesus tradition has gone through a process of modification, and the faith of the later church has made a major contribution to the formation of the tradition, whatever its precise extent may be.

The post-Easter message of salvation

In the post-Easter message of salvation, the eschatological here and now belongs inseparably to Jesus' message of the Kingdom of God and was being realized in him. In the face of unbelief and doubt, the Gospels do not just offer an account of the history of Jesus as it transpired, but they interpret it as God's history with the world, as the decisive, redemptive, and ultimate act and word of God for the world. All titles of sovereignty that faith has assigned to Jesus express the fact that in him the turning point of the ages, the inauguration of salvation, and the nearness and presence of God have arrived. The special character of the Gospel tradition should therefore be understood in this sense. This tradition has not replaced the historical Jesus with a mythical Christ but has made explicit the Christology that was secretly implicit in Jesus' words, works, and way, although without titles of sovereignty and supernatural traits. The question appropriate to the Gospel tradition would, therefore, not be about what has happened to Jesus of Nazareth in the course of the development but, rather, why the first Christians held fast to him. To ask in this way and to accept the answer of the Gospels are matters for faith. It goes beyond the limits of historical research.

(G.Bor.)

The picture of Christ in the early church: The Apostles' Creed

Even before the Gospels were written, Christians were reflecting upon the meaning of what Jesus had been and what he had said and done. It is a mistake, therefore, to suppose that such reflection is a later accretion upon the simple message of the Gospels. On the contrary, the early Christian communities were engaged in witness and worship from the very beginning. The forms of that witness and worship were also the forms of the narratives in the Gospel accounts. From this fact it follows that to understand the Gospel accounts regarding Jesus we must consider the faith of the early church regarding Christ. In this sense it is valid to maintain that there is no distinction between "the Jesus of history" and "the Christ of faith," and that the only way to get at the former is by the latter. Christology, the doctrine about Christ, is then as old as Christianity itself.

To comprehend the faith of the early church regarding Christ, we must turn to the writings of the New Testament, where that faith found embodiment. It was also embodied in brief confessions or creeds, but these have not been preserved for us complete in their original form. What we have are fragments of those confessions or creeds in various books of the New Testament, snatches from them in other early Christian documents, and later forms of them in Christian theology and liturgy. The so-called Apostles' Creed is one such later form. It did not achieve its present form until quite late; just how late is a matter of controversy. But in its earliest ancestry it is very early indeed, perhaps dating back to the 1st century. And its confession regarding Christ is probably the earliest core, around which later elaborations of it were composed. Allowing for such later elaboration, we may say that in the Apostles' Creed we have a convenient summary of what the early church believed about Christ amid all the variety of its expression and formulation. The creeds were a way for Christians to explain what they meant by their acts of worship. When they put "I believe" or "We believe" at the head of what they confessed about God and Christ, they meant that their declarations rested upon faith, not merely upon observation.

PREEXISTENCE

The stages of Christ's career

The statement "I believe" also indicated that Christ was deserving of worship and faith, and that he was therefore on a level with God. At an early date, possibly as early as the words of Paul in Phil. 2:6–11, Christian theology began to distinguish three stages in the career of Jesus Christ: his preexistence with the Father before all things; his Incarnation and humiliation in "the days of His flesh" (Heb. 5:7), and his glorification, beginning with the Resurrection and continuing forever.

Probably the most celebrated statement of the preexistence of Christ is the opening verses of the Gospel of St. John. Here Christ is identified as the incarnation of the Word (Logos) through which God made all things in the beginning, a Word existing in relation to God before the creation. The sources of this doctrine have been sought in Greek philosophy, both early and late, as well as in the Jewish thought of Philo and of the Palestinian rabbis. Whatever its source, the doctrine of the Logos in John is distinctive by virtue of the fact that it identifies the Logos with a specific historical person. Other writings of the New Testament also illustrate the faith of the early Christians regarding the preexistence of Christ. The opening chapters of both Colossians and Hebrews speak of Christ as the preexistent one through whom all things were created, therefore as distinct from the created order of things in both time and preeminence; the preposition "before" in Col. 1:17 apparently refers both to his temporal priority and to his superior dignity. Yet before any theological reflection about the nature of this preexistence had been able to find terms and concepts, the early Christians were worshipping Christ as divine. Phil. 2:6-11 may be a quotation from a hymn used in such worship. Theological reflection told them that if this worship was legitimate, he must have existed with the Father "before all ages."

Jesus Christ. By the time the text of the creed was established, this was the usual designation for the Saviour. Originally, of course, "Jesus" had been his given name, meaning "Yahweh saves," or "Yahweh will save" (see Matt. 1:21), while "Christ" was the Greek translation of the title "Messiah." Some passages of the New Testament still used "Christ" as a title (*e.g.,* Luke 24:26; II John 7), but it is evident from Paul's usage that the title became simply a proper name very early. Most of the Gentiles took it to be a proper name, and it was as "Christians" that the early believers were labelled (Acts 11:26). In the most precise language, the term "Jesus" was reserved for the earthly career of the Lord; but it seems from liturgical sources that it may actually have been endowed with greater solemnity than the name "Christ." Within a few years after the beginnings of the Christian movement, Jesus, Christ, Jesus Christ, and Christ Jesus could be used almost interchangeably, as the textual variants in the New Testament indicate. Only in modern times has it become customary to distinguish sharply among them for the sake of drawing a line between the Jesus of history and the Christ of faith, and this only in certain circles. The theologians and people of many churches still use phrases like "the life of Christ," because "Christ" is primarily a name. It is difficult to imagine how it could be otherwise when the Old Testament implications of the title have become a secondary consideration in its use—a process already evident within the New Testament.

God's only son. The declaration that Jesus Christ is the son of God is one of the most universal in the New Testament, most of whose books refer to him that way. The Gospels do not quote him as using the title for himself in so many words, although sayings like Matt. 11:27 come close to it. There are some instances where the usage of the Gospels appears to echo the more general implications of divine sonship in the Old Testament as a prerogative of Israel or of the true believer. Usually, however, it is evident that the evangelists, like Paul, meant some special honour by the name. The evangelists associated the honour with the story of Jesus' baptism (Matt. 3:17) and transfiguration (Matt. 17:5), Paul with the faith in the Resurrection (Rom. 1:4). From this association some have argued that "Son of God" in the New Testament never referred to the preexistence of Christ. But it is clear in John and in

Paul that this implication was not absent, even though it was not as prominent as it became soon thereafter. What made the implication of preexistence more prominent in later Christian use of the term "Son of God" was the clarification of the doctrine of the Trinity, where "Son" was the name for the eternal Second Person (Matt. 28:19). As the Gospels show, the application of the name "Son of God" to Jesus was offensive to the Jews, probably because it seemed to smack of gentile polytheism. This also made it all too intelligible to the pagans, as early heresies indicate. Facing both the Jews and the Greeks, the apostolic church confessed that Jesus Christ was "God's only Son": the Son of God, in antithesis to Jewish claims that the eternal could have no sons; the only Son, in antithesis to Greek myths of divine procreation.

Kyrios

The Lord. As passages like Rom. 1:4, show, the phrase "Jesus Christ our Lord" was one of the ways the apostolic church expressed its understanding of what he had been and done. Luke even put the title into the mouth of the Christmas angel (Luke 2:11). From the way the name "Lord" (*Kyrios*) was employed during the 1st century it is possible to see several implications in the Christian use of it for Christ. The Christians meant that there were not many divine and lordly beings in the universe, but only one *Kyrios* (I Cor. 8:5–6). They meant that the Roman Caesar was not the lord of all, as he was styled by his worshippers, but that only Christ was Lord (Rev. 17:14). And they meant that Yahweh, the covenant God of the Old Testament, whose name they pronounced as "Lord," had come in Jesus Christ to establish the new covenant (see Rom. 10:12–13). Like "Son of God," therefore, the name *Kyrios* was directed against both parts of the audience to which the primitive church addressed its proclamation. At times it stood particularly for the risen and glorified Christ, as in Acts 2:36; but in passages that echoed the Old Testament it was sometimes the preexistence that was being primarily emphasized (Matt. 22:44). Gradually "our Lord," like "Christ," became a common way of speaking about Jesus Christ, even when the speaker did not intend to stress his lordship over the world.

INCARNATION AND HUMILIATION

Conceived by the Holy Spirit, born of the Virgin Mary. Earlier forms of the creed seem to have read: "Born of the Holy Spirit and of the Virgin Mary." The primary affirmation of this article is that the Son of God, the Word, had become man or, as John's Gospel put it, "flesh" (John 1:14). Preexistence and Incarnation presuppose each other in the Christian view of Jesus Christ. Hence the New Testament assumed his preexistence when it talked about his becoming man; and when it spoke of him as preexistent, it was ascribing this preexistence to him whom it was describing in the flesh. It may be that the reference in the creed to the Virgin Mary was intended to stress primarily her function as the guarantee of Christ's true humanity, but the creed also intended to teach the supernatural origin of that humanity. Although it is true that neither Paul nor John makes reference to it, the teaching about the virginal conception of Jesus, apparently based upon Isa. 7:14, was sufficiently widespread in the 1st century to warrant inclusion in both Matthew and Luke, as well as in creeds that date back to the 1st century. As it stands, the creedal statement is a paraphrase of Luke 1:35. In the New Testament the Holy Spirit was also involved in the baptism and the Resurrection of Jesus.

Suffered under Pontius Pilate, was crucified, dead, and buried. To a reader of the Gospels, the most striking feature of the creed is probably its omission of that which occupied a major part of the Gospels, the story of Jesus' life and teachings. In this respect there is a direct parallel between the creed and the Epistles of the New Testament, especially those of Paul. Judging by the amount of space they devoted to the Passion story, even the writers of the Gospels were apparently more interested in these few days of Jesus' life than they were in anything else he had said or done. The reason for this was the faith underlying both the New Testament and the creed, that the events of Jesus' Passion, death, and Resurrection were the events by which God had accomplished the salvation of human beings. The Gospels found their climax in those events, and the other material in them led up to those events. The Epistles applied those events to concrete situations in the early church. From the way Paul could speak of the Cross (Phil. 2:6–11) and of "the night when he [Jesus] was betrayed" (I Cor. 11:23), it seems that before our Gospels came into existence the church commemorated the happenings associated with what came to be called Holy Week. Some of the earliest Christian art was a portrayal of these happenings, another indication of their importance in the cultic and devotional life of early Christianity. How did the Cross effect the salvation of human beings? The answers of the New Testament and the early church to this question involved a variety of metaphors: Christ offered himself as a sacrifice to God; his life was a ransom for many; his death made mankind alive; his suffering was an example to people when they must suffer; he was the Second Adam, creating a new humanity; his death shows people how much God loves them; and others. Every major atonement theory of Christian theological history discussed below was anticipated by one or another of these metaphors. The New Testament employed them all to symbolize something that could be described only symbolically, that "God was in Christ reconciling the world to himself, not counting their trespasses against them" (II Cor. 5:19).

The happenings of Holy Week

He descended into hell. This phrase was probably the last to be added to the creed. Its principal source in the New Testament was the description in I Pet. 3:18–20 of Christ's preaching to the spirits in prison. Originally the descent into hell may have been identified with the death of Christ, when he entered the abode of the dead in the underworld. But in the time before it entered the creed, the descent was frequently taken to mean that Christ had gone to rescue the souls of the Old Testament faithful from the underworld, from what western Catholic theology eventually called the *limbo patrum.* Among some of the Church Fathers the descent into hell had come to mean Christ's declaration of his triumph over the powers of hell. Despite its subsequent growth in importance, however, the doctrine of the descent into hell apparently did not form an integral part of the apostolic preaching about Christ.

GLORIFICATION

The third day he rose again from the dead. The writers of the New Testament nowhere made the Resurrection of Christ a matter for argument, but everywhere asserted it and assumed it. With it began that state in the history of Jesus Christ that was still continuing, his elevation to glory. They used it as a basis for three kinds of affirmations. The Resurrection of Christ was the way God bore witness to his son, "designated Son of God in power according to the Spirit of holiness by his resurrection from the dead" (Rom. 1:4); this theme was prominent also in the Book of Acts. The Resurrection was also the basis for the Christian hope of life after death (I Thess. 4:14), and without it that hope was said to be baseless (I Cor. 15:12–20). The Resurrection of Christ was also the ground for admonitions to manifest a "newness of life" (Rom. 6:4) and to "seek the things that are above" (Col. 3:1). The writers of the New Testament themselves expressed no doubt that the Resurrection had really happened. But Paul's discussion in I Cor. 15 shows that among those who heard the Christian message there was such doubt, as well as efforts to rationalize the Resurrection. The differences among the Gospels, and between the Gospels and Paul, suggest that from the outset a variety of traditions existed regarding the details of the Resurrection. But such differences only serve to emphasize how universal the faith in the Resurrection was amid this variety of traditions.

The Resurrection

He ascended into heaven, and sitteth on the right hand of God the father almighty. As indicated earlier, the narrative of the Ascension is peculiar to Luke-Acts, but other parts of the New Testament may refer to it. Eph. 4:8–10 may be such a reference, but many interpreters hold that for Paul Resurrection was identical with Ascension. That, they maintain, is why he could speak of the appearance of the risen Christ to him in continuity with the appearances to others (I Cor. 15:5–8) despite the fact that, in the chronology of the creed, the Ascension intervened between them. Session at the right hand of the Father was apparently a Christian interpretation of Ps. 110:1. It implied the elevation—or, as the doctrine of preexistence became clearer, the restoration—of Christ to a position of honour with God. Taken together, the Ascension and the session were a way of speaking about the presence of Christ with the Father during the interim between the Resurrection and the Second Advent. From Eph. 4:8–16, it is evident that this way of speaking was by no means inconsistent with another Christian tenet, the belief that Christ was still present in and with his church. It was, in fact, the only way to state that tenet in harmony with the doctrine of the Resurrection.

The Second Advent

From thence he shall come to judge the quick and the dead. The creed concludes its Christological section with the doctrine of the Second Advent: the First Advent was a coming into the flesh, the Second Advent a coming in glory. Much controversy among modern scholars has been occasioned by the role of this doctrine in the early church. Those who maintain that Jesus erroneously expected the early end of the world have often interpreted Paul as the first of those who began the adjustment to a delay in the end, with John's Gospel as a more advanced stage of that adjustment. Those who hold that the imminence of the end was a continuing aspect of human history as Jesus saw it also maintain that this phrase of the creed was a statement of that imminence, without any timetable necessarily implied. From the New Testament it seems that both the hope of the Second Coming and a faith in the continuing presence of Christ belonged to the outlook of the apostolic church, and that seems to be what the creed meant. The phrase "the quick and the dead" is a summary of passages like I Cor. 15:51–52 and I Thess. 4:15–17.

In order to complete the confession of the creed regarding the glorification of Christ, the Nicene Creed added the phrase: "Of His kingdom there shall be no end." This was a declaration that Christ's return as judge would usher in the full exercise of his reign over the world. Such was the expectation of the apostolic church, based upon what it knew and believed about Jesus Christ.

The dogma of Christ in the ancient councils

The main lines of orthodox Christian teaching about the person of Christ were set by the New Testament and the ancient creeds. But what was present there in a germinal form became a clear statement of Christian doctrine when it was formulated as dogma. In one way or another, the first four ecumenical councils were all concerned with the formulation of the dogma regarding the person of Christ—his relation to the Father, and the relation of the divine and the human in him.

Such a formulation became necessary because teachings arose within the Christian community that seemed to threaten what the church believed and confessed about Christ. Both the dogma and the heretical teachings against which the dogma was directed are therefore part of the history of Jesus Christ.

THE COUNCILS OF NICAEA AND CONSTANTINOPLE

Early heresies. From the outset, Christianity has had to contend with those who misinterpreted the person and mission of Jesus. Both the New Testament and the early confessions of the church referred and replied to such misinterpretations. As the Christian movement gained adherents from the non-Jewish world, it had to explain Christ in the face of new challenges.

These misinterpretations touched both the question of his humanity and the matter of his deity. A concern to safeguard the true humanity of Jesus led some early Christians to teach that Jesus of Nazareth, an ordinary man, was adopted as the Son of God in the moment of his baptism or after his Resurrection; this heresy was called adoptionism. Gnostics and others wanted to protect him against involvement in the world of matter, which they regarded as essentially evil, and therefore taught that he had only an apparent, not a real body; they were called docetists. Most of the struggle over the person of Christ, however, dealt with the question of his relation to the Father. Some early views were so intent upon asserting his identity with the Father that the distinction of his person was lost and he became merely a manifestation of the one God. Because of this idea of Christ as a "mode" of divine self-manifestation, proponents of this view were dubbed "modalists"; from an early supporter of the view it was called "Sabellianism." Other interpretations of the person of Christ in relation to God went to the opposite extreme. They insisted so strenuously upon the distinctness of his person from that of the Father that they subordinated him to the Father. Many early exponents of the doctrine of the Logos were also subordinationists, so that the Logos idea itself became suspect in some quarters. What was needed was a framework of concepts with which to articulate the doctrine of Christ's oneness with the Father and yet distinctness from the Father, and thus to answer the question (Adolf von Harnack): "Is the Divinity which has appeared on earth and reunited men with god identical with that supreme Divinity which governs heaven and earth, or is it a demigod?"

Nicaea. That question forced itself upon the church through the teachings of Arius. He maintained that the Logos was the first of the creatures, called into being by God as the agent or instrument through which he was to make all things. Christ was thus less than God, but more than man; he was divine, but he was not God. To meet the challenge of Arianism, which threatened to split the church, the newly converted emperor Constantine convoked in 325 the first ecumenical council of the Christian church at Nicaea. The private opinions of the attending bishops were anything but unanimous, but the opinion that carried the day was that espoused by the young presbyter Athanasius, who later became bishop of Alexandria. The Council of Nicaea determined that Christ was "begotten, not made," that he was therefore not creature but creator. It also asserted that he was "of the same essence as the Father" (*homoousios to patri*). In this way it made clear its basic opposition to subordinationism, even though there could be, and were, quarrels about details. It was not equally clear how the position of Nicaea and of Athanasius differed from modalism. Athanasius asserted that it was not the Father nor the Holy Spirit, but only the Son that became incarnate as Jesus Christ. But in order to assert this, he needed a more adequate terminology concerning the persons in the Holy Trinity. So the settlement at Nicaea regarding the person of Christ made necessary a fuller clarification of the doctrine of the Trinity, and that clarification in turn made possible a fuller statement of the doctrine of the person of Christ.

The first ecumenical council

Constantinople. Nicaea did not put an end to the controversies but only gave the parties a new rallying point. Doctrinal debate was complicated by the rivalry among bishops and theologians as well as by the intrusion of imperial politics that had begun at Nicaea. Out of the post-Nicene controversies came that fuller statement of the doctrine of the Trinity which was needed to protect the Nicene formula against the charge of failing to distinguish adequately between the Father and the Son. Ratified at the Council of Constantinople in 381, but since lost, that statement apparently made official the terminology developed by the supporters of Nicene orthodoxy in the middle of the 4th century: one divine essence, three divine persons (*mia ousia, treis hypostaseis*). The three persons, Father, Son, and Holy Spirit, were distinct from one another but were equal in their eternity and power. Now it was possible to teach, as Nicaea had, that Christ was "of the same essence as the Father" without arousing the suspicion of modalism. Although this doctrine seemed to make problematical the unity of God, it did provide an answer to the first of the two issues confronted by the church in its doctrine of the person of Christ—the issue of Christ's relation to the Father. It now became necessary to clarify the second issue—the relation of the divine and the human within Christ.

THE COUNCILS OF EPHESUS AND CHALCEDON

The Trinity

By excluding several extreme positions from the circle of orthodoxy, the formulation of the doctrine of the Trinity in the 4th century determined the course of subsequent discussion about the person of Christ. It also provided the terminology for that discussion, since 5th-century theologians were able to describe the relation between the divine and the human Christ by analogy to the relation between the Father and the Son in the Trinity. The term that was found to express this relation in Christ was "nature," *physis.* There were three divine persons in one divine essence; such was the outcome of the controversies in the 4th century. But there were also two natures, one of them divine and the other human, in the one person Jesus Christ. Over the relation between these two natures the theologians of the 5th century carried on their controversy.

The abstract questions with which they sometimes dealt in that controversy, some of them almost unintelligible to a modern mind, must not be permitted to obscure the fact that a basic issue of the Christian faith was at stake: how can Jesus Christ be said to partake of both identity with God and brotherhood with humanity?

The parties. During the half century after the Council of Constantinople several major points of emphasis developed in the doctrine of the person of Christ; characteristically, these are usually defined by the episcopal see that espoused them. There was a way of talking about Christ that was characteristic of the see at Alexandria. It stressed the divine character of all that Jesus Christ had been and done, but its enemies accused it of absorbing the humanity of Christ in his divinity. The mode of thought and language employed at Antioch, on the other hand, emphasized the true humanity of Christ; but its opponents maintained that in so doing it had split Christ into two persons, each of whom maintained his individual selfhood while they acted in concert with each other. Western theology was not as abstract as either of these alternatives. Its central emphasis was a practical concern for human salvation and for as irenic a settlement of the conflict as was possible without sacrificing that concern. Even more than in the 4th century, considerations of imperial politics were always involved in conciliar actions, together with the fear in countries like Egypt that Constantinople might come to dominate them. Thus a decision regarding the relation between the divine and the human in Christ could be simultaneously a decision regarding the political situation. Nevertheless, the settlements at which the councils of the 5th century arrived may be and are regarded as normative in the church long after their political setting has disappeared.

The conflict between Alexandria and Antioch came to a head when Nestorius, taking exception to the use of the title "Mother of God" or, more literally, "God-Bearer" (*Theotokos*) for the Virgin Mary, insisted that she was only "Christ-Bearer." In this insistence the Antiochian emphasis upon the distinction between the two natures in Christ made itself heard throughout the church. The Alexandrian theologians responded by charging that Nestorius was dividing the person of Christ, which they represented as so completely united that, in the famous phrase of Cyril, there was "one nature of the Logos which became incarnate." By this he meant that there was only one nature, the divine, before the Incarnation, but that after the Incarnation there were two natures indissolubly joined in one person; Christ's human nature had never had an independent existence. There were times when Cyril appeared to be saying that there was "one nature of the incarnate Logos" even after the Incarnation, but his most precise formulations avoided this language.

The Council of Ephesus in 431 was one in a series of gatherings called to settle this conflict, some by one party and some by the other. The actual settlement was not accomplished, however, until the calling of the Council of Chalcedon in 451.

The settlement at Chalcedon. The basis of the settlement was the Western understanding of the two natures in Christ, as formulated in the *Tome* of Pope Leo I of Rome. Chalcedon declared: "We all unanimously teach . . . one and the same Son, our Lord Jesus Christ, perfect in deity and perfect in humanity . . . in two natures, without being mixed, transmuted, divided, or separated. The distinction between the natures is by no means done away with through the union, but rather the identity of each nature is preserved and concurs into one person and being." In this formula the valid emphases of both Alexandria and Antioch came to expression; both the unity of the person and the distinctness of the natures were affirmed. Therefore the decision of the Council of Chalcedon has been the basic statement of the doctrine of the person of Christ for most of the church ever since. The western part of the church went on to give further attention to the doctrine of the work of Christ. In the eastern part of the church the Alexandrians and the Antiochians continued the controversies that had preceded Chalcedon, but they clashed now over the question of how to interpret Chalcedon. The controversy over the Monophysite and the Monothelite heresies was an effort to clarify the interpretation of Chalcedon, with the result that the extremes of the Alexandrian position were condemned just as the Nestorian extreme of the Antiochian had been.

The two natures in Christ

Emerging from all this theological discussion was an interpretation of the person of Christ that affirmed both his oneness with God and his oneness with humanity while still maintaining the oneness of his person. Interestingly, the liturgies of the church had maintained this interpretation at a time when the theologians of the church were still struggling for clarity; and the final solution was a scientifically precise restatement of what had been present germinally in the liturgical piety of the church. In the formula of Chalcedon that solution finally found the framework of concepts and of vocabulary that it needed to become intellectually consistent. In one sense, therefore, what Chalcedon formulated was what Christians had been believing from the beginning; but in another sense it represented a development from the earlier stages of Christian thought.

The interpretation of Christ in Western faith and thought

With the determination of the orthodox teaching of the church regarding the person of Christ, it still remained necessary to clarify the doctrine of the work of Christ. While it had been principally in the East that the discussion of the former question was carried on—though with important additions from the West, as we have seen—it was the Western Church that provided the most detailed answers to the question: granted that this is what Jesus Christ was, how are we to describe what it is that he did?

DOCTRINES OF THE PERSON AND WORK OF CHRIST

The medieval development. The most representative spokesman of the Western Church on this question, as on most others, was St. Augustine. His deep understanding of the meaning of human sin was matched by his detailed attention to the meaning of divine grace. Central to that attention was his emphasis upon the humanity of Jesus Christ as man's assurance of his salvation, an emphasis to which he gave voice in a variety of ways. The humanity of Christ showed how God elevated the humble; it was the link between the physical nature of human beings and the spiritual nature of God; it was the sacrifice which the human race offered to God; it was the foundation of a new humanity, recreated in Christ as the old humanity had been created in Adam—in these and other ways Augustine sought to describe the importance of the Incarnation for the redemption of man. By combining this stress upon the humanity of Christ as the Saviour with a doctrine of the Trinity that was orthodox but nevertheless highly creative and original, St. Augustine put his mark indelibly upon Western piety and theology, which, in Anselm and in the reformers, sought further for adequate language in which to describe God's deed of reconciliation in Jesus Christ.

St. Augustine

During the formative centuries of Christian dogma, there had been many ways of describing that reconciliation, most of them having some precedent in biblical speech. One of the most prominent pictures of the reconciliation was that connected with the biblical metaphor of ransom: Satan held the human race captive in its sin and corruptibility, and the death of Christ was the ransom paid to the Devil as the price for setting mankind free. A related metaphor was that of the victory of Christ: Christ entered into mortal combat with Satan for the human race, and the winner was to be lord; although the Crucifixion appeared to be Christ's capitulation to the enemy, his Resurrection broke the power of the Devil and gave the victory to Christ, granting to mankind gift of immortality. From the Old Testament and the Epistle to the Hebrews came the image of Christ as the sacrificial victim who was offered up to God as a means of stilling the divine anger. From the sacrament of penance came the idea, most fully developed by St. Anselm, that the death of Christ was a vicarious satisfaction rendered for mankind. Like the New Testament, the Church Fathers could admonish their hearers to learn from the death of Christ how to suffer patiently. They could also point to the suffering and death of Christ as the supreme illustration of how much God loves mankind. As in the New Testament, therefore, so in the tradition of the church there were many figures of speech to represent the miracle of the reunion between man and God effected in the God-man Christ Jesus.

Common to all these figures of speech was the desire to do two things simultaneously: to emphasize that the reunion was an act of God, and to safeguard the participation of man in that act. Some theories were so "objective" in their emphasis upon the divine initiative that man seemed to be almost a pawn in the transaction between God in Christ and the Devil. Other theories so "subjectively" concentrated their attention upon man's involvement and man's response that the full scope of the redemption could vanish from sight. It was in Anselm of Canterbury that Western Christendom found a theologian who could bring together elements from many theories into one doctrine of the Atonement, summarized in his book, *Cur Deus homo?* According to this doctrine, sin was a violation of the honour of God. God offered man life if he rendered satisfaction for that violation; but the longer man lived, the worse the situation became. Only a life that was truly human and yet had infinite worth would have been enough to give such a satisfaction to the violated honour of God on behalf of the entire human race. Such a life was that of Jesus Christ, whom the mercy of God sent as a means of satisfying the justice of God. Because he was true man, his life and death could be valid for men; because he was true God, his life and death could be valid for *all* men. By accepting the fruits of his life and death, mankind could receive the benefits of his satisfaction. With some relatively minor alterations, Anselm's doctrine of Atonement passed over into the theology of the Latin church, forming the basis of both Roman Catholic and orthodox Protestant ideas of the work of Christ. It owed its acceptance to many factors, not the least of them being the way it squared with the liturgy and art of the West. The crucifix has become the traditional symbol of Christ in the Western Church, reinforcing and being reinforced by the satisfaction theory of the Atonement.

Scholastic theology

Scholastic theology, therefore, did not modify traditional ways of speaking about either the person or the work of Christ as sharply as it did, for example, some of the ways the Church Fathers had spoken about the presence of the body and blood of Christ in the Eucharist. The major contribution of the scholastic period to the Christian conception of Jesus Christ appears to lie in the way it managed to combine theological and mystical elements. Alongside the growth of Christological dogma and sometimes in apparent competition with it was the development of a view of Christ that sought personal union with him rather than accurate concepts about him. Such a view of Christ appeared occasionally in the writings of Augustine, but it was in men like Bernard of Clairvaux that it attained both its fullest expression and its most adequate harmonization with the dogmatic view. The relation between the divine and the human natures in Christ, as formulated in ancient dogma, provided the mystic with the ladder he needed to ascend through the man Jesus to the eternal Son of God, and through him to a mystical union with the Holy Trinity; this had been anticipated in the mystical theology of some of the Greek fathers. At the same time the dogma saved mysticism from the pantheistic excesses to which it might otherwise have gone; for the doctrine of the two natures meant that the humanity of the Lord was not an expendable element in Christian piety, mystical or not, but its indispensable presupposition and the continuing object of its adoration, in union with his deity. As a matter of fact, another contribution of the medieval development was the increased emphasis of St. Francis of Assisi and his followers upon the human life of Jesus. These brotherhoods cultivated a more practical and ethical version of mystical devotion, to be distinguished from speculative and contemplative mysticism. Their theme became the imitation of Christ in a life of humility and obedience. With it came a new appreciation of that true humanity of Christ which the dogma had indeed affirmed, but which theologians had been in danger of reducing to a mere dogmatic concept. As Henry Thode and others have suggested, this new appreciation is reflected in the way painters like Giotto began to portray Jesus, in contrast with their Western predecessors and especially with the stylized picture of Christ in Byzantine icon painting.

Luther, Calvin, and Zwingli

The Reformation and classical Protestantism. The attitude of the reformers toward traditional conceptions of the person and work of Christ was conservative. Insisting for both religious and political reasons that they were orthodox, they altered little in the Christological dogma. Martin Luther and John Calvin gave the dogma a new meaning when they related it to their doctrine of justification by grace through faith. Because of his interpretation of sin as the captivity of the will, Luther also revived the patristic metaphor of the Atonement as the victory of Christ; it is characteristic of him that he wrote hymns for both Christmas and Easter but not for Lent. The new attention to the Bible that came with the Reformation created interest in the earthly life of Jesus, while the Reformation idea of "grace alone" and of the sovereignty of God even in his grace made the deity of Christ a matter of continuing importance.

In the ideas about the Lord's Supper set forth by Huldrych Zwingli, Luther thought he saw a threat to the orthodox doctrine of Christ, and he denounced those doctrines vehemently. As this controversy progressed, Luther interpreted the ancient dogma of the two natures to mean that the omnipresence of the divine nature was communicated to the human nature of Christ, and that therefore Christ as both God and man was present everywhere and at all times. Although he repudiated both Luther's and Zwingli's theories, Calvin was persuaded that the ancient Christological dogma was true to the biblical witness and he permitted no deviation from it. All this is evidence for the significance that "Jesus Christ, true God begotten of the Father from eternity, and also true man, born of the Virgin Mary," to use Luther's formula, had in the faith and theology of all the reformers.

At one point the theology of the reformers did serve to bring together several facets of the biblical and the patristic descriptions of Jesus Christ. That was the doctrine of the threefold office of Christ, systematized by Calvin and developed more fully in Protestant orthodoxy: Christ as prophet, priest, and king. Each of these symbolized the fulfillment of the Old Testament and represented one aspect of the church's continuing life. Christ as prophet fulfilled and elevated the prophetic tradition of the Old Testament, while continuing to fulfill his prophetic office in the ministry of the Word. Christ as priest brought to an end the sacrificial system of the Old Testament by being both the priest and the victim, while he continues to function as intercessor with and for the church. Christ as king was the royal figure to whom the Old Testament had pointed, while exercising his rule among men now through those whom he has appointed. In each of the three, Protestants differed from one another according to their theological, ethical, or liturgical positions. But the threefold office enabled Protestant theology to take into account the complexity of the biblical and patristic pictures of Christ as no oversimplified theory was able to do, and it is probably the chief contribution of the reformers to the theological formulation of the doctrine of the person and work of Christ.

THE DEBATE OVER CHRISTOLOGY IN MODERN CHRISTIAN THOUGHT

Few Protestant theologians in the middle of the 20th century were willing to endorse the ancient dogma of the two natures in Christ as unconditionally as the reformers had done, for between the Reformation and modern theology there intervened a debate over Christology that altered the perspective of most Protestant denominations and theologians. By the 20th century there was a wider gap between the theology of the reformers and that of many modern Protestants than there had been between the theology of the reformers and that of their Roman Catholic opponents.

The Socinian criticism

Origins of the debate. The earliest criticism of orthodox dogma came in the age of the Reformation, not from the reformers but from the "left wing of the Reformation," from Michael Servetus (1511?–53) and the Socinians. This criticism was directed against the presence of nonbiblical concepts and terms in the dogma, and it was intent upon safeguarding the true humanity of Jesus as a moral example. There were many inconsistencies in this criticism, such as the willingness of Servetus to call Jesus "Son of God" and the Socinian custom of addressing prayer and worship to him. But it illustrates the tendency, which became more evident in the Enlightenment, to use the Reformation protest against Catholicism as a basis for a protest against orthodox dogma as well. While that tendency did not gain much support in the 16th century because of the orthodoxy of the reformers, later criticism of orthodox Christology was able to wield the "Protestant principle" against the dogma of the two natures on the grounds that this was a consistent application of what the reformers had done. Among the ranks of the Protestant laity, the hymnody and the catechetical instruction of the Protestant churches assured continuing support for the orthodox dogma. Indeed, the doctrine of Atonement by the vicarious satisfaction of Christ's death has seldom been expressed as amply as it was in the hymns and catechisms of both the Lutheran and the Reformed churches. During the period of Pietism in the Protestant churches, this loyalty to orthodox teaching was combined with a growing emphasis upon the humanity of Jesus, also expressed in the hymnody of the time.

When theologians began to criticize orthodox ideas of the person and work of Christ, therefore, they met with opposition from the common people. Albert Schweitzer dates the development of a critical attitude from the work of H.S. Reimarus (1694–1768), but Reimarus was representative of the way the Enlightenment treated the traditional view of Jesus. The books of the Bible were to be studied just as other books are, and the life of Jesus was to be drawn from them by critically sifting and weighing the evidence of the Gospels. The Enlightenment thus initiated the modern interest in the life of Jesus, with its detailed attention to the problem of the relative credibility of the Gospel records. It has been suggested by some historians that the principal target of Enlightenment criticism was not the dogma of the two natures but the doctrine of the vicarious Atonement. The leaders of Enlightenment thought did not make a sudden break with traditional ideas, but gave up belief in miracles, the Virgin Birth, the Resurrection, and the Second Advent only gradually. Their principal importance for the history of the doctrine of Christ consists in the fact that they made the historical study of the sources for the life of Jesus an indispensable element of any Christology.

Strauss and Renan

The 19th century. Although the Enlightenment of the 18th century was the beginning of the break with orthodox teachings about Jesus Christ, it was only in the 19th century that this break attracted wide support among theologians and scholars in many parts of Christendom—even, for a while, among the Modernists of the Roman Catholic Church. Two works of the 19th century were especially influential in their rejection of orthodox Christology. One was the *Life of Jesus,* first published in 1835 by David Friedrich Strauss; the other, bearing the same title, was first published by Ernest Renan in 1863. Strauss's work paid more attention to the growth of Christian ideas—he called them "myths"—about Jesus as the basis for the picture we have in the Gospels, while Renan attempted to account for Jesus' career by a study of his inner psychological life in relation to his environment. Both works achieved wide circulation and were translated into other languages, including English. They took up the Enlightenment contention that the sources for the life of Jesus were to be studied as other sources are, and what they constructed on the basis of the sources was a type of biography in the modern sense of the word. In addition to Strauss and Renan, the 19th century saw the publication of a plethora of books about the life and teachings of Jesus. Each new hypothesis regarding the problem of the Synoptic Gospels implied a reconstruction of the life and message of Jesus.

The fundamental assumption for most of this work on the life and teachings of Jesus was a distinction between the "Jesus of history" and the "Christ of faith." Another favourite way of putting the distinction was to speak of the religion *of* Jesus in antithesis to the religion *about* Jesus. This implied that Jesus was a man like other men, but with a heightened awareness of the presence and power of God. Then the dogma of the church had mistaken this awareness for a metaphysical statement that Jesus was the Son of God and had thus distorted the original simplicity of his message. Some critics went so far as to question the very historicity of Jesus, but even those who did not go that far questioned the historicity of some of the sayings and deeds attributed to Jesus in the Gospels.

In part this effort grew out of the general concern of 19th-century scholarsip with the problem of history, but it also reflected the religious and ethical assumptions of the theologians. Many of them were influenced by the moral theories of Kant in their estimate of what was permanent about the teachings of Jesus, and by the historical theories of Hegel in the way they related the original message of Jesus to the Christian interpretations of that message by later generations of Christians. The ideas of evolution and of natural causality associated with the science of the 19th century also played a part through the naturalistic explanations of the biblical miracles. And the historians of dogma, climaxing in Adolf von Harnack (1851–1931), used their demonstration of the dependence of ancient Christology upon non-Christian sources for its concepts and terminology to reinforce their claim that Christianity had to get back from the Christ of dogma to the "essence of Christianity" in the teachings of Jesus about the fatherhood of God and the brotherhood of man.

The 20th century. At the beginning of the 20th century the most influential authorities on the New Testament were engaged in this quest for the essence of Christianity and for the Jesus of history. But that quest led in the early decades of the 20th century to a revolutionary conclusion regarding the teachings of Jesus, namely, that he had expected the end of the age to come shortly after his death and that his teachings as laid down in the Gospels were an "interim ethic," intended for the messianic community in the brief span of time still remaining before the end. The effort to apply those teachings in modern life was criticized as a dangerous modernization. This thesis of the "consistent eschatology" in Jesus' message was espoused by Johannes von Weiss (1863–1914) and gained wide circulation through the writings of Albert Schweitzer.

Form criticism

The years surrounding World War I also saw the development of a new theory regarding the composition of the Gospels. Because of its origin, this theory is usually called form criticism (German *Formgeschichte*). It stressed the forms of the Gospel narratives—parables, sayings, miracle stories, Passion accounts, etc.—as an indication of the oral tradition in the Christian community out of which the narratives came. While the attention of earlier scholars had been concentrated on the authenticity of Jesus' teachings as transmitted in the Gospels, this new theory was less confident of being able to separate the authentic from the later elements in the Gospel records, though various proponents of it did suggest criteria by which such a separation might be guided. The studies of form criticism made a life of Jesus in the old biographical sense impossible, just as consistent eschatology had declared impossible the codification of a universal ethic from the teachings of Jesus. Some adherents of form criticism espoused an extreme skepticism regarding any historical knowledge of Jesus' life at all, but the work of men like Martin Dibelius and even Rudolf Bultmann showed that such skepticism was not warranted by the conclusions of this study.

Influenced by these trends in New Testament study, Protestant theology by the middle of the 20th century was engaged in a reinterpretation of the Christology of the early church. Some Protestant churches continued to repeat the formulas of ancient dogma, but even there the critical study of the New Testament documents was beginning to call those formulas into question. The struggles of the evangelical churches in Germany under Adolf Hitler caused some theologians to realize anew the power of the ancient dogma of the person of Christ to sustain faith, and some of them were inclined to treat the dogma with less severity. But even they acknowledged that the formulation of that dogma in static categories of person, essence, and nature was inadequate to the biblical emphasis upon actions and events rather than upon states of being. Karl Barth for the Reformed tradition, Lionel Thornton for the Anglican tradition, and Karl Heim for the Lutheran tradition were instances of theologians trying to reinterpret classical Christology. While yielding nothing of their loyalty to the dogma of the church, Roman Catholic theologians like Karl Adam were also endeavouring to state that dogma in a form that was meaningful to modern men. The doctrine of the work of Christ was receiving less attention than the doctrine of Christ's person. In much of Protestantism, the concentration of the 19th century upon the teachings of Jesus had made it difficult to speak of more than the prophetic office. The priestly office received least attention of all; and, therefore, despite the support accorded to efforts like that of Gustaf Aulén to reinterpret the metaphor of the Atonement as Christ's victory over his enemies, Protestant theology in the middle of the 20th century was still searching for a doctrine of the Atonement to match its newly won insights into the doctrine of the person of Christ.

In a curious way, therefore, the figure of Jesus Christ has become both a unitive and a divisive element in Christendom. All Christians are united in their loyalty to him, even though they express their loyalty in a variety of doctrinal and liturgical ways. But doctrine and liturgy also divide Christian communions from one another. It has not been the official statements about Christ that have differed widely among most communions. What has become a sharp point of division is the amount of historical and critical inquiry that is permitted where the person of Christ is involved. Despite their official statements and confessions, most Protestant denominations had indicated by the second half of the 20th century that they would tolerate such inquiry, differ though they did in prescribing how far it would be permitted to go. On the other hand, the exclusion of Modernism by the Roman Catholic Church in 1907–10 drew definite limits beyond which the theological use of the methods of critical inquiry was heretical. Within those limits, however, Roman Catholic biblical scholars were engaging in considerable critical literary study, at the same time that critical Protestant theologians were becoming more sympathetic to traditional Christological formulas. (J.J.Pe.)

BIBLIOGRAPHY

Times and environment: BO REICKE, *The New Testament Era* (1968, reissued 1978; originally published in German, 1964); ANTHONY E. HARVEY, *Jesus and the Constraints of History* (1982), a study of the constraints imposed on Jesus by contemporary conditions.

The life and ministry of Jesus: Valuable surveys of premodern material are provided by ROBERT M. GRANT, *The Earliest Lives of Jesus* (1961), for the early patristic period; and HARVEY K. MCARTHUR, *The Quest Through the Centuries* (1966), especially for the 14th and 16th centuries. Modern historical study of the Gospels dates from the 18th century and was characterized in the 19th and early 20th centuries by the Life-of-Jesus movement. The literature is authoritatively surveyed in ALBERT SCHWEITZER, *The Quest of the Historical Jesus* (1910, reissued 1968; originally published in German, 1906); and C.C. MCCOWN, *The Search for the Real Jesus* (1940). JOHN F. O'GRADY, *Models of Jesus* (1981); and JOHN FERGUSON, *Jesus in the Tide of Time: A Historical Study* (1980), are studies of interpretations of Jesus in different ages and cultures. MAURICE GOGUEL, *The Life of Jesus* (1933, reissued 1976; originally published in French, 1932), is a biography especially valuable for the inclusion of detailed evidence frequently assumed in subsequent works. JOSEPH KLAUSNER, *Jesus of Nazareth* (1925, reissued 1979; originally published in Hebrew, 1922), remains important for its collection of relevant rabbinic materials. SHIRLEY J. CASE, *Jesus: A New Biography* (1927, reprinted 1968), in spite of its title, is a socio-historical treatment of the subject. Other lives include VINCENT TAYLOR, *The Life and Ministry of Jesus* (1954); ETHELBERT STAUFFER, *Jesus and His Story* (1960, originally published in German, 1957); and DAVID FLUSSER, *Jesus* (1969; originally published in German, 1968). The first half of the 20th century witnessed a marked decline in the appearance of biographies and studies of the historical Jesus. Several factors contributed to the decline, of which the most important was the rise of form criticism in the first quarter of the century as exemplified in MARTIN DIBELIUS, *From Tradition to Gospel* (1934, reissued 1971; trans. of the rev. 2nd German ed., 1933); and RUDOLF BULTMANN, *The History of the Synoptic Tradition*, rev. ed. (1968, reissued 1972; originally published in German, 1921). Bultmann's *Jesus and the Word* (1934, reissued 1958; originally published in German, 1926) presents a so-called encounter with the message of Jesus but without sharp differentiation of that message from the church's "earliest tradition." The period 1950–70 produced a resurgence of interest in the historical Jesus. Stimulus for the resurgence is often credited to a paper by ERNST KASEMANN, "The Problem of the Historical Jesus," in W.J. MONTAGUE (trans.), *Essays on New Testament Themes* (1964). The "new quest," as it came to be called after JAMES M. ROBINSON, *A New Quest of the Historical Jesus* (1959, reissued 1983), arose as a question concerning continuity between Jesus and his message, on the one hand, and the early church's proclamation of Christ, on the other. Some of the alternative positions concerning the possibility of a biography of Jesus may be seen in CARL E. BRAATEN and ROY A. HARRISVILLE (eds.), *The Historical Jesus and the Kerygmatic Christ* (1964), a collection of essays by major figures. Distinctive positions are to be found in ERNST FUCHS, *Studies of the Historical Jesus* (1964; originally published in German, 1964); JOHN KNOX, *The Church and the Reality of Christ* (1962); HUGH ANDERSON, *Jesus and Christian Origins* (1964); LEANDER E. KECK, *A Future for the Historical Jesus* (1971, reissued 1981); FREDERICK F. BRUCE, *Jesus and Christian Origins Outside the New Testament* (1974), an annotated list of sources; JAMES P. MACKEY, *Jesus, the Man and the Myth: A Contemporary Christology* (1979); GEORGE VERMES, *Jesus the Jew: a Historian's Reading of the Gospels* (1974), the thesis that Jesus was a Galilean Ḥasid; and EDWARD SCHILLEBEECKX, *Jesus: An Experiment in Christology* (1979; originally published in Dutch, 1975), a review and interpretation of modern scholarship. Accurate popular presentations are in HEINZ ZAHRNT, *The Historical Jesus* (1963; originally published in German, 1960); and in the moderately conservative JOACHIM JEREMIAS, *The Problem of the Historical Jesus* (1964, reissued 1972; originally published in German, 1960). The entire trend of Gospel study since the rise of form criticism is challenged by BIRGER GERHARDSSON in *Memory and Manuscript* (1961), who argues that Jesus himself taught his disciples, in the manner of a Jewish rabbi, to memorize and transmit traditions in a fixed form. J. ARTHUR BAIRD, *Audience Criticism and the Historical Jesus* (1969), explores the use of the computer in literary analysis but is somewhat ambiguous with respect to total historical methodology. A standard study is GÜNTHER BORNKAMM, *Jesus of Nazareth* (1960, reissued 1975; originally published in German, 1956). ROLAND H. BAINTON, *Behold the Christ* (1974), discusses works of art representative of various interpretations of Christ.

The message of Jesus: Many of the titles listed in the preceding section include discussions of the teachings of Jesus. NORMAN PERRIN, *Rediscovering the Teaching of Jesus* (1967, reissued 1976), is an important technical introduction. THOMAS W. MANSON, *The Sayings of Jesus* (1949, reissued 1979), is a valuable commentary on the so-called Q material. Since JOHANNES WEISS, *Jesus' Proclamation of the Kingdom of God* (1971; originally published in German, 1892), it has been recognized that the eschatological Kingdom of God was the centre of Jesus' message. Whether the message implied a wholly futuristic expectation, a "realized" eschatology, or a future kingdom with present manifestations is still discussed. Convenient summaries of alternative interpretations are presented in NORMAN PERRIN, *The Kingdom of God in the Teaching of Jesus* (1963; reissued 1975); and GÖSTA LUNDSTRÖM, *The Kingdom of God in the Teaching of Jesus* (1963; originally published in Swedish, 1947). Perrin's work also includes a survey of interpretations of the phrase "Son of man" and other titles related to Jesus' vocation. Representative modern studies include WERNER G. KÜMMEL, *Promise and Fulfillment* (1957; originally published in German, 1945); RUDOLF SCHNACKENBURG, *God's Rule and Kingdom* (1963; originally published in German, 1959); GEORGE ELDON LADD, *Jesus and the Kingdom* (1964), written from a conservative perspective; and H.E. TÖDT, *The Son of Man in the Synoptic Tradition* (1965; originally published in German, 1959). Modern investigation of the parables derives from C.H. DODD, *The Parables of the Kingdom*, rev. ed. (1961). A definitive work is JOACHIM JEREMIAS, *The Parables of Jesus*, 3rd rev. ed. (1972; originally published in German, 1947). DAN O. VIA, *The Parables* (1967, reissued 1977), is one of several books in which literary critical insights are employed to advance interpretation beyond strictly historical critical study. ETA LINNEMANN, *Jesus of the Parables* (1967; originally published in German, 1961), advances views along the lines proposed by Ernst Fuchs. MARTIN DIBELIUS, *The Sermon on the Mount* (1940); HANS WINDISCH, *The Meaning of the Sermon on the Mount* (1951; originally published in German, 1929); and WILLIAM D. DAVIES, *The Setting of the Sermon on the Mount* (1964, reissued 1976), offer varying interpretations of the most familiar body of Jesus' teaching. See also MILAN MACHOVEČ, *A Marxist Looks at Jesus* (1976; originally published in German, 1972), a sympathetic study; JON SOBRINO, *Christology at the Crossroads: A Latin American Approach* (1978; originally published in Spanish, 1976), representative of "liberation theology"; ROSEMARY R. RUETHER, *To Change the World: Christology and Cultural Criticism* (1981), a feminist Christology; CHARLES B. KETCHAM, *A Theology of Encounter: The Ontological Ground for a New Christology* (1978), an existentialist's Christology; and RUSSEL PREGEANT, *Christology Beyond Dogma* (1978), a hermeneutical study of Matthew.

The sufferings and death of Jesus: EDUARD LOHSE, *History of the Suffering and Death of Jesus Christ* (1967; originally published in German, 1964), provides a convenient overview of the problems. Of the massive literature on the trial and Crucifixion, JOSEPH BLINZLER, *The Trial of Jesus* (1959; trans. from 2nd rev. German ed., 1955); and PAUL WINTER, *On the Trial of Jesus*, 2nd ed. (1974), have been very influential. The range of scholarly opinion about the trial is well represented by eight responsible essays in *Judaism*, 20:6–74 (1971).

The story of Jesus and faith in Jesus: REGINALD H. FULLER, *The Mission and Achievement of Jesus* (1954, reissued 1967), and *The Foundations of New Testament Christology* (1965), partly because of the author's shift in viewpoint, exhibit a range of opinion on the relation of Jesus' sense of vocation to the church's Christology. A more difficult study is that of FERDINAND HAHN, *The Titles of Jesus in Christology* (1969; originally published in German, 1963).

Joan of Arc

Joan of Arc (French Jeanne d'Arc) was a peasant girl who, believing that she was acting under divine guidance, led the French army in a momentous victory at Orléans that repulsed an English attempt to conquer France during the Hundred Years' War. Captured a year afterward, Joan was burned by the English and their French collaborators as a heretic. She became the greatest national heroine of her compatriots. Her achievement was a decisive factor in the awakening of French national consciousness. She was canonized in 1920.

Giraudon—Art Resource/EB Inc.

St. Joan, equestrian miniature from the manuscript *La Vie des femmes célèbres* by Antoine Dufour, c. 1505. In the Musée Archéologique Thomas Dobrée, Nantes, Fr.

Joan was born *c.* 1412, the daughter of a tenant farmer, at Domrémy, on the borders of the duchies of Bar and Lorraine. In her mission of expelling the English and their Burgundian allies from the Valois kingdom of France, she felt herself to be guided by the "voices" of St. Michael, St. Catherine, and St. Margaret. She possessed many attributes characteristic of the female visionaries who were a noted feature of her time. These qualities included extreme personal piety, a claim to direct communication with the saints, and a consequent reliance upon individual experience of God's presence beyond the ministrations of the priesthood and the confines of the institutional church. But to these were added remarkable mental and physical courage, as well as a robust common sense. Known as La Pucelle, or the Maid of Orléans, Joan became in the following centuries a focus of unity for the French people, especially at times of crisis.

JOAN'S MISSION

The crown of France at the time was in dispute between the dauphin Charles, son and heir of the Valois king Charles VI, and the Lancastrian English king Henry VI. Henry's armies were in alliance with those of Philip the Good, duke of Burgundy (whose father, John the Fearless, had been assassinated in 1419 by partisans of the Dauphin), and were occupying much of the northern part of the kingdom. The apparent hopelessness of the Dauphin's cause at the end of 1427 was increased by the fact that, five years after his father's death, he still had not been crowned. Reims, the traditional place for the investiture of French kings, was well within the territory held by his enemies. As long as the Dauphin remained unconsecrated, the rightfulness of his claim to be king of France was open to challenge.

Joan's village of Domrémy was on the frontier between the France of the Anglo-Burgundians and that of the Dauphin. The villagers had already had to abandon their homes before Burgundian threats. Led by her voices, Joan traveled in May 1428 from Domrémy to Vaucouleurs, the nearest stronghold still loyal to the Dauphin, where she asked the captain of the garrison, Robert de Baudricourt, for permission to join the Dauphin. He did not take the 16-year-old girl and her visions seriously, and she returned home. Joan went to Vaucouleurs again in January 1429. This time her quiet firmness and piety gained her the respect of the people; and the captain, persuaded that she was neither a witch nor feebleminded, allowed her to go to the Dauphin at Chinon. She left Vaucouleurs about February 13, dressed in men's clothes and accompanied by six men-at-arms. Crossing territory held by the enemy, and traveling for 11 days, she reached Chinon.

Meeting with the Dauphin

Joan went at once to the castle occupied by the dauphin Charles. He was uncertain whether to receive her, and his counselors gave him conflicting advice; but two days later he granted her an audience. Charles had hidden himself among his courtiers, but Joan made straight for him and told him that she wished to go to battle against the English and that she would have him crowned at Reims. On the Dauphin's orders she was immediately interrogated by ecclesiastical authorities in the presence of Jean, duc d'Alençon, a relative of Charles, who showed himself well-disposed toward her. For three weeks she was further questioned at Poitiers by eminent theologians who were allied to the Dauphin's cause. These examinations, the record of which has not survived, were occasioned by the ever-present fear of heresy following the end of the Great Schism in 1417. Joan told the ecclesiastics that it was not at Poitiers but at Orléans that she would give proof of her mission; and forthwith, on March 22, she dictated letters of defiance to the English. In their report the churchmen suggested that in view of the desperate situation of Orléans, which had been under English siege for months, the Dauphin would be well-advised to make use of her.

Joan returned to Chinon. At Tours, during April, the Dauphin provided her with a military household of several men; Jean d'Aulon became her squire, and she was joined by her brothers Jean and Pierre. She had her standard painted with an image of Christ in Judgment and a banner made bearing the name of Jesus. When the question of a sword was brought up, she declared that it would be found in the church of Sainte-Catherine-de-Fierbois, and one was in fact discovered there.

Troops numbering several hundred men were mustered at Blois, and on April 27 they set out for Orléans. The city, besieged since Oct. 12, 1428, was almost totally surrounded by a ring of English strongholds. When Joan and one of the French commanders, La Hire, entered with supplies on April 29, she was told that action must be deferred until further reinforcements could be brought in.

Relief of Orléans

On the evening of May 4, when Joan was resting, she suddenly sprang up, apparently inspired, and announced that she must go and attack the English. Having herself armed, she hurried out to the east of the city toward an English fort where, indeed, an engagement of which she had not been told was taking place. Her arrival roused the French, and they took the fort. The next day Joan addressed another of her letters of defiance to the English. On the morning of May 6 she crossed to the south bank of the river and advanced toward another fort; the English immediately evacuated it in order to defend a stronger position nearby, but Joan and La Hire attacked them there and took it by storm. Very early on May 7 the French advanced against the fort of Les Tourelles. Joan was wounded but quickly returned to the fight, and it was thanks in part to her example that the French commanders maintained the attack until the English capitulated. Next day the English were seen to be retreating, but, because it was a Sunday, Joan refused to allow any pursuit.

Joan left Orléans on May 9 and met Charles at Tours. She urged him to make haste to Reims to be crowned. Though he hesitated because some of his more prudent counselors were advising him to undertake the conquest of Normandy, Joan's importunity ultimately carried the day. It was decided, however, first to clear the English out of the other towns along the Loire River. Joan met her friend the Duc d'Alençon, who had been made lieutenant general of the French armies, and they moved off together, taking a town and an important bridge. They next attacked Beaugency, whereupon the English retreated into the castle. Then, notwithstanding the opposition of the Dauphin and Georges de La Trémoille, one of his favourites, and despite the reserve of Alençon, Joan received the Constable de Richemont, who was under suspicion at the French court. After making him swear fidelity, she accepted his help. Shortly thereafter the castle of Beaugency was surrendered.

The French and English armies came face to face at Patay on June 18, 1429. Joan promised success to the French, saying that Charles would win a greater victory that day than any he had won so far. The victory was indeed complete; the English army was routed and with it, finally, its reputation for invincibility.

Battle of Patay

Instead of pressing home their advantage by a bold attack upon Paris, Joan and the French commanders turned back to rejoin the Dauphin, who was staying with La Trémoille at Sully-sur-Loire. Again Joan urged upon Charles the need to go on swiftly to Reims. He vacillated, however; and as he meandered through the towns along the Loire, Joan accompanied him, arguing all the while in an attempt to vanquish his hesitancy and prevail over the counselors who advised delay. She was not unaware of the dangers and difficulties involved but declared them of no account. Finally she won Charles to her view.

From Gien, where the army began to assemble, the Dauphin sent out the customary letters of summons to the coronation. Joan wrote two letters: one of exhortation to the people of Tournai, always loyal to Charles, the other a challenge to Philip the Good, duke of Burgundy. She and the Dauphin set out on the march to Reims on June 29. Before arriving at Troyes, Joan wrote to the inhabitants, promising them pardon if they would submit. They countered by sending a friar, the popular preacher Brother Richard, to take stock of her; but though he returned full of enthusiasm for the Maid and her mission, the townsfolk decided after all to remain loyal to the Anglo-Burgundian regime. The Dauphin held a council, and Joan proposed that the town be attacked. The next morning she began the assault, and the citizens at once asked for terms. The royal army then marched on to Châlons. Despite an earlier decision to resist, the Count-Bishop handed the keys of the town to Charles. On July 16 the royal army reached Reims, which opened its gates. The coronation took place on July 17, 1429. Joan was present at the consecration, standing with her banner not far from the altar. After the ceremony she knelt before Charles, calling him her king for the first time. That same day she wrote to the Duke of Burgundy, adjuring him to make peace with the King and to withdraw his garrisons from the royal fortresses.

Coronation of the Dauphin

Charles VII left Reims on July 20, and for a month the army paraded through Champagne and the Île-de-France. On August 2 the King decided on a retreat from Provins to the Loire, a move that implied abandoning any plan to attack Paris. The loyal towns that would thus have been left to the enemy's mercy expressed some alarm. Joan, who was opposed to Charles's decision, wrote to reassure the citizens of Reims on August 5, saying that the Duke of Burgundy, then in possession of Paris, had made a fortnight's truce, after which it was hoped that he would yield Paris to the King. In fact, on August 6, English troops prevented the royal army from crossing the Seine at Bray, much to the delight of Joan and the commanders, who hoped that Charles would attack Paris. Everywhere acclaimed, Joan was now, according to a 15th-century chronicler, the idol of the French. She herself felt that the purpose of her mission had been achieved.

Near Senlis, on August 14, the French and English armies again confronted each other. This time only skirmishes took place, neither side daring to start a battle, though Joan carried her standard up to the enemy's earthworks and openly challenged them. Meanwhile Compiègne, Beauvais, Senlis, and other towns north of Paris surrendered to the King. Soon afterward, on August 28, a four months' truce for all the territory north of the Seine was concluded with the Burgundians.

Joan, however, was becoming more and more impatient; she thought it essential to take Paris. She and Alençon were at Saint-Denis on the northern outskirts of Paris on August 26, and the Parisians began to organize their defenses. Charles arrived on September 7, and an attack was launched on September 8, directed between the gates of Saint-Honoré and Saint-Denis. The Parisians could be in no doubt of Joan's presence among the besiegers; she stood forward on the earthworks, calling on them to surrender their city to the King of France. Wounded, she continued to encourage the soldiers until she had to abandon the attack. Though the next day she and Alençon sought to renew the assault, they were ordered by Charles's council to retreat.

Attack on Paris

Charles VII retired to the Loire, Joan following him. At Gien, which they reached on September 22, the army was disbanded. Alençon and the other captains went home; only Joan remained with the King. Later, when Alençon was planning a campaign in Normandy, he asked the King to let Joan rejoin him, but La Trémoille and other courtiers dissuaded him. Joan went with the King to Bourges, where many years later she was to be remembered for her goodness and her generosity to the poor. In October she was sent against Saint-Pierre-le-Moûtier; through her courageous assault, with only a few men, the town was taken. Joan's army then laid siege to La Charité-sur-Loire; short of munitions, they appealed to neighbouring towns for help. The supplies arrived too late, and after a month they had to withdraw.

Compiègne

Joan then rejoined the King, who was spending the winter in towns along the Loire. Late in December 1429 Charles issued letters patent ennobling Joan, her parents, and her brothers. Early in 1430 the Duke of Burgundy began to threaten Brie and Champagne. The inhabitants of Reims became alarmed, and Joan wrote in March to assure them of the King's concern and to promise that she would come to their defense. When the Duke moved up to attack Compiègne, the townsfolk determined to resist, and in late March or early April Joan left the King and set out to their aid, accompanied only by her brother Pierre, her squire Jean d'Aulon, and a small troop of men-at-arms. She arrived at Melun in the middle of April, and it was no doubt her presence that prompted the citizens there to declare themselves for Charles VII.

Joan was at Compiègne by May 14, 1430. There she found Renaud de Chartres, archbishop of Reims, and Louis I de Bourbon, comte de Vendôme, a relative of the King. With them she went on to Soissons, where the townspeople refused them entry. Renaud and Vendôme therefore decided to return south of the Marne and Seine rivers; but Joan refused to accompany them, preferring to return to her "good friends" in Compiègne.

CAPTURE, TRIAL, AND EXECUTION

On her way back Joan heard that John of Luxembourg, the captain of a Burgundian company, had laid siege to Compiègne. Hurrying on, she entered Compiègne under cover of darkness. The next afternoon, May 23, she led a sortie and twice repelled the Burgundians but was eventually outflanked by English reinforcements and compelled to retreat. Remaining until the last to protect the rear guard while they crossed the Oise River, she was unhorsed and could not remount. She gave herself up and, with her brother Pierre and Jean d'Aulon, was taken to Margny, where the Duke of Burgundy came to see her. In telling the people of Reims of Joan's capture, Renaud de Chartres accused her of rejecting all counsel and acting willfully. Charles, who was working toward a truce with the Duke of Burgundy, made no attempts to save her.

Imprisonment

John of Luxembourg sent Joan and Jean d'Aulon to his castle in Vermandois. When she tried to escape in order to return to Compiègne, he sent her to one of his more distant castles. There, though she was treated kindly, she became more and more distressed at the predicament of Compiègne. Her desire to escape became so great that she jumped from the top of a tower, falling unconscious into the moat. She was not seriously hurt, and when she had recovered, she was taken to Arras, a town adhering to the Duke of Burgundy.

News of her capture had reached Paris on May 25. Next day the University of Paris, which had taken the English side, requested the Duke of Burgundy to turn her over for judgment either to the chief inquisitor or to the bishop of Beauvais, Pierre Cauchon, in whose diocese she had been seized. The university wrote also, to the same effect, to John of Luxembourg; and on July 14 the Bishop of Beauvais presented himself before the Duke of Burgundy asking, on his own behalf and in the name of the English king, that the Maid be handed over in return for a payment of 10,000 francs. The Duke passed on the demand to John of Luxembourg, and by Jan. 3, 1431, she was in the Bishop's hands. The trial was fixed to take place at Rouen. Joan was moved to a tower in the castle of Bouvreuil, which was occupied by the Earl of Warwick, the English commander at Rouen. Though her offenses against the Lancastrian monarchy were common knowledge, Joan was brought to trial before a church court because the University of Paris, as arbiter in matters concerning the faith, insisted that she be tried as a heretic. Her beliefs were not strictly orthodox, according to the criteria for orthodoxy laid down by many theologians of the period. She was no friend of the church militant on Earth (which perceived itself as in spiritual combat with the forces of evil), and she threatened its hierarchy through her claim that she communicated directly with God by means of visions or voices. Further, her trial might serve to discredit Charles VII by demonstrating that he owed his coronation to a witch, or at least a heretic. Her two judges were to be Cauchon, bishop of Beauvais, and Jean Lemaître, the vice-inquisitor of France.

Interrogation

The trial. Beginning Jan. 13, 1431, statements taken in Lorraine and elsewhere were read before the Bishop and his assessors; they were to provide the framework for Joan's interrogation. Summoned to appear before her judges on February 21, Joan asked for permission to attend mass beforehand, but it was refused on account of the gravity of the crimes with which she was charged, including attempted suicide in having jumped into the moat. She was ordered to swear to tell the truth and did so swear, but she always refused to reveal the things she had said to Charles. Cauchon forbade her to leave her prison, but Joan insisted that she was morally free to attempt escape. Guards were then assigned to remain always inside the cell with her, and she was chained to a wooden block and sometimes put in irons. Between February 21 and March 24 she was interrogated nearly a dozen times. On every occasion she was required to swear anew to tell the truth, but she always made it clear that she would not necessarily divulge everything to her judges since, although nearly all of them were Frenchmen, they were enemies of King Charles. The report of this preliminary questioning was read to her on March 24, and apart from two points she admitted its accuracy.

When the trial proper began a day or so later, it took two days for Joan to answer the 70 charges that had been drawn up against her. These were based mainly on the contention that her whole attitude and behaviour showed blasphemous presumption: in particular, that she claimed for her pronouncements the authority of divine revelation; prophesied the future; endorsed her letters with the names of Jesus and Mary, thereby identifying herself with the novel and suspect cult of the Name of Jesus; professed to be assured of salvation; and wore men's clothing. Perhaps the most serious charge was of preferring what she believed to be the direct commands of God to those of the church.

On March 31 she was questioned again on several points about which she had been evasive, notably on the question of her submission to the church. In her position, obedience to the court that was trying her was inevitably made a test of such submission. She did her best to avoid this trap, saying she knew well that the church militant could not err, but it was to God and to her saints that she held herself answerable for her words and actions. The trial continued, and the 70 charges were reduced to 12, which were sent for consideration to many eminent theologians in both Rouen and Paris.

Meanwhile, Joan fell sick in prison and was attended by two doctors. She received a visit on April 18 from Cauchon and his assistants, who exhorted her to submit to the church. Joan, who was seriously ill and obviously thought she was dying, begged to be allowed to go to confession and receive Holy Communion and to be buried in consecrated ground. But they continued to badger her, receiving only her constant response "I am relying on our Lord," "I hold to what I have already said." They became more insistent on May 9, threatening her with torture if she did not clarify certain points. She answered that even if they tortured her to death she would not reply differently, adding that in any case she would afterward maintain that any statement she might make had been extorted from her by force. In face of this commonsense fortitude her interrogators, by a majority of 10 to three, decided on May 12 that torture would be useless. Joan was informed on May 23 of the decision of the University of Paris that if she persisted in her errors she would be turned over to the secular authorities; only they, and not the church, could carry out the death sentence of a condemned heretic.

Abjuration, relapse, and execution. Apparently nothing further could be done. Joan was taken out of prison for the first time in four months on May 24 and conducted to the cemetery of the church of Saint-Ouen, where her sentence was to be read out. First she was made to listen to a sermon by one of the theologians in which he violently attacked Charles VII, provoking Joan to interrupt him because she thought he had no right to attack the King, a "good Christian," and should confine his strictures to her. After the sermon was ended, she asked that all the evidence on her words and deeds be sent to Rome. But her judges ignored her appeal to the Pope, to whom, under God, she would be answerable, and began to read out the sentence abandoning her to the secular power. Hearing this dreadful pronouncement, Joan quailed and declared she would do all that the church required of her. She was presented with a form of abjuration, which must already have been prepared. She hesitated in signing it, eventually doing so on condition that it was "pleasing to our Lord." She was then condemned to perpetual imprisonment or, as some maintain, to incarceration in a place habitually used as a prison. In any case, the judges required her to return to her former prison.

Revocation of abjuration

The vice-inquisitor had ordered Joan to put on women's clothes, and she obeyed. But two or three days later, when the judges and others visited her and found her again in male attire, she said she had made the change of her own free will, preferring men's clothes. They then pressed other questions, to which she answered that the voices of St. Catherine and St. Margaret had censured her "treason" in making an abjuration. These admissions were taken to signify relapse, and on May 29 the judges and 39 assessors agreed unanimously that she must be handed over to the secular officials.

The next morning, May 30, 1431, Joan received from Cauchon permission, unprecedented for a relapsed heretic, to make her confession and receive Communion. Accompanied by two Dominicans, she was then led to the Place du Vieux-Marché. There she endured one more sermon, and the sentence abandoning her to the secular arm—that is, to the English and their French collaborators—was read out in the presence of her judges and a great crowd. The executioner seized her, led her to the stake, and lit the pyre. A Dominican consoled Joan, who asked him to hold high a crucifix for her to see and to shout out the assurances of salvation so loudly that she should hear him above the roar of the flames. To the last she maintained that her voices were sent of God and had not deceived her. According to the rehabilitation proceedings of 1456, few witnesses of her death seem to have doubted her salvation, and they agreed that she died a faithful Christian. A few days later the English king and the University of Paris formally published the news of Joan's execution.

Rehabilitation

Almost 20 years afterward, on his entry into Rouen in 1450, Charles VII ordered an inquiry into the trial. Two years later the cardinal legate Guillaume d'Estouteville made a much more thorough investigation. Finally, on the order of Pope Calixtus III following a petition from the d'Arc family, proceedings were instituted in 1455–56 that revoked and annulled the sentence of 1431. Joan was canonized by Pope Benedict XV on May 16, 1920; her feast day is May 30. The French parliament, on June 24, 1920, decreed a yearly national festival in her honour; this is held the second Sunday in May.

CHARACTER AND IMPORTANCE

Joan of Arc's place in history is assured. Perhaps her contribution to the history of human courage is greater than her significance in the political and military history of France. She was victimized as much by a French civil conflict as by a war with a foreign power. The relief of Orléans was undoubtedly a notable victory, which secured the loyalty of certain regions of northern France to the régime of Charles VII. But the Hundred Years' War continued for a further 22 years after her death, and it was the defection of Philip the Good of Burgundy from his alliance with the Lancastrians in 1435 that provided the foundation upon which the recovery of Valois France was to be based. The nature of Joan's mission, moreover, is a source of controversy among historians, theologians, and psychologists. Innumerable points about her campaigns and about the motives and actions of her supporters and enemies are subject to dispute: for instance, the number and dates of her visits to Vaucouleurs, Chinon, and Poitiers; how she was able to win the confidence of the Dauphin at their first meeting at Chinon; whether Charles's perambulations after his coronation at Reims represented triumphant progress or scandalous indecision; what her judges meant by "perpetual imprisonment"; whether, after her recantation, Joan resumed men's clothes of her own free will and at the bidding of her voices or, as one later story has it, because they were forced upon her by her English jailers.

Assessment

Later generations have tended to distort the significance of Joan's mission according to their own political and religious viewpoints rather than seeking to set it in the troubled context of her time. The effects of the Great Schism within the Western Church (1378–1417) and the decline of papal authority during the Conciliar Movement (1409–49) made it difficult for persons to seek independent arbitration and judgment in cases relating to the faith. The verdicts of the Inquisition were liable to be coloured by political and other influences; and Joan was not the only victim of an essentially unjust procedure, which allowed the accused no counsel for the defense and which sanctioned interrogation under duress. Her place among the saints is secured, not perhaps by the somewhat dubious miracles attributed to her, but by the heroic fortitude with which she endured the ordeal of her trial and, except for one lapse toward its end, by her profound conviction of the justice of her cause, sustained by faith in the divine origin of her voices. In many ways a victim of internal strife within France, condemned by judges and assessors who were almost entirely northern French in origin, she has become a symbol of national consciousness with whom all French people, of whatever creed or party, can identify.

BIBLIOGRAPHY. Thereis an extremely large literature on Joan of Arc but no truly definitive biography. Important works include: JULES QUICHERAT (ed.), *Procès de condamnation et de réhabilitation de Jeanne d'Arc,* 5 vol. (1841–49, reprinted 1965); *Le Procès de condamnation de Jeanne d'Arc,* 2 vol., trans. and annotated by PIERRE TISSET (1960–70); PAUL DONCOEUR (ed.), *La Minute française des interrogatoires de Jeanne la Pucelle* (1952); *La Réhabilitation de Jeanne la Pucelle,* 3 vol., trans. and annotated by PAUL DONCOEUR and YVONNE LANHERS (1956–61), containing the recorded testimony from the trial of more than 100 witnesses who had known Joan of Arc; and PIERRE LANÉRY D'ARC, *Les Mémoires et consultations en faveur de Jeanne d'Arc* (1889), dealing with the memoirs of jurists and theologians, intended to facilitate her rehabilitation. In English, DANIEL S. RANKIN and CLAIRE QUINTAL (eds. and trans.), *The First Biography of Joan of Arc with the Chronicle Record of a Contemporary Account* (1964); and RÉGINE PERNOUD, *Joan of Arc by Herself and Her Witnesses* (1964, reissued 1982; originally published in French, 1962), and *The Retrial of Joan of Arc: The Evidence at the Trial for Her Rehabilitation* (1955; originally published in French, 1953), are particularly recommended. Other works include GEORGES DUBY and ANDRÉE DUBY (eds.), *Les Procès de Jeanne d'Arc* (1973); JOHN H. SMITH, *Joan of Arc* (1973), which explores her military activities; HENRI GUILLEMIN, *Joan, Maid of Orléans* (1973; originally published in French, 1970); PHILIPPE WOLFF, "Le Théologien Pierre Cauchon, de sinistre mémoire," in *Économies et sociétés au Moyen Âge: Mélanges offerts à Edouard Perroy,* pp. 553–570 (1973); EDWARD LUCIE-SMITH, *Joan of Arc* (1976); WALTER S. SCOTT, *Jeanne d'Arc* (1974); MARINA WARNER, *Joan of Arc: The Image of Female Heroism* (1981, reprinted 1982), a psychohistorical approach with an important survey of the posthumous history of Joan's legend; FRANCES GIES, *Joan of Arc: The Legend and the Reality* (1981), an examination of her life and the literature about her; and COLLOQUE D'HISTOIRE MÉDIÉVALE. ORLÉANS, FRANCE, 1979, *Jeanne d'Arc: une époque, un rayonnement* (1982), an informative collection of scholarly papers. For the political background to Joan's career see M.G.A. VALE, *Charles VII* (1974); C.T. ALLMAND, *Lancastrian Normandy, 1415-1450* (1983); and ROGER G. LITTLE, *The Parlement of Poitiers: War, Government and Politics in France, 1418-1436* (1984).

(Y.L./M.G.A.V.)

Johannesburg

Johannesburg, the largest city of the Republic of South Africa and one of the largest cities on the African continent, is South Africa's economic metropolis. It stands on the southern slopes of the Witwatersrand—commonly called the Rand—a rocky watershed of east–west ridges surrounded by the Transvaal highveld. The greater metropolitan area, also known as Greater Johannesburg, consists of a municipal and a magisterial area.

Johannesburg grew at a remarkable speed after the discovery of gold on the Rand in 1886. It now lies at the centre of the country's gold-mining industry. Johannesburg is the only city of its size that is not situated on a coast, a lakeshore, or on a river.

The article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. Johannesburg is situated in the most thickly populated part of southern Africa. The mines, many of which were almost worked out in the early 1970s, lie to the south and southwest of the city. The city's height at 5,709 feet (1,740 metres) above sea level is measured at Joubert Park near the centre of the city; the highest point, however, is the site of the Republic Observatory, which stands at an altitude of 5,932 feet.

Climate. Johannesburg has an attractive climate; in July the mean temperature is about 50° F (10° C); in December it is 68° F (20° C). Rainfall averages about 33 inches (850 millimetres) a year. The sun shines for an average of nine hours a day in summer and more than seven hours in winter. The city's air and water are relatively pure in most parts; water shortages occur periodically, especially during the winter months.

The city centre

The city layout. The streets of the centre of the city, laid out on a grid plan established in 1886, are narrow, and city blocks are short. More imaginative street planning is in evidence in the suburbs, where many of the streets are tree lined.

There are about 400 suburbs altogether; most of the city's workers live either in the suburbs or in the surrounding dormitory towns. The population is segregated by colour, in accordance with South Africa's official policy of apartheid (literally "apartness"), so that the nonwhite groups—black Africans, Coloureds, Asians—are each restricted to residence in certain areas, which are located particularly in the west and southwest. The major buildings in the city are grouped around Eloff Street, the main shopping area that leads south from the railway station, with Von Brandis, Joubert, and Rissik streets parallel to it. Other streets in the vicinity include Commissioner, Market, Fox, and Main streets (which constitute the commercial area) and Jeppe and Bree streets. Joubert Park, Sturrock Park, George Harrison Park (named for the Australian who first discovered gold-bearing quartz in the area in 1886), Turffontein Race Course, sports grounds, and mine dumps are all in the area. Johannesburg is the meeting place of roads from major towns in South Africa, such as Pretoria, Bloemfontein, Cape Town, and Durban. Not included in the municipal area, though formerly administered by the municipality, is Soweto (South-Western Townships), an independent city originally constituted as a group of towns inhabited by black Africans. Soweto has an area of 26 square miles (67 square kilometres) and is linked with Johannesburg by road and rail. Housing, water, sanitation, and lighting are supplied, and sports fields, schools, and other facilities are provided by the city council. All shopkeepers in the area are black Africans. Nearby townships include Coronationville and Westbury, reserved for Coloured people, and—about 20 miles (32 kilometres) away—Lenasia, a township for Asians.

Architectural styles

It is customary in Johannesburg to demolish and rebuild houses, rather than to follow the European practice of renovating or altering existing buildings. There are numerous blocks of flats or apartments for all types of income groups, and many South Africans have been able to maintain houses with gardens. Architectural styles of all kinds are in evidence. Buildings of the style developed from 18th-century Cape architecture are noticeable in residential areas, whereas business buildings and many houses are derived from modern European or United States models. Public buildings are modern adaptations of older European styles, particularly English, Dutch, and Renaissance Italian. By the early 1970s it had become the practice to construct apartment buildings (called flats in Johannesburg) so that residential density can be increased; it was still the custom, however, to build single-story homes for married nonwhites. Subsidized housing is available for eligible (*i.e.,* low-income) groups of the population, and large building projects are features of Soweto and other areas to the south of the city.

THE PEOPLE

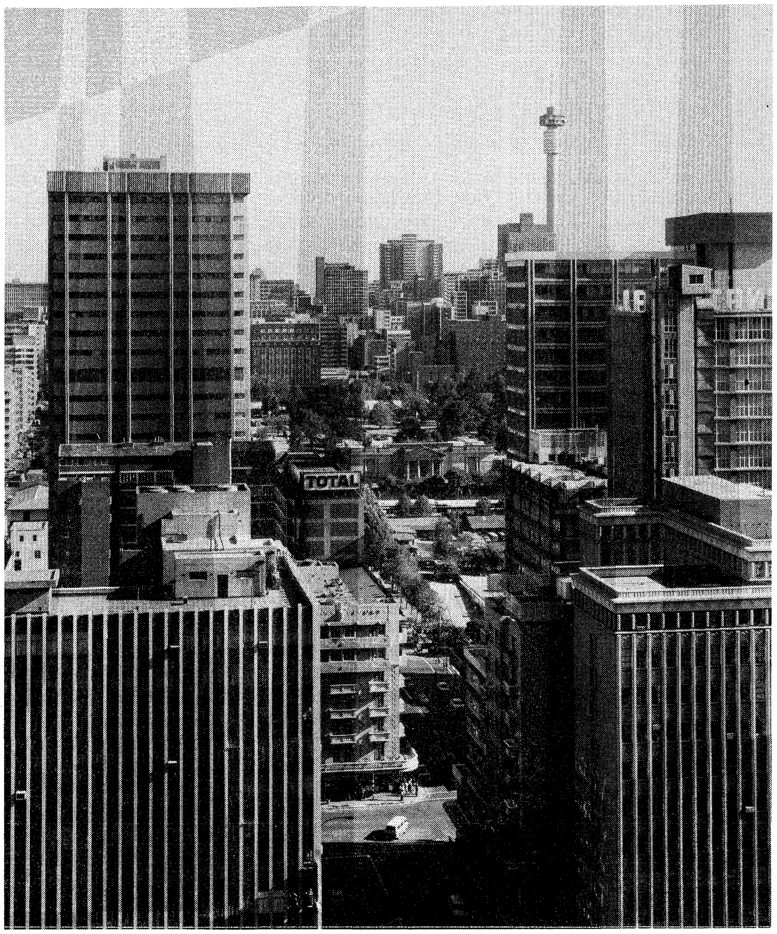
About 60 percent of the city's population is black. Whites comprise approximately one-third of the residents, and the remainder are Coloureds and Asians.

The city is cosmopolitan. Although English- and Afrikaans-speaking groups dominate the white section of the population, minority groups of Germans, Dutch, Hungarians, Italians, French, Scandinavians, Swiss, Poles, and others are also to be found, while Jews from all of these groups have also made Johannesburg their home. Africans speaking Zulu, Xhosa, Pedi, Venda, and Tswana also inhabit the city, while Asian groups include Japanese, Chinese, and Indians, speaking various languages and adhering to various faiths. The Coloured people of the city represent an intermixture of many groups, both white and nonwhite; they speak English or Afrikaans or both, and belong to Christian churches.

THE ECONOMY

South Africa's business centre

Commerce and industry. Because Johannesburg is the centre of the South African business world, more affluent people are to be found in this city than elsewhere in the country. Gold mining is now conducted elsewhere, but most of the business and administrative headquarters of the gold-mining companies are located there. Many secondary industries, particularly of the heavier type, are to be found in the city, and opportunities for employment of many kinds are available. There are many banking, industrial, and commercial concerns, both South African and foreign; the city is also the home of the Johannesburg



Downtown Johannesburg. Joubert Park is in the middle distance, and the J.G. Strijdom (or Hillbrow) Tower for telecommunications is in the right background.

Colour Library International

Stock Exchange, which was founded in 1887. Because of the city's importance, it has a number of branch offices of governmental institutions, as well as consular offices and other institutions (such as banks, building societies, and insurance companies) usually located only in capital cities.

Transportation. Highways have been built in many parts of the city to accommodate the large amount of traffic that enters and leaves the city each day. In the suburbs local public transport is provided by bus services, whereas the railroad serves commuters living outside municipal boundaries to the east, west, and south. Jan Smuts (international) Airport, from which scheduled domestic flights also leave, is situated 14 miles northeast of the city. Charter planes fly from Rand Airport, nine miles to the east; and Baragwanath Aerodrome, seven miles to the southeast, has facilities for gliding and other aerial sports.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. Johannesburg is a municipally controlled city. It has a city council with 47 councillors, elected by popular vote every five years; a mayor is chosen by the councillors from among themselves each year. The city council is the local government unit and is responsible to the Transvaal provincial council, which is in turn responsible to the central government. City councillors represent the same political parties that are represented in the national Parliament.

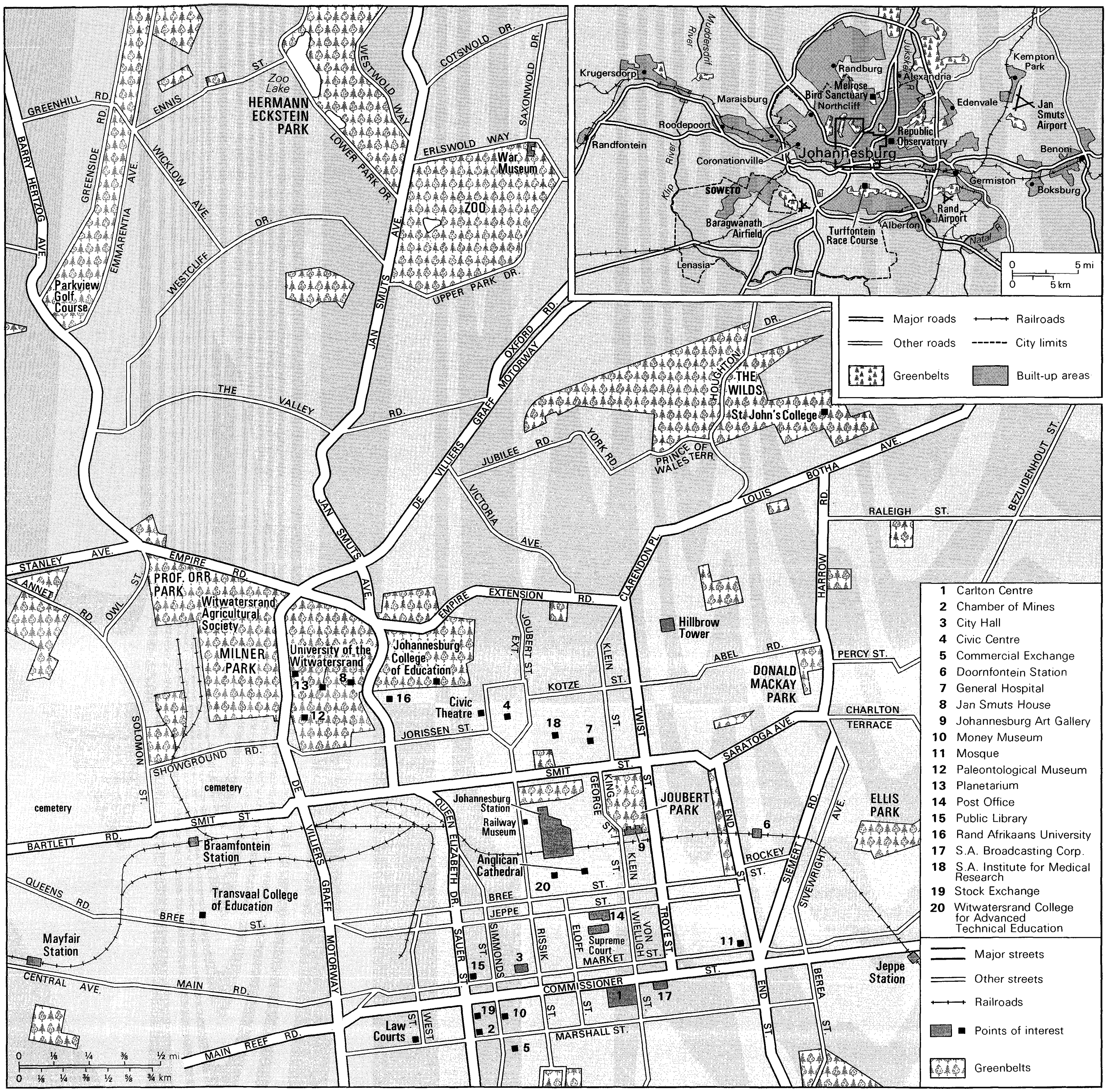
Public services. The local transport, electricity and gas utilities, fire-fighting services, and sanitation are run by the municipality. Water is provided by the municipality through the agency of the Rand Water Board. The railways are government owned, with the city's large railway station constituting the heart of the nation's rail system. Electricity for the mines and railways is provided by a public utility with headquarters in the city.

Police services are provided by the nationally administered South African Police service, but traffic is regulated by the municipality's traffic officers. The city operates produce and livestock markets and the abattoir. The Witwatersrand Agricultural Society organizes a major show every year. It attracts agricultural and industrial exhibitors from throughout the country as well as from overseas. Trade is promoted by the Johannesburg Chamber of Commerce; the Commercial Exchange, a unique institution, owes its existence to commerce generated by the mining industry. Mining operations themselves are regulated by the South African Chamber of Mines, which also has its headquarters in the city.

Health. In addition to Johannesburg General Hospital, medical facilities for whites include many private nursing homes and institutions for sick children, maternity cases, and mine workers. Medical facilities for nonwhites are also extensive, and include clinical, curative, midwifery, and dental services; most mines also provide hospitalization for employees. Medical research is pursued at the South African Institute for Medical Research (founded 1912) as well as at the medical school of the University of the Witwatersrand (founded 1922).

Education. The University of the Witwatersrand itself is mainly intended for English-speaking students; higher education in Afrikaans is provided at the Rand Afrikaans University (founded 1966). The Witwatersrand Technikon (founded 1925) specializes in technological training related to mining and other subjects. In addition, there are throughout the city a large number of schools, both private and governmental, for pupils of all ages, from kindergarten to school-leaving age. Teacher-training courses are given at the Johannesburg College of Education (for English-speaking students), at the Goudstadse Onderwyskollege (for Afrikaans-speakers), at the Transvaal College of Education (for Asians), and at the Rand College of Education (for Coloured students).

Teacher-training colleges



Central Johannesburg and (inset) its metropolitan area.

CULTURAL LIFE

Cultural institutions include the Johannesburg Public Library and a number of museums with collections relating to South African history, military history, medicine, archaeology, geology, costumes, transport, railroads, and Judaica. The Johannesburg Art Gallery exhibits modern European paintings as well as the work of South African artists. Commercial art galleries and antique shops are scattered throughout the city centre and in the northern districts. Branch library services are available in the suburbs and townships.

There are a great number of sporting facilities, including golf courses, tennis courts, rugby and soccer grounds, swimming baths, and cricket grounds, as well as privately owned sports clubs of various kinds and commercial establishments providing facilities for ice hockey, horse racing, and ice skating. The Zoological Gardens have a large collection of wild animals, some fine gardens, and a lake. The larger parks constitute popular picnic areas, and there are numerous open areas in the suburbs. South African wild flowers are to be seen in the Wilds Botanic Gardens, which consists of 45 acres (about 17 hectares) in the northern suburbs; wild birds may be seen at the Melrose Bird Sanctuary. The open space known as the Civic Centre has some fine examples of sculpture; the Civic Theatre is the home of opera, ballet, music, and drama. Throughout the city, many cinemas, halls, and theatres are to be found, as well as a number of recreation centres. There are also a planetarium and some observatories, including the Republic Observatory, which is of international importance. Kelvin House is the headquarters of many of South Africa's scientific and technical societies, including the South African Chemical Institute. A number of societies cater to the interests of photographers, philatelists, chess players, horticulturists, students of Jewish affairs, archaeologists, and other specialized groups.

History

EARLY SETTLEMENT

The discovery of gold

Johannesburg was founded as a result of the discovery of gold in the Witwatersrand area of the Transvaal in 1886. The farms from Driefontein in the east to Roodepoort in the west were proclaimed as public diggings, and gold-mining operations began. The centre of contemporary Johannesburg now stands on land that was then government ground and was known as Randjeslaagte or Rantjeslaagte. There has been much controversy about the origin of the name Johannesburg; the earliest known official statement about the Johannes, for whom the city was named, dated from 1896, when it was recorded that it was called after Johann Rissik, acting surveyor general, and Christian Johannes Joubert, head of the mines department of the Zuid-Afrikaansche Republiek (South African Republic), in whose territory the goldfields were located. The two men had been appointed as commissioners to investigate the mining situation. By November 1886, according to the provisions of the act known as the Gold Law, the first members of a Goldfield's Diggers Committee were elected and became the city's first local government authority. This body was later called the Sanitary Committee but by 1897 had become a Stadsraad, or Town Council. In 1887, the year after the discovery of gold, the mining syndicates—forerunners of the mining companies of today—had begun to operate in Johannesburg; and houses, offices, churches, and shops had begun to spring up.

In 1889 the new city's expansion was halted when the miners reached a zone of pyrite (an ore of sulfur, associated with gold and sometimes called fool's gold) from which it was not known how to extract the gold. As a result, a number of mines closed down before three Scotsmen from Glasgow, the brothers R.W. Forrest and W. Forrest and J.S. MacArthur, discovered the cyanide method of gold extraction. This process saved the Rand, and Johannesburg's growth was resumed.

Although postal facilities of a sort were available from the outset and a telegraph office was opened in 1887, the first telephone service was not installed until 1894. Goods had to be carried in by ox wagon, and passengers were transported by coach and horses; the only railroad in the locality, from Johannesburg to nearby Boksburg, was primarily used for the transport of coal to the mines. In 1892, however, the city was linked by rail to the port of Cape Town, some 800 miles to the southwest. By 1896 the inhabitants of Johannesburg, most of whom were English-speaking and did not have the franchise, came to wish for representation in the government of the country. As a result, differences between the rulers of the country in Pretoria and the mining men in Johannesburg became acute. When the South African War between the United Kingdom and the Zuid-Afrikaansche Republiek broke out in 1899, many residents left Johannesburg. By agreement, the mines were handed over undamaged when the British took the city in 1900. At the conclusion of the war in 1902, by which time the city's population numbered 100,000, labour for the mines was so scarce that Chinese workers were imported. Due to political pressures, however, the Chinese were all repatriated by 1910, the year in which Johannesburg, together with the remainder of the Transvaal, became a part of the Union of South Africa.

THE MODERN CITY

After World War I, labour unrest occasioned various strikes in the mines. In a strike of 1922, white miners opposed the use of Africans for semiskilled work. The strike was suppressed but at a cost of more than 200 lives. In 1928 Johannesburg was declared to be a city; it celebrated its 50th anniversary in 1936. World War II checked further development from 1939 to 1945, after which a building boom occurred, with urban expansion taking place to the north and northwest. By 1970 the city's municipal boundaries were extended to include an area of 104 square miles.

Expansion of the city after World War II

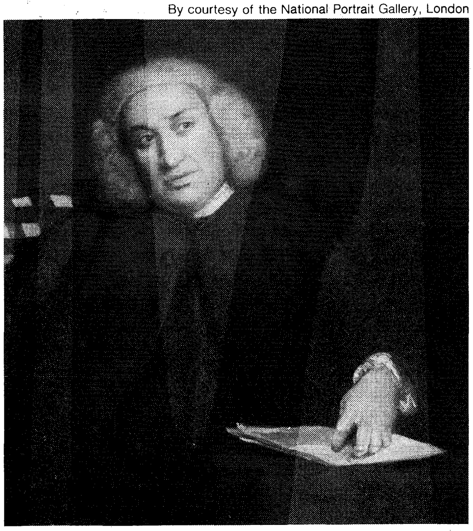
BIBLIOGRAPHY. A.H. SMITH (ed.), *Pictorial History of Johannesburg* (1956); H.A. CHILVERS, *Out of the Crucible: Being the Romantic Story of the Witwatersrand Goldfields and of the Great City Which Arose in Their Midst,* and an *Epilogue, 1929–48: The Incredible City,* by ALEXANDER CAMPBELL (1948); JOHANNESBURG, CITY COUNCIL, *The City of Johannesburg: Official Guide,* 3rd ed. (1962), fairly complete, but somewhat dated; JAMES GRAY, *Payable Gold . . . An Intimate Record of the History of the Discovery of the Payable Witwatersrand Goldfields and of Johannesburg in 1886 and 1887* (1937), and with ETHEL L. GRAY, *A History of the Discovery of the Witwatersrand Goldfields* (1940), a sequel to *Payable Gold . . . ;* JOHANNESBURG, CITY TREASURER'S DEPARTMENT, *Vade-Mecum,* 53rd ed. (1983), gives official factual information, largely municipal, arranged alphabetically by subject; L.E. NEAME, *City Built on Gold* (1970), a historical approach largely based on information from the daily newspaper *The Star,* which dates from 1887; ERIC ROSENTHAL, *Gold! Gold! Gold!: The Johannesburg Gold Rush* (1970), a journalist's account of the history and development of the city and its mines; ALICE M. RALLS and RUTH E. GORDON, *Daughter of Yesterday* (1975), a popular written history of early Johannesburg; ARNOLD BENJAMIN, *Lost Johannesburg* (1979), a look at historic buildings and sites.

(A.H.S.)

Samuel Johnson

The English poet, critic, essayist, and lexicographer Samuel Johnson became famous not only for his writings but also for his forceful, witty conversation. After Shakespeare, Johnson is possibly the best known figure and the most frequently quoted in the whole range of English literature.

By courtesy of the National Portrait Gallery, London



Dr. Johnson, detail of an oil painting by Sir Joshua Reynolds, 1756. In the National Portrait Gallery, London.

EARLY LIFE AND INFLUENCES

He was born at Lichfield, Staffordshire, on September 18 (new style; September 7, old style), 1709. His father, Michael Johnson, was a prominent citizen of Lichfield and was sheriff of the city at the time of Samuel's birth. As a bookseller, he conducted a substantial but not very profitable business. Samuel was not a healthy child. His eyes were weak and he was the victim of a tubercular infection in the glands of the neck, commonly known then as "the King's Evil." In the hope that the cure for this disease lay in the royal touch, Mrs. Johnson travelled to London with Samuel in March 1712, and the boy was duly touched by Queen Anne. His memories of the ceremony were naturally slight, but he retained "a sort of solemn recollection of a lady in diamonds and a long black hood." It is not recorded that he derived any physical benefit from the royal touch, but the gold amulet that the Queen hung round his neck remained there until his death.

Michael Johnson was a high churchman with Jacobite sympathies, favouring the Stuart rather than the Hanoverian succession. His wife, Sarah Ford, was a devout woman with leanings toward Calvinism. She taught her son to learn the collect for the day by heart and expounded to him the contrast between heaven and hell. In 1717 Johnson entered Lichfield grammar school and began the study of Latin. One of his schoolfellows, Edmund Hector, who was to become his lifelong friend, recalling in later years Johnson's "uncommon abilities for learning," wrote: "His ambition to excel was great, though his application to books . . . was very trifling . . . his dislike to business was so great that he would procrastinate his exercises to the last hour." It was a disposition that remained with Johnson throughout his long literary career.

When he was promoted to the upper school, Johnson came under the discipline of the headmaster, a scholar but a tyrant who beat his boys indiscriminately—"to save them from the gallows." As Johnson later said, "My master whipt me very well. Without that, Sir, I should have done nothing." After a brief period at the grammar school at Stourbridge, where he was a student and also took some part in the teaching of the younger boys, Johnson helped his father in the bookshop. Rambling along his father's shelves, he read widely and with the instinct of an incipient scholar—"not voyages and travels, but all literature, Sir, all ancient writers, all manly . . ." Thus, by the time that he entered Pembroke College, Oxford, in 1728, he was familiar with many works unknown at the universities; and when he was first introduced to William Jorden, tutor of the college, his own contribution to the conversation was a quotation from the Latin grammarian and philosopher Macrobius. Johnson thought little of Jorden's scholarship but praised his goodness of heart. William Adams, fellow and later master of Pembroke, meant more to Johnson and remained his friend through many years. He had some good friends, too, among the undergraduates of his own generation. He was near to his old schoolfellow, John Taylor of Ashbourne, at Christ Church, and among Pembroke men he had the reputation of being gay and frolicsome. But it was a forced gaiety. Frustrated and embittered by poverty, he defied authority and "thought to fight his way by his literature and his wit." One example of his scholastic facility survives from his undergraduate days—a Latin translation of Alexander Pope's "Messiah," which was included in a *Miscellany* published at Oxford in 1731.

Friends and influences at Oxford

How the impecunious Michael Johnson was enabled to send his son to college is not wholly clear. Possibly a small legacy received by Mrs. Johnson may have helped. But, in any event, Johnson was compelled to leave Oxford in December 1729 after a residence of 13 months. His prospects were poor. He had no degree or other qualification; his father's business was declining; an application for an ushership (assistant teacher) at Stourbridge was unsuccessful. His father died at the end of 1731, and in the following year Samuel accepted a post as undermaster in the grammar school at Market Bosworth. The work brought him neither health nor happiness, and he resented the arrogance with which he was treated by Sir Wolstan Dixie, in whose house he lived. Through the influence of his old friend Edmund Hector, who had become a surgeon in Birmingham, Johnson secured the task of translating into English the French version of *A Voyage to Abyssinia* by Father Jerome Lobo. Johnson's preface gives an early indication of his instinctive sympathy with the natives of an invaded country. The invaders of Abyssinia were missionaries who preached the gospel with swords in their hands. The preface also gives an authentic foretaste of Johnson's prose style:

> The Reader . . . will discover, what will always be discover'd by a diligent and impartial Enquirer, that wherever Human Nature is to be found, there is a mixture of Vice and Virtue, a contest of Passion and Reason, and that the Creator doth not appear Partial in his Distributions, but has balanced in most Countries their particular Inconveniences by particular Favours.

Marriage and early literary endeavours

Johnson's sojourn in Birmingham brought him something more than a fee of five guineas for his first book—it brought him a wife in the person of Elizabeth, widow of Harry Porter, a mercer. She was 20 years older than Johnson, and it is not easy to determine the grounds of mutual attachment. Johnson, in later years, used to speak of his wife's beautiful blonde hair; and whatever the lady thought of Johnson's looks ("lean and lank . . . the scars of the scrofula deeply visible") or of his "convulsive starts and odd gesticulations," she at least appreciated the vigour and good sense of his conversation. The marriage took place at Derby in 1735, and the bride brought with her a reputed dowry of £700. On the strength of this and with the encouragement of a friend who was registrar of the ecclesiastical court of Lichfield, Johnson decided to set up at Edial, near Lichfield, a school on his own at which young gentlemen could be boarded and taught the Latin and Greek languages. He prepared an elaborate curriculum, but only a few young gentlemen, among whom was David Garrick, came as pupils, and at the end of two years Johnson had to admit failure.

But in spite of ill health, of melancholia, of his lack of a degree, and of the collapse of his school, the desire to be known as a scholar and a writer remained clear in his mind. While still a schoolboy he had written of

> . . . the young Authour, panting after fame,
> And the long honours of a lasting name

and in the year before his marriage he had written to the founder of the *Gentleman's Magazine,* offering to fill a column, on reasonable terms, with poetry, dissertations, critical remarks on ancient and modern authors, and other material. Furthermore, with plenty of time on his hands at Edial, he embarked, again with the encouragement of friends, on the writing of a tragedy, based on the story of Sultan Mahomet (Mehmed) II and the beautiful Greek maiden Irene, as told by Richard Knolles in his *General History of the Turks.* But the work was not finished, and Johnson, facing the fact that he must write something for which an editor or a bookseller would pay, decided to seek his fortune in London. So, in company with David Garrick, he rode to London in March 1737. Arthur Murphy, in "An Essay on the Life and Genius of Samuel Johnson LL.D." (1792), wrote:

> Two such candidates for fame perhaps never, before that day, entered the metropolis together. . . . They brought with them genius, and powers of mind, peculiarly formed by nature for the different vocations to which each of them felt himself inclined. . . . In three or four years afterwards Garrick came forth with talents that astonished the publick. . . . Johnson was left to toil in the humble walks of literature.

Contributions to the *Gentleman's Magazine*

The *Gentleman's Magazine* offered him the first opportunity of humble toil. Edward Cave, the publisher, quickly recognized his journalistic ability, and Johnson contributed a number of pieces in prose and verse—odes, epigrams, reviews, as well as a series of concise biographies. In the later part of 1737 he returned to Lichfield, finished his tragedy *Irene,* and brought his wife to London. Meanwhile, his ambition to be a writer of something more than ephemeral pieces for periodicals remained, and in 1738 his first substantial poem, *London,* was published. It was written in imitation of the third *Satire* of Juvenal, and in it Johnson embodied his protest against political corruption:

> Here let those reign, whom pensions can incite
> To vote a patriot black, a courtier white;

against the dangers of the London streets:

> Their ambush here relentless ruffians lay,
> And here the fell attorney prowls for prey;
> Here falling houses thunder on your head,
> And here a female Atheist talks you dead.

and, with more acutely personal feeling, against the miseries of the unknown and impecunious author:

> This mournful truth is ev'ry where confess'd,
> SLOW RISES WORTH, BY POVERTY DEPRESS'D.

London, published anonymously, had an immediate success. It went quickly into three editions and won high praise from Alexander Pope. But Johnson's fee was only 10 guineas, and again he thought of schoolmastering as an alternative to being "starved to death in translating for booksellers." The headmastership of Appleby School in Leicestershire was offered to him, subject to his obtaining the degree of M.A., but negotiations for its conferment broke down both at Oxford and at Dublin; similarly, his lack of a degree in law frustrated his application for permission to practice as an advocate. To this period of embittered disappointment belong his two most violent and satirical strictures upon the government of Walpole: the first was *Marmor Norfolciense* (1739), an essay upon a Latin rhyme supposed to have been discovered in Walpole's county, Norfolk; and the second, *A Compleat Vindication of the Licensers of the Stage* (1739), an ironical defense of the suppression of Henry Brooke's play *Gustavus Vasa.* Both satires are the protests of an angry young man rebelling against authority and striving, as he had striven at Oxford, to fight his way out by his wit. Johnson was no sentimental Jacobite, but his scorn of the Hanoverian government was never more bitterly explicit than in *Marmor Norfolciense:*

Satires against the Walpole ministry

> Then o'er the World shall Discord stretch her wings;
> Kings change their Laws, and Kingdoms change their
> Kings.

The story that a warrant was issued for Johnson's arrest has never been verified, but, meanwhile, as a member of Cave's staff, he was required to treat contemporary politics in more sober style. Reports of parliamentary debates had been a feature of the *Gentleman's Magazine* since 1732, but shortly after Johnson's arrival in London the House of Commons forbade publication of its proceedings. Cave, however, contrived to continue the publication as "Debates in the Senate of Magna Lilliputia"; and in the first instance Johnson was employed to assist in editing and expanding the reports, but from 1740 to 1743 the "Debates" were entirely Johnson's own work. He was not a reporter in the modern sense and was only once inside the House of Commons. Sometimes he had a few notes supplied by other reporters, sometimes nothing more than the subject of debate and the names of the speakers; sometimes, he confessed, the "Debates" were the mere coinage of his own imagination, and in later years he had some prickings of conscience about his freedom of invention. As a journalistic feat, they were a remarkable tour de force. He would shut himself up in a room at Cave's headquarters in St. John's Gate and deliver three columns for the *Magazine* in an hour. His often-quoted remark that he took care that the Whig dogs—members of the political party opposed to the Tories—should not have the best of it gives a false impression of what he was doing. His reports have been well described as leading articles on both sides of the question, rather than records of the cut and thrust of debate. The short speech put into the mouth of Walpole at the end of the debate on the motion for his removal from office is a remarkably objective record of a minister speaking with dignity and restraint in his own defense.

Friendship with Richard Savage

One of Johnson's closest companions in his early years in London was Richard Savage. Savage had in his time been actor, playwright, and poet. He claimed to have been nobly born and had had many friends among the great whose hospitality he persistently abused. Two qualities at least he shared with Johnson—poverty and patriotic indignation against the Walpole administration. When their fortunes were at their lowest, they would spend whole nights in "a perambulation round the squares of Westminster . . . when all the money they could both raise was less than sufficient to purchase for them the shelter and sordid comforts of a night cellar." When Savage died, Johnson lost no time in commemorating his friend. Written con amore and at a white heat, the *Account of the Life of Mr Richard Savage,* published anonymously in 1744, was the first of Johnson's prose works to captivate the public. The novelist Henry Fielding, who would publish *Tom Jones* five years later, called it the best treatise in the language on the excellencies and defects of human nature; Joshua Reynolds, then only 21 but destined to become one of England's greatest painters, could not put the book down until he had finished it.

RECOGNITION AND MATURE CAREER

The theatre. In addition to a great variety of hack work, Johnson was now turning his attention to Shakespeare. In 1745 he published his *Miscellaneous Observations on the Tragedy of Macbeth,* coupled with preliminary proposals for his own edition of the plays, but he was deflected from further Shakespearean work by the suggestion that he should compile a dictionary of the English language. That a syndicate of booksellers should have chosen Johnson for so gigantic a task is a tribute to the position he had made for himself during his nine years in London.

In those same years David Garrick had made more rapid and more brilliant progress. Deserting the law for the theatre, he had made his mark as an actor in 1741 and by 1747 had become the manager of Drury Lane Theatre. For the first performances under Garrick's management Johnson wrote a prologue, and, although he commonly spoke with contempt of actors and their profession, he was willing, for friendship's sake, to plead their cause:

> Ah let not Censure term our Fate our Choice,
> The Stage but echoes back the publick Voice,
> The Drama's Laws the Drama's Patrons give,
> For we that live to please, must please to live.

At this time, moreover, Johnson had a personal interest in dramatic production. His tragedy *Irene* had long lacked a publisher or a producer. Garrick, when he came into power, agreed to do his best for his old master, and *Irene* was produced "with a display of Eastern magnificence" in 1749. It ran for nine nights. It was essentially a moralist's play, and although it was said to be universally admired, it has never been revived.

The Vanity of Human Wishes

More permanent was Johnson's second didactic poem, *The Vanity of Human Wishes* (1749), in which the careers of Galileo, Wolsey, Charles XII of Sweden, and others are shown to illustrate the hazards of political ambition, the futility of military conquest, and the miseries of authorship:

> There mark what ills the scholar's life assail,
> Toil, envy, want, the patron, and the jail.
> See nations, slowly wise and meanly just,
> To buried merit raise the tardy bust.
> If dreams yet flatter, once again attend,
> Hear Lydiat's life, and Galileo's end.

Though at the time of its publication *The Vanity of Human Wishes* had a much less rapid sale than *London,* it is Johnson's greatest poem. Its manner is that of its period. But Johnson's panorama of the rise and fall of scholars and philosophers, of statesmen and kings, in the modern as well as the ancient world, is inspired by that "high seriousness" that endows the poem with universality. But the record of disillusionment is not the end.

> Must helpless man, in ignorance sedate,
> Roll darkling down the torrent of his fate?

No, says Johnson, he must pray for the love and patience and faith of the Christian.

The "Rambler." The contract for *A Dictionary of the English Language* had been signed in 1746, and Johnson published his *Plan of a Dictionary of the English Language* the following year. He made no complaint of what the booksellers offered him, but it was not enough to keep the wolf from the door. So, in 1750, he embarked upon the *Rambler,* a twopenny sheet published twice a week and containing a single anonymous essay. The *Rambler* is of fundamental importance in any estimate of Johnson's approach to literature. Before he embarked on the work he prayed that he might promote the glory of God and the salvation of himself and others. The Rambler, in short, was not an entertainer but an instructor, and 19th-century critics tended to dismiss his essays as lay sermons.

Johnson had recently founded his first club (the Ivy Lane Club) and from his tavern chair would take a prominent part in contributing to what he called "colloquial entertainment." But a printed essay demanded "more accurate thought and more laboured beauties." Talking for victory was legitimate; with the printed word came the moralist's responsibility, and in Johnson's literary creed the basic article of belief was that it was always a writer's duty to make the world better. This sense of responsibility determined the style of the essays. He created no character comparable to that of Addison and Steele's Sir Roger de Coverley in the *Spectator,* and the essays bear little relation to current events or current literature. On the other hand, they frequently reflect the social and literary conditions of the time: there is scorn for the virtuoso and the overdomesticated hostess; an objective picture of the prostitute's life; a vigorous protest against the death sentence for robbery; and, inevitably, a grimly humorous presentation of the journalist's lot. Moreover, they show a remarkable understanding of human frustration and blocked wills. As has recently been pointed out, in the writings of Johnson there is the closest anticipation of the theories of Freud, if not his language, before the 20th century.

The *Rambler* appeared twice a week for two years (1750–52). A few days after the issue of the last number, Johnson's constitutional melancholy was deepened by the death of his wife. The full story of his married life must necessarily be conjectural. None of Johnson's friends in his early London period appears to have met Mrs. Johnson, and there are but few contemporary references to her. That the marriage was based on mutual admiration and affection is reasonably clear. From the beginning Mrs. Johnson had relished the quality of her husband's conversation, and she had a special admiration for the *Rambler;* Johnson, on his part, appreciated her intelligent reading of comedy. But, domestically, they were not well suited. Johnson was as insistent about the quality of his food as he was careless and untidy about the house. What was more serious was that in her later years Mrs. Johnson became addicted to strong liquor and to drugs and was unwilling to satisfy her husband's physical desires. Something of this may be read between the lines of Johnson's prayers and meditations about her death. Nevertheless, Johnson's affection, tinged with some remorse, remained sincere; and its sincerity is not impaired by the discovery that on April 22, 1753, he purposed "to try . . . to seek a new wife without any derogation from dear Tetty's memory." The precise direction of his search is not known, but it is interesting that he contemplated an adventure that he was later to describe as the triumph of hope over experience. "Formosae, cultae, ingeniosae, piae" ("a woman of beauty, elegance, ingenuity, and piety")—so Johnson described his Tetty on the gravestone that he placed in Bromley Parish Church more than 30 years later. In lapidary inscriptions, as he said, a man is not upon oath.

Mrs. Johnson's death

In 1752 Johnson was a lonely man, but he had made some good friends. At the Ivy Lane Club he was delighted "to pass those hours in a free and unrestricted interchange of sentiments, which otherwise had been spent at home in painful reflection." There were 10 members, including Sir John Hawkins, who lived to be Johnson's executor and biographer; John Hawkesworth, editor of the *Adventurer,* to which Johnson himself contributed a number of essays; John Ryland, one of the few of Johnson's early friends who lived to attend his funeral; John Payne, publisher of the *Rambler;* and Richard Bathurst, the physician whom Johnson loved better than any other creature. It was at the Ivy Lane Club that Johnson celebrated the publication of the first book of his friend Mrs. Charlotte Lennox by "a whole night spent in festivity." About five in the morning, according to Hawkins, "Johnson's face shone with meridian splendour, though his drink had been only lemonade."

The dictionary. In 1747 Johnson published his *Plan of a Dictionary of the English Language.* At the suggestion of Robert Dodsley, the author and bookseller, it was dedicated to Lord Chesterfield. At first Chesterfield showed some interest and made some suggestions for revision, but when he paid no further attention to the work, Johnson would not be obsequious—and Johnson did not forget. In April 1753 he was beginning work upon the second volume of the *Dictionary,* and he had still to write the preface, the grammar, and the history.

Two years later the work was finished. It had occupied Johnson and his amanuenses for eight and one-half years, and its accomplishment was described by Sir James Murray, editor in chief of the *Oxford English Dictionary,* as a marvellous one. It surpassed earlier dictionaries not in bulk but in precision of definition and in literary illustration. Its 40,000 words were, in fact, rather less than those in the work of his predecessor, Nathaniel Bailey; but what distinguished Johnson's work was the range of reading by which he exemplified the different shades of the meaning of a particular word. "I applied myself," he wrote in his preface, "to the perusal of our writers . . . noting whatever might be of use to ascertain or illustrate any word or phrase." Certain books (Bacon's *Essays,* South's *Sermons,* and others) have survived with Johnson's underlinings and indications of the passages chosen for quotation. His assistants copied the quotations on separate slips, and these were pasted below Johnson's own definitions. At the beginning Johnson had looked forward to hours that he would revel away in feasts of literature and ransack the obscure recesses of northern learning. But he soon realized that these were "the dreams of a poet doomed at last to wake a lexicographer." To the weariness of copying was

Character and reputation of the *Dictionary* added the vexation of expunging, and Johnson, having set practical limits to his work, was acutely conscious of its imperfections. His orthography was admittedly controvertible and his etymology uncertain; and, even while the dictionary was hastening to publication, some words, as he said, were budding and some were falling away. Nevertheless, the claim made in the final paragraph of his preface—one of the finest examples of his prose style—was abundantly justified:

> In this work, when it shall be found that much is omitted, let it not be forgotten that much likewise is performed; and though no book was ever spared out of tenderness to the author, and the world is little solicitous to know whence proceeded the faults of that which it condemns; yet it may gratify curiosity to inform it, that the *English Dictionary* was written with little assistance of the learned, and without any patronage of the great; not in the soft obscurities of retirement, or under the shelter of academick bowers, but amidst inconvenience and distraction, in sickness and in sorrow.

Today the ordinary reader may tend to remember the few "wild blunders"; the "risible absurdities"; and the frankly personal prejudices shown in the definitions of "oats" ("A grain, which in England is generally given to horses, but in Scotland supports the people"), "excise," "pension," and other well-known examples. But Johnson's enrichment of the wordbooks of his predecessors by judicious definition and by linguistic illustration from English literature was an enduring monument in the history of lexicography.

On the title page of the *Dictionary,* Johnson had the satisfaction of describing himself as a Master of Arts of the University of Oxford, and it is significant that the award was made in consideration of the religious and moral value of his essays. The degree, in short, was conferred upon the Rambler.

The *Dictionary* was well received, and opposition came only from "the Criticks of the coffeehouse whose outcries are soon dispersed in the air and are thought on no more." But there had been one outcry of enthusiastic praise that Johnson did not allow to be dispersed. Shortly before the work appeared, in two papers in the *World,* Lord Chesterfield, seeking to make amends for his previous neglect, hailed Johnson as the supreme dictator of the English language and so provoked the most famous of all Johnson's letters:

> The notice which you have been pleased to take of my labours, had it been early, had been kind; but it has been delayed till I am indifferent, and cannot enjoy it; till I am solitary, and cannot impart it; till I am known, and do not want it. . . .

Journalism. Johnson was soon at work upon an abridged edition of the *Dictionary,* but fame as a lexicographer had not relieved him of pecuniary distress. In March 1756 he was under arrest for the sum of £5 18*s.* Normally he would have appealed to one of the printers for whom he worked, but they were not available, and it was the printer-novelist Samuel Richardson who sent him six guineas. So, of necessity, his activity as a journalist continued. He edited Sir Thomas Browne's *Christian Morals* (1756), wrote prefaces to William Payne's *Introduction to the Game of Draughts* (1756) and Richard Rolt's *New Dictionary of Trade and Commerce* (1756), and contributed many articles to journals. In the *Literary Magazine,* Jonas Hanway's "Essay on Tea" provoked Johnson's description of himself as a "hardened and shameless tea-drinker" who had drunk of it for 20 years without hurt and therefore believed it not to be poison.

A more important review was that of Soame Jenyns' *Free Inquiry into the Nature and Origin of Evil* (1757), in which, with massive irony, Johnson demolished the conjecture of a superior race of beings deceiving and tormenting men for their own pleasure:

> The only end of writing is to enable the readers better to enjoy life, or better to endure it: and how will either of those be put more in our power by him who tells us, that we are puppets, of which some creature not much wiser than ourselves manages the wires.

Nor could Johnson swallow the bland assertion that poverty was generally compensated by better health and a more exquisite relish of the smallest enjoyments. "Life," he retorted, "must be seen before it can be known." These were topics on which he wrote with peculiar depth of feeling, and, as always when he felt deeply, his style became more simple.

Political writings

Neither Johnson's interests nor his writings were confined to the problems of literature and ethics. To the *Literary Magazine* of 1756 he contributed two articles on the political situation—"An Introduction to the Political State of Great Britain" and "Observations on the present State of Affairs." In both he writes with scorn of the power politics inherent in both English and French colonialism. The dispute between the two countries in America was "the quarrel of two robbers for the spoil of a passenger," and even those who had settled in the New World on the fairest terms had no other merit than that of "a scrivener who ruins in silence over a plunderer that seizes by force." Nor was he afraid of asserting that the French sent out better governors and that they treated the natives better than did the English. It was ridiculous, he wrote, to imagine that the friendship of nations, whether civil or barbarous, could be gained or kept but by kind treatment; and it was this basic mistrust of the motives and methods of colonizers that provoked his later and better known outburst against colonial claims. Meanwhile, he had projects of his own in hand or in view. He issued his proposals for his edition of Shakespeare in 1756, emphasizing, not for the first time, that his motive was not desire of fame but want of money. In April 1757 he wrote to his old friend Edmund Hector, in Birmingham, that the subscriptions, if slightly disappointing, were satisfactory, but early in 1758 he was obliged to borrow £40 from a bookseller.

"The Idler" essays

About the same time he undertook to contribute a weekly essay, to be entitled "The Idler," to the *Universal Chronicle.* Of these essays it may be said that the reader who has been nurtured in the tradition that while Johnson's talk is magnificent, his writings (other than the *Lives of the English Poets*) are unreadable, would be well advised to turn his attention to "The Idler." Though the moralist and the social reformer (especially on such topics as debtors' prisons and vivisection) are still evident, he is willing to turn aside to the human comedy: for instance, to the female bargain hunter ("whatever she thinks cheap, she holds it the duty of an economist to buy") or to the publisher ("Some never dealt with authors; others had their hands full; some had never known such a dead time; others had lost by all that they had published for the last twelvemonth"). There are also some character sketches, of which the most memorable is that of "Mr. Sober":

> Mr. Sober's chief pleasure is conversation; there is no end of his talk or his attention; to speak or to hear is equally pleasing; for he still fancies that he is teaching or learning something, and is free for the time from his own reproaches.
>
> But there is one time at night when he must go home, that his friends may sleep; and another time in the morning, when all the world agrees to shut out interruption. These are the moments of which poor Sober trembles at the thought.

It is a rare and convincing piece of self-portraiture. Every week for two years, with a few exceptions, Johnson delivered his "Idler" essay to the printer.

Rasselas. In the midst of his work and worry came news of his mother's illness. On January 13, 1759, he contrived to send her 12 guineas. But he knew that he would need more, and on the same day that he wrote his last tender tribute to his mother (January 20), he implored Strahan, the printer, to let him have £30 on account of "a thing he was preparing for the press." The thing was to be entitled *The Prince of Abissinia,* better known as *Rasselas.* Written in the evenings of a week with the impending expenses of his mother's funeral in mind, it explores and exposes the vanity of the human search for happiness. The setting was no doubt prompted by Johnson's recollection of Jerome Lobo's *Voyage to Abyssinia,* and the work is addressed to those who "listen with credulity to the whispers of fancy and pursue with eagerness the phantoms of hope." Impelled by such eagerness, Rasselas, with his sister, leaves his happy valley because its pleasure has ceased to please and because he is fired with a desire to do something. He meets with men of varied occupations and interests and earnestly explores their manner of life—scholars, astronomers, shepherds, hermits, poets. With Imlac, the poet, he ranges over many of the basic problems of art and life but gains little satisfaction from the answers he receives. At his first entry into city life, Rasselas meets everywhere with gaiety and happiness, but closer association reveals a picture of levity and intemperance. From the extravagance of youth he turns to an inspiring lecture on morality, only to find that the lecturer's philosophy collapses at the first stroke of personal misfortune. When he passes hopefully to scenes of pastoral simplicity, he finds the shepherds cankered with discontent, since they are condemned to labour for the luxury of the rich; when he seeks advice from a hermit, he is disappointed to be told that "the life of a solitary man will be certainly miserable, but not certainly devout." So the journey of disillusionment continues, until Imlac protests: "While you are making the choice of life, you neglect to live"—which is, perhaps, the most important moral to be drawn from the tale.

The moral of *Rasselas*

Of all Johnson's writings, none is more intensely characteristic than *Rasselas.* It is a remarkable example of his fluent productivity when he had a definite object and a definite date before him, but speed did not debase either the solemnity of his subject or the dignity of its treatment. Johnson allowed his imagination to wander into Ethiopia and Egypt, but fundamentally *Rasselas* is a spiritual autobiography. Furthermore, it was the one prose work of Johnson's that obtained an immediate and wide popularity in his lifetime. It satisfied the taste of the 18th-century reader for "impressive truth in splendid fiction drest"—and not only of the English reader. In a variety of translations its fame spread over Europe and beyond.

Johnson's circle. In spite of his struggles and sorrows, Johnson was, by this time, no longer a lonely hack writer. His writings, though they did not make his fortune, had brought him many friends. Joshua Reynolds had been enthralled by Johnson's *Life of Savage* and was no less enthusiastic about his conversation and his counsel; Charles Burney, a leading English music historian of his time, warmly commended the *Dictionary* and encouraged the edition of Shakespeare; Bennet Langton, a Lincolnshire gentleman and scholar, came to London at an early age for the express purpose of meeting the author of the *Rambler.* Johnson also cherished his Oxford friendships. He spent some time there in 1754 and again in 1755 and 1759 and, after the conferment of his degree, took a keen pleasure in wearing his gown. It was at Trinity in 1759 that Bennet Langton introduced him to Topham Beauclerk, who was descended from Charles II and Nell Gwyn.

Introductions, reviews, and dedications

Even so, Johnson was still at the beck and call of authors, editors, and friends for the writing of introductions, reviews, and, especially, dedications. Of James Bennet's edition of Roger Ascham's works he was virtually the editor and contributed a life of Ascham, a famous scholar and Latin secretary to Elizabeth I; he wrote the loyal address presented to George III on his accession by the painters, sculptors, and architects; he revised a pamphlet on the proper route for the coronation procession; he reviewed a work on Mary, queen of Scots, and wrote a dedication for Giuseppe Baretti's *Italian Dictionary.* Meanwhile his edition of Shakespeare tarried. In 1762 the unexpected happened. He was informed of the gracious intention of his majesty the King to confer on him a pension of £300 a year. Could he in decency take it, after defining "pension" in his *Dictionary* as "pay given to a state hireling for treason to his country"? At least, he felt, he must consult his friends. But Reynolds reassured him, and Lord Bute, the prime minister, told him that the award was made not for anything he might do but for what he had already done. So, from a full heart, Johnson thanked his lordship for sparing him "the shame of solicitation, and the anxiety of suspense."

Meeting with James Boswell. The following year was to provide yet another landmark in Johnson's life, for on May 16, 1763, his accidental meeting with James Boswell in the back parlour of Thomas Davies' bookshop in Covent Garden inaugurated one of the most famous companionships in history. Boswell, the eldest son of Lord Auchinleck, a Scottish judge, had studied law at Edinburgh and Glasgow and had a passionate desire to taste the felicities of London life. His mind was dominated by two ambitions—to meet famous men and to be a famous author himself. Staying in a country house in the previous year, he had read aloud, with enthusiastic comment, some of the *Rambler* essays; and, in the list of celebrities whom he wished to meet, the author of the *Rambler* stood high. A little daunted by Johnson's brusque rejoinders at their first meeting, Boswell was nevertheless encouraged by Davies to persevere; and a week later he waited, with some apologies, upon Johnson. Johnson cut him short. "I am obliged," he said, "to any man who visits me." So Boswell stayed and listened in rapture. "His conversation," he wrote in his diary, "is as great as his writing." He could give no higher praise. To Johnson new friendships were always welcome, and especially with young people (Johnson was then 53 and Boswell only 22), for such friendships lasted longest. To the charm of Boswell's enthusiasm he quickly succumbed. Together they enjoyed suppers at the Mitre Tavern and excursions on the Thames, but the signal mark of Johnson's favour was his offer to accompany Boswell in the coach to Harwich, whence Boswell was to embark on his continental tour. Johnson gave him much sound advice. In particular, he urged him to keep a diary, and it is to Boswell's diary that the world owes its intimate knowledge of Johnson. Many others left valuable records of his life and character, but it was Boswell who Johnsonized the land.

Foundation of The Club

In 1764 Johnson was happy to concur with Joshua Reynolds' suggestion for the foundation of what is still the most famous of London dining clubs—The Club. Among the original members, besides Reynolds and Johnson, were Edmund Burke, Topham Beauclerk, Bennet Langton, and Oliver Goldsmith, and nine years later it was one of the proudest moments of Boswell's life when he was admitted to membership.

Edition of Shakespeare. Meanwhile, Johnson had been working on his edition of Shakespeare. Adumbrated in 1745 and formally announced in 1756, it had occupied him for much more than the two years in which he had hoped to complete it, and the long delay had provoked some typical satire from the pen of the scurrilous poet Charles Churchill:

> He for subscribers baits his hook,
> And takes their cash—but where's the book?

The book, in eight volumes, appeared at length in 1765. Johnson had been a student of Shakespeare all his life. He was no idolater, and his basic criticism is that of the moralist:

> He is so much more careful to please than to instruct, that he seems to write without any moral purpose . . . . He carries his persons indifferently through right and wrong, and at the close dismisses them without further care, and leaves their examples to operate by chance. This fault the barbarity of his age cannot extenuate; for it is always a writer's duty to make the world better.

As an editor Johnson set out, first, to correct textual corruptions; second, to elucidate obscurities of language; and, last, in treating of Shakespeare's sources, to examine the very books that Shakespeare consulted. Of the critics' complaint of Shakespeare's neglect of the unities of time and place he made short work. The demand for these unities arose from "the supposed necessity of making the drama credible," but in Johnson's view no drama was either credible or credited. "The truth is that the spectators are always in their senses, and know, from the first act to the last, that the stage is only a stage, and that the players are only players." Johnson recognized that Shakespeare wrote his plays not for the reader at his desk but for an audience in the theatre. But Johnson himself could never appreciate the contribution made by the actor to dramatic interpretation. For him the actor was a reciter who said his piece "with just gesture and elegant modulation," and it was not in the theatre but in his study that he was most deeply moved by "the perpetual tumult of indignation, pity and hope." "He that peruses Shakespeare," he wrote about *Macbeth,* "looks round alarmed and starts to find himself alone," and he was so deeply shocked by his first reading of Cordelia's death in *King Lear* that he could not bear to read it again until he revised the play as an editor. The romantic critics of the 19th century belittled Johnson's work on Shakespeare, but the prophecy by the English critic Sir Walter Raleigh (1861–1922) that Johnson would receive more respect in the 20th century has been abundantly fulfilled by modern editors. It was in the year of the publication of his Shakespeare that Johnson received the degree of LL.D. from Trinity College, Dublin; his own university did him a similar honour 10 years later.

The house in Gough Square

Johnson's household. In whatever literary work Johnson was engaged his greatest terror was solitude, and the composition of his household reflects his efforts to avoid it. He had many habitations in London of which the most famous is the house in Gough Square, just north of Fleet Street, in which he lived from 1749 to 1759. When Boswell first called upon him in 1763 he was at 1, Inner Temple Lane. From there he moved back to Johnson's Court (7, Fleet Street) in 1765 and in 1776 to Bolt Court, which was his home until his death. The house in Gough Square has been well preserved and is now a Johnson museum. There the *Dictionary* was compiled; there the *Rambler,* the *Idler,* possibly *Rasselas,* and much else was written; there his wife died in 1752. Even before her death and long before the grant of his pension, he had begun to make his house a refuge for the poor and unfortunate.

Anna Williams, daughter of Zachariah Williams, who had received Johnson's help in writing his *Longitude at Sea,* came to London in the hope of being cured of a cataract but later became totally blind. She was a constant visitor in Mrs. Johnson's lifetime and, after her death, came to live in the house. For the rest of her life she had either a room in Johnson's house or lodgings near at hand. Blindness made her peevish in manner, but Johnson's regard and affection for her remained constant. When she died in 1783, he mourned a companion to whom he had had recourse for domestic amusement for 30 years. Shortly after Mrs. Johnson's death, Johnson's beloved friend Richard Bathurst presented to him Francis Barber, a Negro slave who had been freed by Bathurst's father. Johnson made Francis his friend as well as his servant, sent him to school at Bishop's Stortford, and provided for him handsomely in his will. Another humble friend was Robert Levett, "an obscure practiser in physick amongst the lower people," who was given a room at the top of the house in Johnson's Court. Johnson insisted that Levett was indebted to him for nothing more than houseroom, a share of a penny loaf at breakfast, and an occasional Sunday dinner. His death provoked one of Johnson's most moving poems:

> Well try'd through many a varying year,
> See Levett to the grave descend;
> Officious, innocent, sincere,
> Of every friendless name the friend.

Yet another beneficiary was Mrs. Desmoulins, daughter of Dr. Swynfen, Johnson's godfather, who joined the household in the early London years. After Mrs. Johnson's death she appears to have been in charge of the cooking. But she quarrelled constantly with Anna Williams, and anarchy reigned in the kitchen.

Friendship with the Thrales

It was from this domestic background that in 1765 Johnson was introduced to a family that was to provide one of the most comforting friendships of his life. Henry Thrale, owner of a Southwark brewery, had been married to Hester Lynch Salusbury in 1763. Two years later he was elected member of Parliament for Southwark, and his wife became a famous hostess and, in particular, "the provider and conductress of Dr. Johnson." For the first time in his life Johnson, who was invited not only to dine but to spend weeks, or even months, at the Thrales's country house at Streatham, was free to enjoy the luxury of solid comfort. No man, as he said, is a hypocrite in his pleasures, and at Streatham he could enjoy them all—a good library, intelligent conversation, pretty women, late hours, tasteful cookery. Furthermore, he became the confidant and counsellor of both parents and children; a room was set apart for him at the Southwark house as well as at Streatham; he was one of the family.

Political pamphlets. Already Johnson's pension had made two important differences in his way of life: he was no longer obliged to write for a living, and he could afford the time and money for holidays. Not that he was entirely idle as a writer: Shakespeare and the *Dictionary* called for revision; friends called for dedications and prologues (he wrote one for Goldsmith's *Good-Natur'd Man*); and he was also moved to write a series of political pamphlets. Of these tracts, the general view was, for long, that they exhibited the sad spectacle of a great man giving vent, unworthily, to the abusive expression of his conservative prejudice against the rights of the people. Today, more trouble is taken to understand Johnson's point of view. The pamphlets of the 1770s are seen not as a sudden whim of his later years but as part of the expression of a lifelong concern with political morality, which in the 1730s and 1740s had produced pamphlets vigorously denouncing the alleged tyranny and corruption of the Walpole government and, in the 1750s, a large amount of acute journalistic analysis (and condemnation) of the commercial and imperialist motives underlying the Seven Years' War.

The False Alarm (1770) was written not so much to defend the right of the House of Commons to refuse readmittance to a member already expelled as to protest against the absurdity of raising the controversy to the level of a constitutional crisis. What the country was arguing about, he said, was whether Middlesex should be represented, or not, by a criminal from jail. To raise the cry that liberty was in danger was to raise a false alarm.

Thoughts on the Late Transactions respecting Falkland's Islands (1771) was a reply to the anonymous political controversialist Junius and others, who had been urging the British government to resist the Spanish claim to those islands by aggressive action, and who now reproached it for having avoided war by a diplomatic compromise that left Britain in possession of the islands, though without fully establishing its title to them as against Spain's. Johnson gives a thorough history of the rival claims and of the diplomatic manoeuvring and then, in vigorous rhetoric, denounces the jingoism of his opponents, charging them with warmongering for commercial gain.

> These are the men who, without virtue, labour, or hazard . . . rejoice when obstinacy or ambition adds another year to slaughter and devastation; and laugh from their desks at bravery and science, while they are adding figure to figure, and cipher to cipher, hoping for a new contract from a new armament, and computing the profits of a siege or tempest.

The Patriot (1774), a short piece written just before the election of that year, defines the qualities of the true patriot and contrasts them with the behaviour of those who had unjustifiably arrogated the title to themselves. As for the supposed defenders of liberty of conscience, had they not opposed the act of Parliament that gave the Catholics of Quebec the right to practice their own religion?

Taxation No Tyranny (1775) is a reply to the resolutions passed by the American Continental Congress of 1774, which assert a position similar to that contained in the Declaration of Independence in the year 1776. Here it is important to remember that throughout his life Johnson had shown little sympathy toward European colonization and colonists in general and had published bitter denunciations of the exploitation (and worse) of native populations by the Portuguese in Africa, the Spanish in Central and South America, and the "English barbarians that cultivate the southern islands of America"—the West Indies. In an "Idler" essay (81; 1759) Johnson puts into the mouth of an Indian chief a fierce tirade against the American colonists' oppression of the Indians and Negroes. By comparison, the colonists' complaints of "oppression" by the British government (in the form of taxation) seem to Johnson trivial. For mercenary reasons, they had settled on the other side of the Atlantic under English protection and with the authority of English charters. That they had no representation at Westminster was their own fault. They had the protection of English arms. Why should they not pay for it? And, if there were 3,000,000 Whigs in America "fierce for liberty," why were the loudest yelps for liberty heard among the drivers of Negroes?

For Johnson, these tracts were an opportunity to expound his essentially pragmatic philosophy of politics. He ascribed divine right neither to kings nor to people. Talk

Johnson's pragmatic philosophy of politics

about liberty in the abstract or about "natural" rights he dismissed as cant. What was essential for a civilized community was a stable government and respect for its laws. But he clearly recognized that there was an ultimate safeguard. If government abused its power, mankind would not bear it; against a tyrant the people would rise and cut off his head.

The politics of the 1770s did not deter Johnson from the enjoyment of his newfound liberties. With the Thrales he spent weeks at Brighton as well as at Streatham, and his "annual midland ramble" included long visits to Oxford, to Lichfield, and to his old friend John Taylor at Ashbourne.

Journey to the western islands of Scotland. In 1773 Boswell persuaded Johnson to accompany him on a more exciting expedition. Johnson had wanted to visit the Hebrides, the western islands of Scotland, for longer than he could remember and now felt that Boswell's "gaiety of conversation and civility of manners were sufficient to counteract the inconveniences of travel." The travellers left Edinburgh on August 18 and followed the coast road. At St. Andrews Johnson's sadness at the sight of archiepiscopal ruins was mitigated by the kindness of the professors; at Aberdeen he was made a freeman of the city; dining with the governor of Ft. George, he talked learnedly about gunpowder. After Inverness, horses were substituted for the post chaise, and by the shore of Loch Ness Johnson encountered an old woman, who could speak very little English, boiling goat's flesh in a kettle. They crossed to Skye from Glenelg. There they had rain for the first time and were warned that a succession of three dry days was not to be expected for many months. In the island of Raasay they found "nothing but civility, elegance and plenty," and Johnson was delighted with the patriarchal life that he had come to see. Similarly, at Dunvegan he "tasted lotus." On the voyage to the island of Mull in the Inner Hebrides they were driven by a stormy sea into Coll, where, to Boswell's delight, Johnson strutted about one night with a broadsword and target (shield). But they were weather-bound for a week, and Johnson began to long for the mainland—"to go on with existence." It was Boswell, not Johnson, who insisted on visiting Iona, a small island southwest of Mull, but Johnson's comment is one of the most famous paragraphs in all his writings: "That man is little to be envied, whose patriotism would not gain force upon the plain of Marathon, or whose piety would not grow warmer among the ruins on Iona." On the way back to Edinburgh a visit was paid to Boswell's father at Auchinleck, where, to Boswell's distress, Johnson did not succeed in avoiding controversial topics. After a fortnight's stay in Edinburgh, Johnson returned to London and declared that his tour was the most pleasant journey he had ever made.

Boswell's diary

From many points of view the tour was a triumph. Johnson had his 64th birthday at Dunvegan, and his endurance of the physical strain of riding his pony up the rough tracks of Mam Rattachan, of sleeping in Highland huts, and of being tossed in small boats on stormy seas was in itself remarkable. Socially, with rare exceptions, the tour was a complete success, both from Johnson's point of view and from that of his hosts. For Boswell, of course, it was a glorious opportunity for him to write what was, in effect, the first installment of his *Life* of Johnson. Johnson himself was moved, for once, to write a book for its own sake. It was at Anoch in Glenmoriston, on August 31, 1773, that the travellers "entered a narrow valley not very flowery, but sufficiently verdant." While the horses grazed, Johnson sat down on a bank "such as a writer of Romance might have delighted to feign" and conceived the thought not of a romance but of an account of his journey. Of course, he did not attempt to compete with Boswell, whose diary he read, and approved, at intervals. For an intimate personal record of the "minute particulars" of Johnson's behaviour and conversation, the reader naturally turns to Boswell's *Journal of a Tour to the Hebrides,* published after Johnson's death. Johnson, in his own *Journey to the Western Islands of Scotland* (1775), was concerned to describe the customs, religion, education, trade, and agriculture of a society that was new to him. His narrative is far from being impersonal, but it contains no gossip. It was in his many letters to Mrs. Thrale that he wrote more freely. In one letter he offered her a definition of travel highly characteristic of his general approach to life: "The use of travelling is to regulate imagination by reality, and instead of thinking how things may be, to see them as they are."

Not long after his return, Johnson was saddened by the death of Oliver Goldsmith. They had met perhaps as early as 1760, and Johnson had been one of the first to recognize the quality of Goldsmith's writings. He had contributed a famous couplet to *The Traveller,* and *She Stoops to Conquer* had been dedicated to him. He frequently crushed Goldsmith in argument at The Club and elsewhere, but his final judgment was sincere: "Let not his frailties be remembered," he wrote, "he was a very great man."

Tours of Wales and France with the Thrales

Johnson's intimacy with the Thrales at this time is clearly illustrated by his joining them in two long tours—to North Wales in 1774 and to France in the following year. On the way to Wales he was able to introduce them to his Lichfield friends, and some time was spent in the Vale of Clwyd, where Mrs. Thrale's ancestors had lived. On the return journey Johnson was interested to see Matthew Boulton's "enginery" at Birmingham, where Boulton built steam engines in partnership with James Watt, and at Oxford the party was entertained in the hall of University College. In France he was interested more in people than in places. He admired some of the cathedrals, but his most significant comments were on social conditions, particularly on the gulf between rich and poor. There was no provision for the maintenance of the poor, and there was no comfortable middle class.

"The Lives of the English Poets." In 1777 three booksellers waited upon Johnson and asked him to write a series of *Lives* for an "elegant and accurate" edition of the English poets that they and others had in preparation. Johnson agreed. The choice of the poets had already been made, and he strongly objected to the volume being described as "Johnson's Poets." Only five names (and some of them rather odd names) were added at his suggestion. Though he said that he wrote his pieces "dilatorily and hastily, unwilling to work, and working with vigour and haste," it was in this work that he came nearest to actual enjoyment of writing.

The *Lives* are not a series but a miscellany. They follow no plan; they have no uniformity of design. On certain major poets (Milton, Dryden, Pope, and others) Johnson wrote long essays that remain as part of the stock-in-trade of English criticism. He was led on, he said, by "the honest desire of giving useful pleasure." Thus, in writing of Abraham Cowley, he seized the opportunity of a lengthy examination of the characteristics of the Metaphysical Poets, those poets who were "more desirous of being admired than understood." On the other hand, in a laconic piece of 300 words on Richard Duke, he concluded that his poems "were not below mediocrity." When he came to James Thomson (one of his own additions) he praised his *Seasons* highly, but of the poem *Liberty* he wrote: "When it first appeared, I tried to read and soon desisted. I have never tried again . . . ."

Criticism of Milton and Gray

The biographical part of literature was what Johnson loved best, and he made no attempt to separate the poetry from the man who had written it. If he disapproved of the man, he found it difficult to be a detached critic of his work. The *Lives* of Milton and Gray might be cited as two instances. Milton's religion and politics are scornfully exposed—he was not a member of any church and politically he was an acrimonious and surly republican. And some of this prejudice comes through in the discussion of Milton's poetry. To be sure, Johnson had high praise for *Comus* and "L'Allegro" and "Il Penseroso," and thought *Paradise Lost* one of the great "productions of the human mind." But he did not wish it longer, and insisted on pointing out major faults. In his criticism of "Lycidas" (a judgment that was to irritate the Romantics, though even this has been persuasively defended by modern critics), Johnson made a firm break with the older pastoral tradition, which he found stale and appropriate only to descriptions of country life.

Personal dislike is also evident in the *Life* of Gray. To the *Elegy* Johnson paid a famous tribute of genuine admiration, but he approached his other poems with "the neutrality of a stranger and the coldness of a critick." In fact, he was far from neutral in his harsh unfavourable criticisms of Gray's linguistic inversions and antiquarian epithets. It was Gray's attitude to life and literature that irritated Johnson. One who had spent 30 years in Grub Street could have little patience with a fastidious scholar who asked for leisure to be good and wrote only when he was in the humour.

LAST YEARS

Death of Henry Thrale

Publication of the *Lives* was completed in 1781, and a storm of criticism broke out. But Johnson was too old a campaigner to be disturbed by criticism, and, in fact, the *Lives* remains the best loved, or the least neglected, of Johnson's works. What distressed him far more that year than unfavourable reviews was the death of Henry Thrale. No loss since that of his wife had so much oppressed him; he felt like a man beginning a new course of life. To some extent he was saved from brooding over his loss by his attention to his duties as executor. Always interested in trade and commerce, he had enjoyed discussing brewery policy with Thrale in his lifetime, and now he was dealing not with parcels of boilers and vats but with "the potentiality of growing rich beyond the dreams of avarice."

To his disappointment, but to Mrs. Thrale's delight, the brewery was sold to "a knot of rich Quakers" for £135,000. For a time Johnson continued to regard Streatham as his home, but in August 1782 Mrs. Thrale told him of her decision to sell the house. On his last visit in October he humbly thanked God for the comforts he had enjoyed there. A room was allotted to him in the house that Mrs. Thrale took in Argyll Street, but the intimacy of the Streatham days was weakening. For long a victim of asthma and dropsy, Johnson had a paralytic stroke in June 1783, from which he did not properly recover until the spring of the following year. Mrs. Thrale wrote to him with "the attention and tenderness of ancient time," but she dared not tell him that she had for long been trying to decide whether or not to marry Gabriel Piozzi, an Italian musician. When, at the end of June 1784, Johnson heard of her decision to marry, he wrote a letter of angry condemnation; but a week later he recognized that Mrs. Thrale's marriage (which was, in fact, a happy one) was her own affair and thanked her for the kindness that had soothed 20 years of a life radically wretched.

Johnson was now bereft of many old friends. He was a sick man, but to the end he fought untiringly against his worst enemy—solitude—and the melancholy that it induced. In December 1783 he had formed yet another club in Essex Street, which was conveniently near his house in Bolt Court, and in the following July he set out upon his last Midland ramble. At Lichfield everyone was pleased to see him; he stayed for two months at Ashbourne, where he was comfortable but "hungry for conversation," for John Taylor, his friend from school days, went to bed at nine; at Birmingham he talked over old times with Edmund Hector; at Oxford he was welcomed to his old college by the master. On November 16 he returned to London. Though his mind remained alert, his bodily state grew worse, and he died on December 13, 1784. A week later he was buried in Westminster Abbey. One journalist noted that there was only one man of hereditary title among the mourners; but he added, rightly, that he who was followed by Reynolds and Burke did not go unhonoured to the grave.

REPUTATION AND CHARACTER

The history of Johnson's fame is curious. In his own day he was acknowledged as an outstanding writer and thinker, and his fame still persists. Yet today his reputation rests on a dual tradition. There is the "folk image," based largely on Boswell's accounts, and perpetuated by Macaulay, which stresses Johnson the witty talker, who often takes the wrong side of an argument. This is the colourful eccentric, the bear with the heart of gold, whom everybody quotes but nobody reads, remembered chiefly as a character in a great book. Then there is Johnson the man of letters, so admired in his own day and now, after a long eclipse, once again dominating the scene. This second tradition does not rule out Johnson the talker, so admirably described by Boswell, but it concentrates on the Great Cham's own writings.

Johnson the moralist

In scores of recent books the emphasis is not on Johnson's amusing foibles and forceful remarks but rather on his vigorous reasoning intelligence, his keen understanding of human frailty, his detestation of hypocrisy, and his practice of applying ethical standards to nations as well as to individuals. Many of Johnson's so-called obtuse prejudices, when properly understood, derive from his insistence on making strict moral judgments. Johnson was not a sentimental Tory reactionary. He violently opposed war and pleaded for more humane treatment of prisoners of war. He attacked censorship in the 1737 licensing act; he argued for mitigation of laws against prostitutes and debtors and the death penalty for forgery. He was the consistent champion of the poor and oppressed and argued for the rights of blacks and other indigenous peoples. Yet at the same time he saw the necessity of a strong central government in order to withstand the pressures of wealthy mercantile interests.

Moreover, it is now evident that Johnson was not, as was formerly assumed, a strictly neoclassical literary critic; rather, his approach tended to be empirical. He rejected any slavish following of rules and insisted on judging each work separately, making his own decision as to its merits and defects. Johnson is now seen to be the father of 20th-century New Criticism.

Although he scorned those optimists who had a naïve faith in the perfectibility of human institutions, Johnson himself was not a complete pessimist. Along with his skeptical doubts, he had an enormous zest for living. Participation and struggle were always necessary, and while there were some areas where improvement might be possible, everything had to have a strong ethical base. Johnson was not only a moralist but a Christian moralist. The sinfulness of man and his need of redemption by the Passion of Jesus Christ were the basis of his personal faith. His constitutional melancholy deprived him of a feeling of joy in his religion, and the sin of which he was most deeply conscious was idleness. In his *Prayers and Meditations* (1785), published after his death, his repentance and his good resolutions, constantly repeated but seldom kept, show the sincerity of his heart-searching and his humility. But his faith prevailed, and in his last days he refused to take opiates because he had prayed that he might render his soul to God unclouded.

Born in a bookshop, Johnson had the qualities, and the conscience, of a scholar. By the writing of books he strove to earn his daily bread; by the reading of books he sought to enlarge the range of his ideas and of his scholarship. And what, he asked in later years, should books teach but the art of living? Few men have left finer examples of the art of living than Samuel Johnson. (S.R.S./J.L.Cl.)

MAJOR WORKS

VERSE: *London: A Poem* (1738); *The Vanity of Human Wishes* (1749); *Irene: a Tragedy* (1749).

PROSE: *Marmor Norfolciense* (1739), satirical essay; *A Compleat Vindication of the Licensers of the Stage* (1739), ironical defense of the suppression of Henry Brooke's play *Gustavus Vasa; An Account of the Life of Mr Richard Savage* (1744); *Miscellaneous Observations on the Tragedy of Macbeth* (1745); *The Plan of a Dictionary of the English Language* (1747); *The Rambler* (1750–52); *A Dictionary of the English Language,* 2 vol. (1755); *The Prince of Abissinia,* 2 vol. (1759), better known as *Rasselas; The Idler* (1758–60); *The Plays of William Shakespeare,* 8 vol. (1765); *The False Alarm* (1770); *Thoughts on the Late Transactions respecting Falkland's Islands* (1771); *The Patriot* (1774); and *Taxation No Tyranny* (1775), four political tracts; *A Journey to the Western Islands of Scotland* (1775); *Prefaces, Biographical and Critical, to the Works of the English Poets,* 10 vol. (1779–81, rev. as *The Lives of the most eminent English Poets,* 4 vol., 1781); *Prayers and Meditations, composed by Samuel Johnson, LL.D.* (1785), published posthumously by George Strahan.

BIBLIOGRAPHY. The standard bibliography is WILLIAM P. COURTNEY and DAVID NICHOL SMITH, *A Bibliography of Samuel*

Johnson (1915, reprinted 1967), supplemented by ROBERT W. CHAPMAN and ALLEN T. HAZEN, "Johnsonian Bibliography: A Supplement to Courtney," in OXFORD BIBLIOGRAPHICAL SOCIETY, *Proceedings and Papers,* vol. 5, pt. 3, p. 116 (1938). ROBERT B. ADAM, *The R.B. Adam Library Relating to Dr. Samuel Johnson and His Era,* 4 vol. (1929–30), contains facsimilies of many manuscripts and letters. The most extensive Johnson collection is now in the Hyde Library, near Somerville, New Jersey, described in GABRIEL AUSTIN (ed.), *Four Oaks Library,* 2 vol. (1967). JAMES L. CLIFFORD and DONALD J. GREENE, in *Samuel Johnson: A Survey and Bibliography of Critical Studies* (1970), with about 4,000 entries, list the most important works on Johnson from his day through 1968.

Collected editions: The Oxford edition, *The Works of Samuel Johnson,* 9 vol. (1825), with two supplementary volumes of *Debates,* has long been the most cited, but its text is deficient by modern standards. The collection was reprinted under the title *Dr. Johnson's Works,* 11 vol. (1970). The new Yale edition under the general editorship of ALLEN T. HAZEN and JOHN H. MIDDENDORF undertakes to supply a more satisfactory text. To this edition belong *Diaries, Prayers, and Annals,* ed. by EDWARD L. MCADAM, JR., with DONALD and MARY HYDE (1958); *The Idler and The Adventurer,* ed. by WALTER J. BATE, JOHN M. BULLITT, and LAWRENCE F. POWELL (1963); *Poems,* ed. by EDWARD L. MCADAM, JR., with GEORGE MILNE (1964); *Johnson on Shakespeare,* ed. by ARTHUR SHERBO, 2 vol. (1968); *The Rambler,* ed. by WALTER J. BATE and ALBRECHT B. STRAUSS, 3 vol. (1969); *A Journey to the Western Islands of Scotland,* ed. by MARY LASCELLES (1971); *Political Writings,* ed. by DONALD J. GREENE (1977); and *Sermons,* ed. by JEAN H. HAGSTRUM and JAMES GRAY (1978). HELEN H. NAUGLE and PETER B. SHERRY (eds.), *A Concordance to the Poems of Samuel Johnson* (1973), is based on *The Poems of Samuel Johnson,* ed. by DAVID NICHOL SMITH and EDWARD L. MCADAM (1941), and on the Yale edition of the *Poems.* Until superseded, the standard edition of *The Lives of the English Poets* is that by GEORGE B. HILL, 3 vol. (1905, reprinted 1967); and of the correspondence, *The Letters of Samuel Johnson with Mrs. Thrale's Genuine Letters to Him,* ed. by R.W. CHAPMAN, 3 vol. (1952). The most extensive anthology is *Johnson, Prose and Poetry,* selected by MONA WILSON (1950, reprinted 1967).

Biography and criticism: The number of biographical and critical studies of Johnson is enormous. Apart from Boswell's *Journal of a Tour to the Hebrides* (1785) and *Life of Johnson* (1791), both of which are available in many editions, there are a number of contemporary authorities, including the official *Life of Samuel Johnson, LL.D.* by SIR JOHN HAWKINS (1787, reprinted 1974); HESTER LYNCH PIOZZI, *Anecdotes of the Late Samuel Johnson, LL.D., During the Last Twenty Years of His Life* (1786; ed. by S.C. ROBERTS, 1925); ARTHUR MURPHY, *An Essay on the Life and Genius of Samuel Johnson, LL.D.* (1792, reprinted 1970); ROBERT ANDERSON, *The Life of Samuel Johnson, LL.D., with Critical Observations on His Works* (3rd expanded ed., 1815, reprinted 1974). Portions of these and many other anecdotes are included in GEORGE B. HILL (ed.), *Johnsonian Miscellanies,* 2 vol. (1897, reprinted 1966). Two of the most famous essays on Johnson are those of THOMAS MACAULAY: one a review of J.W. Croker's edition of Boswell (*Edinburgh Review,* 1831), in which the depreciation of Johnson as a writer set the fashion for many years; the other, a more balanced account, in the *Encyclopædia Britannica,* 8th ed. (1856). Among shorter Victorian studies are those by LESLIE STEPHEN, *Samuel Johnson* (1878, reprinted 1968); and FRANCIS R. GRANT, *Life of Samuel Johnson* (1887, reprinted 1972). See also JAMES T. BOULTON (ed.), *Johnson: The Critical Heritage* (1971); and MARY HYDE, *The Impossible Friendship: Boswell and Mrs. Thrale* (1972).

In the 20th century much new biographical evidence has been assembled by ALEYN L. READE in *Johnsonian Gleanings,* 11 vol. (1909–52, reprinted 1968): ALLEN T. HAZEN, *Samuel Johnson's Prefaces and Dedications* (1937, reprinted 1973); EDWARD L. MCADAM, JR., *Dr. Johnson and the English Law* (1951); BENJAMIN B. HOOVER, *Samuel Johnson's Parliamentary Reporting* (1953); and in the various volumes of the *Yale Edition of the Private Papers of James Boswell* (ed. by FREDERICK A. POTTLE *et al.*), particularly *The Correspondence and Other Papers of James Boswell Relating to the Making of the Life of Johnson,* ed. by MARSHALL WAINGROW (1969). Much of this new evidence is used in JOSEPH W. KRUTCH, *Samuel Johnson* (1944, reprinted 1963); JAMES L. CLIFFORD, *Young Sam Johnson* (1955, reprinted 1981), and *Dictionary Johnson: Samuel Johnson's Middle Years* (1979); PAUL FUSSELL, *Samuel Johnson and the Life of Writing* (1971); CHRISTOPHER HIBBERT, *The Personal History of Samuel Johnson* (1971); and GEORGE IRWIN, *Samuel Johnson: A Personality in Conflict* (1971).

Special studies include WALTER RALEIGH, *Six Essays on Johnson* (1910, reprinted 1965), an early revaluation correcting Macaulay's view; BERTRAND H. BRONSON, *Johnson Agonistes, and Other Essays* (1946, reissued 1965); WILLIAM K. WIMSATT, JR., *The Prose Style of Samuel Johnson* (1941, reprinted 1972); JEAN H. HAGSTRUM, *Samuel Johnson's Literary Criticism* (1952, reprinted 1967); JEAN H. SLEDD and GWIN J. KOLB, *Dr. Johnson's Dictionary: Essays in the Biography of a Book* (1955, reprinted 1974); ARTHUR SHERBO, *Samuel Johnson, Editor of Shakespeare* (1956, reprinted 1978); EDWARD A. BLOOM, *Samuel Johnson in Grub Street* (1957); DONALD J. GREENE, *The Politics of Samuel Johnson* (1960, reprinted 1973); MAURICE J. QUINLAN, *Samuel Johnson: A Layman's Religion* (1964); CHESTER F. CHAPIN, *The Religious Thought of Samuel Johnson* (1968); and RICHARD B. SCHWARTZ, *Samuel Johnson and the New Science* (1971), and *Samuel Johnson and the Problem of Evil* (1975). A study of one aspect of Johnson's critical thought is presented in LEOPOLD DAMROSCH, *Samuel Johnson and the Tragic Sense* (1972); and a discussion of the fusion of religion and morality is found in JAMES GRAY, *Johnson's Sermons: A Study* (1972). There have also been various collections of shorter articles: FREDERICK W. HILLES (ed.), *New Light on Dr. Johnson: Essays on the Occasion of His 250th Birthday* (1959, reprinted 1967); *Johnson, Boswell and Their Circle: Essays Presented to Lawrence Fitzroy Powell* (1965); DONALD J. GREENE (ed.), *Samuel Johnson: A Collection of Critical Essays* (1965); and JAMES L. CLIFFORD (ed.), *Twentieth Century Interpretations of Boswell's Life of Johnson* (1970). Other studies include CAREY MCINTOSH, *The Choice of Life: Samuel Johnson and the World of Fiction* (1973); ROBERT D. STOCK, *Samuel Johnson and Neoclassical Dramatic Theory: The Intellectual Context of the "Preface to Shakespeare"* (1973), and (ed.), *Samuel Johnson's Literary Criticism* (1974); PETER QUENNELL, *Samuel Johnson: His Friends and Enemies* (1973); and DOROTHY MARSHALL, *Dr. Johnson's London* (1968).

General evaluations that stress Johnson the moralist, thinker, and critic are: WALTER J. BATE, *The Achievement of Samuel Johnson* (1955, reprinted 1978), and *Samuel Johnson* (1977, reprinted 1979), a psychological study; ROBERT B. VOITLE, JR., *Samuel Johnson the Moralist* (1961); PAUL K. ALKON, *Samuel Johnson and Moral Discipline* (1967); ARIEH SACHS, *Passionate Intelligence: Imagination and Reason in the Works of Samuel Johnson* (1967). DONALD J. GREENE, *Samuel Johnson* (1970), is an excellent introduction and analysis. Later studies include THOMAS M. CURLEY, *Samuel Johnson and the Age of Travel* (1976); WILLIAM EDINGER, *Samuel Johnson and Poetic Style* (1977); JOHN P. HARDY, *Samuel Johnson: A Critical Study* (1979); JOHN WAIN, *Samuel Johnson,* 2nd ed. (1980), a comprehensive biographical and literary account; and ISRAEL SHENKER, *In the Footsteps of Johnson and Boswell* (1982), an imaginative, humorous account of places visited by Johnson and Boswell.

Jordan

The Hashemite Kingdom of Jordan (al-Mamlakah al-Urdunnīyah al-Hāshimīyah), an Arab state of Southwest Asia, is a young nation that occupies an ancient land associated with the civilizations of antiquity. It is bounded to the north by Syria, to the east by Iraq, to the southeast and south by Saudi Arabia, and to the west by Israel and the West Bank. Jordan has 12 miles (19 kilometres) of coastline on the Gulf of Aqaba in the southwest, where al-ʿAqabah, its only port, is located. The total area of undisputed territory is 34,443 square miles (89,206 square kilometres). Jordan's capital and largest city is Amman.

As an international entity, Jordan came into being after World War I, gaining its independence from the United Kingdom as a hereditary constitutional monarchy in 1946. King Hussein ibn Talal ascended the throne in 1953; he attempted to maintain Jordan's traditional policy of friendship with the West despite strong local and international pressures. Although Jordan has meagre natural resources, its most pressing problems are political. The intellectual and ideological divisions and frustrations that pervade the Arab world are reflected within Jordan—all the more so because the country has provided a haven for thousands of Arab refugees who fled from their homes as a result of the Arab-Israeli wars.

An economically developing country, Jordan throughout its existence has had to depend upon outside aid. This came first from the United Kingdom, later from the United States, and then—since 1967—from other Western and Arab countries, including Saudi Arabia, Kuwait, and the United Arab Emirates. Despite population pressures, a paucity of natural resources, and an influx of refugees, Jordan was able to achieve strong economic growth prior to the 1967 war. To sustain this progress, Jordan initiated state planning in 1964, with a seven-year plan envisaging a self-sufficient economy. The 1967 Six-Day War with Israel, however, and the resulting influx of still more refugees, temporarily halted some of the measures included in the plan, overstrained the economy, which was already in difficulty, and increased Jordan's dependence on outside aid. Since then Jordan has successfully implemented a series of government plans. The object of these plans has been the revitalization of Jordan's economy, the reduction of its need for outside aid, and the promotion of general economic and social development.

This article is divided into the following sections:

Physical and human geography

THE LAND

Relief and drainage. There are three major physiographic regions in Jordan: the Jordan desert, the East Bank uplands, and the Jordan rift valley (a branch of the great African rift-valley system).

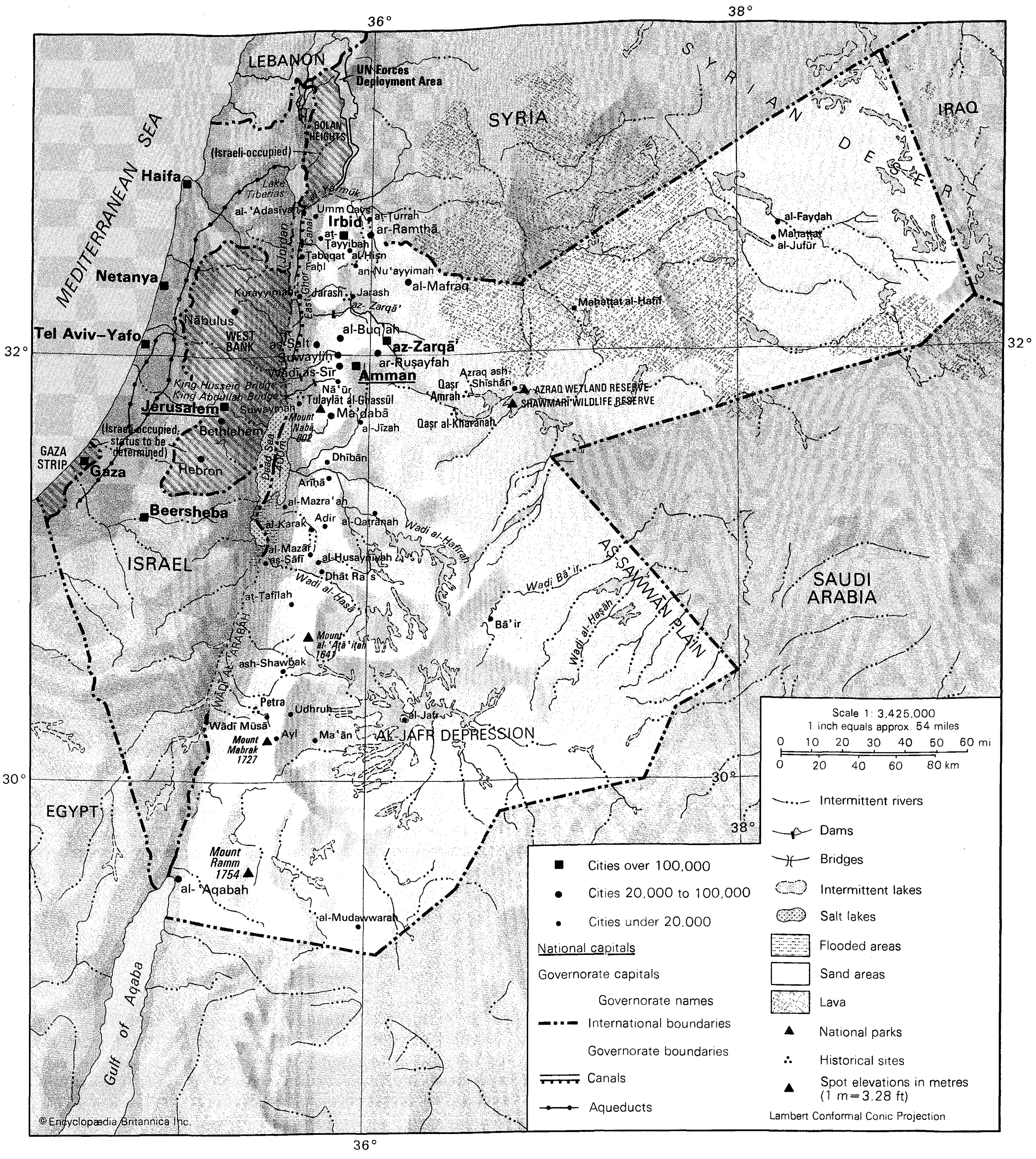
The desert is located in the eastern and southern parts of the country, occupying more than four-fifths of its territory. The desert's northern part is composed of volcanic lava and basalt, and its southern part of outcrops of sandstone and granite. It is much eroded, primarily by wind. The East Bank uplands, an escarpment overlooking the rift valley, has an average altitude of between 2,000 and 3,000 feet (600 and 900 metres); the elevation increases to about 5,750 feet in the south. There are outcrops of sandstone, chalk, limestone, and flint extending to the extreme south, where igneous rocks solidified from the molten state predominate. In the northern uplands several valleys and perennial streams run west; around al-Karak they run west, east, and north; south of al-Karak nonperennial valley streams run east toward al-Jafr Depression.

Jordan Valley

The Jordan Valley, some 1,312 feet below sea level at the Dead Sea, contains the lowest point on the Earth's surface. Meandering south, the Jordan River drains the waters of the Sea of Galilee (Lake Tiberias), the Yarmūk, and the valley streams of both plateaus into the Dead Sea. The soil of its lower reaches is very saline, and the shores of the Dead Sea consist of salt marshes that do not support vegetation. The Dead Sea occupies the central area of the valley. To its south, Wadi al-ʿArabah, a completely desolate region, is thought to contain mineral resources.

Climate. The climate varies from the Mediterranean type in the west to the desert type in the east and south, but the land is generally arid. The proximity of the Mediterranean Sea is the major climatic influence, although this influence is modified by continental air masses and by altitude. Average monthly temperatures at the capital in the north range between 46° and 78° F (8° and 26° C), while at al-ʿAqabah in the far south they range between 60° and 91° F (16° and 33° C). The prevailing winds throughout the country are westerly to southwesterly, but spells of hot, dry, dusty winds blowing from the southeast off the Arabian Peninsula frequently occur. Known locally as the khamsin, these winds bring the country its most uncomfortable weather. They blow most often in the early and late summer and can last for several days at a time before terminating abruptly as the wind direction changes and much cooler air follows. Rainfall occurs in the short, cool winters, decreasing from 16 inches (400 millimetres) annually in the northwest near the Jordan River to less than four inches in the south. The average rainfall in the East Bank uplands totals about 14 inches annually. The valley itself has a yearly average of eight inches, and the desert regions receive less than two inches. Occasional snow and frost occur in the uplands but are rare in the rift valley.

Plant and animal life. The plant and animal life of Jordan falls into three distinct types: that associated with



MAP INDEX

Political subdivisions

ʻAmmān	31 40 N 36 30 E
Balqā', al-	31 50 N 35 40 E
Irbid	32 25 N 35 50 E
Karak, al-	31 10 N 35 45 E
Maʻān	30 10 N 36 20 E
Mafraq, al-	32 10 N 37 45 E
Ṭafīlah, aṭ-	30 45 N 35 40 E
Zarqā', az-	32 05 N 36 10 E

Cities and towns

ʻAdasīyah, al-	32 40 N 35 37 E
Adir	31 12 N 35 46 E
Amman (ʻAmmān)	31 57 N 35 56 E
ʻAqabah, al-	29 31 N 35 00 E
Arīḥā	31 25 N 35 47 E
Ayl	30 13 N 35 32 E
Azraq ash-Shīshān	31 50 N 36 49 E
Bā'ir	30 46 N 36 41 E
Buqʻah, al-	32 05 N 35 53 E
Dhāt Ra's	31 00 N 35 46 E
Dhībān	31 30 N 35 47 E
Fayḍah, al-	32 35 N 38 13 E
Ḥiṣn, al-	32 29 N 35 53 E
Ḥusaynīyah, al-	31 02 N 35 43 E
Irbid	32 33 N 35 51 E
Jafr, al-	30 18 N 36 13 E
Jarash	32 17 N 35 54 E
Jīzah, al-	31 42 N 35 57 E
Karak, al-	31 11 N 35 42 E
Karāmah, al-	31 57 N 35 35 E
Kurayyimah	32 16 N 35 36 E
Maʻān	30 12 N 35 44 E
Ma'dabā	31 43 N 35 48 E
Mafraq, al-	32 21 N 36 12 E
Maḥaṭṭat al-Ḥafīf	32 12 N 37 08 E
Maḥaṭṭat al-Jufūr	32 30 N 38 12 E
Mazār, al-	31 04 N 35 42 E
Mazraʻah, al-	31 16 N 35 31 E
Mudawwarah, al-	29 19 N 35 59 E
Nāʻūr	31 53 N 35 50 E
Nuʻayyimah, an-	32 25 N 35 55 E
Qaṭrānah, al-	31 15 N 36 03 E
Ramthā, ar-	32 34 N 36 00 E
Ruṣayfah, ar-	32 01 N 36 03 E
Ṣāfī, aṣ-	31 02 N 35 28 E
Ṣalt, aṣ-	32 03 N 35 44 E
Shawbak, ash-	30 32 N 35 34 E
Ṣuwayliḥ	32 02 N 35 50 E
Suwaymah	31 47 N 35 36 E
Ṭabaqat Faḥl	32 27 N 35 37 E
Ṭafīlah, aṭ-	30 50 N 35 36 E
Ṭayyibah, aṭ-	32 33 N 35 43 E
Ṭurrah, aṭ-	32 38 N 35 59 E
Udhruḥ	30 20 N 35 36 E
Umm Qays	32 39 N 35 41 E
Wādī as-Sīr	31 57 N 35 49 E
Wādī Mūsā	30 19 N 35 29 E
Zarqā', az-	32 05 N 36 06 E

Physical features and points of interest

Allenby Bridge, see King Hussein Bridge	
Aqaba, Gulf of (Khalīj al-ʻAqaba)	29 00 N 34 40 E
ʻArabah, Wadi al-, *valley*	30 10 N 35 10 E
ʻAṭāʻiṭah, Mount al-	30 40 N 35 39 E
Azraq Wetland Reserve	31 51 N 36 51 E
Bādiyat ash-Shām, see Syrian Desert	
Bā'ir, Wadi	31 12 N 37 31 E
Batrā', al-, see Petra	
Dead Sea (al-Baḥr al-Mayyit)	31 30 N 35 30 E
East Ghor Canal (Qanāt al-Ghawr ash-Sharqīyah)	32 41 N 35 38 E
Ḥafīrah, Wadi al-	31 15 N 36 02 E
Ḥasā, Wadi al-	30 49 N 35 48 E
Ḥaṣāh, Wadi al-	30 56 N 37 35 E
Jafr Depression, al-	30 17 N 36 20 E
Jarash, *historical site*	32 17 N 35 54 E
Jordan River (Nahr al-Urdunn)	31 46 N 35 33 E
King Abdullah Bridge	31 48 N 35 33 E
King Hussein Bridge (Allenby Bridge)	31 52 N 35 32 E
Mabrak, Mount	30 13 N 35 29 E

Mayyit, al-Baḥr al-, see Dead Sea	
Mediterranean Sea	32 30 N 34 40 E
Nabā, Mount	31 46 N 35 45 E
Petra (al-Batrā'), *historical site*	30 20 N 35 26 E
Qaṣr al-Kharānah, *historical site*	31 44 N 36 28 E
Qaṣr 'Amrah, *historical site*	31 48 N 36 35 E
Ramm Mount	29 35 N 35 24 E
Ṣawwān Plain, aṣ-	30 45 N 37 15 E
Shawmarī Wildlife Reserve	31 46 N 36 48 E
Syrian Desert (Bādiyat ash-Shām)	32 00 N 40 00 E
Tulaylāt al-Ghassūl, *historical site*	31 48 N 35 36 E
Urdunn, Nahr al-, see Jordan River	
Yarmūk, al-, *river*	32 38 N 35 34 E
Zarqā, az-, *river*	32 07 N 35 33 E

the Mediterranean, with the steppe (treeless plains), and with the desert. In the uplands the Mediterranean type predominates, while in the drier steppe region sagebrush predominates. Grassland is the most prevalent vegetation on the steppe, however, but some isolated trees and shrubs, such as lotus fruit and the Mount Atlas pistachio, also occur. In the desert scant vegetation grows in depressions and on the sides and floors of the valleys.

There is a great variety of animal life, including wild boars, as well as the ibex, a species of wild goat found in the gorges and in the 'Ayn al-Azraq oasis. Hares, jackals, foxes, wildcats, hyenas, wolves, gazelles, mole rats, mongooses, and a few panthers also occur. Among the domesticated animals, horses, mules, donkeys, camels, cattle, sheep, and goats are most common. Centipedes, scorpions, and various types of lizards are also found. Birds include the golden eagle and the vulture, while wild fowl include the pigeon and the partridge.

Nomads

Settlement patterns. The landscape falls into two regions—the desert zone and the cultivated zone—each of which is associated with its own mode of living. The nomads (Bedouin, or Badu) generally inhabit the desert and some areas of the steppe and the uplands. The number of nomads has decreased dramatically because of successful government efforts at resettlement. The eastern Bedouin are principally camel breeders and herders, while the western Bedouin are sheep and goat herders. There are some seminomads, in whose existence the modes of life of the desert and the cultivated zones merge. These people adopt a nomadic existence, living in tents only in the winter months after they have planted their lands, upon which some have also built modest homes; they return to their homes again in the spring or at harvest time. The two largest nomadic tribes of Jordan are Banū Sakhr and al-Ḥuwayṭāt. The grazing grounds of both are entirely within Jordan, as is the case with the smaller tribe of as-Sirhān. Other, lesser tribes include Banū Hasan, al-Banū Khalid, al-Ajarmeh, al-Adwan, Banū Attiyeh, al-Hajayah, and as-Sleet, as well as the smaller tribes of al-Hawazim, as-Sulaylat, and ash-Sherarat, which traditionally were obliged to pay protection money to larger tribes. The Rwalah tribe, which is not indigenous, passes through Jordan in its yearly wandering from Syria to Saudi Arabia.

Including nomads, rural residents represent about a third of the population. The average village is a cluster of houses and other buildings, including an elementary school and a mosque, with pasturage on the outskirts. A medical dispensary and a post office may be found in the larger villages, together with a general store and a small café, whose owners are usually part-time farmers. Kinship relationships are patriarchal, while extended-family ties govern social relationships and tribal organization. Increases in the literacy rate and the influence of the mass media, in addition to extensive migration from rural to urban areas, have had a marked influence.

Of the total population, more than half live in the dozen or so major cities and towns. Amman has a population of more than three-quarters of a million, but the smaller towns have only a few thousand inhabitants. Most towns have hospitals, banks, government and private schools, mosques, churches, libraries, and entertainment facilities, and some have institutions of higher learning and newspapers. Amman and az-Zarqā', and to some extent Irbid, have urban characteristics, while smaller towns are more reluctant to accept modernizing influences.

THE PEOPLE

Languages and dialects

Ethnic and religious groups. The majority of the people are Arabic-speaking. In addition to the difference between the written, or classic, Arabic and the colloquial form, there are various dialects with local inflections and accents. The Qaysī-Yemeni dichotomy—a pre-Islāmic split that was introduced to the area with the Arab conquests and that cut across religious and ecological lines—was at one time an important broad social division. The Arabs, whether Muslim or Christian, used to trace their ancestry from the north Arabian Qaysī (Ma'dī, Nizārī, Adnanī, or Ismā'īlī) tribes or from the south Arabian Yemeni (Banū Kalb or Qahtani) tribes. Only a few tribes and towns have continued to be aware of the split.

The vast majority of the people are Sunnite Muslim, and a small percentage are Christian. Among Christians adherents of the Rūm, or Greek Orthodox church, form a majority. Other Christian groups include the Rūm, or Greek, Catholics, also called the Melchites, or Eastern Catholics, who recognize the supremacy of the pope; the Roman Catholic community, headed by a patriarch appointed by the pope; and the small Syrian Orthodox, or Jacobite, church, whose members use Syriac in their liturgy. Most non-Arab Christians are Armenians; the majority belong to the Gregorian, or Armenian, Orthodox church, the rest to the Armenian Catholic church. There are several Protestant denominations representing relatively recently formed communities whose converts came almost entirely from other Christian sects.

The Druze, an offshoot of the Ismā'īlī Shī'ite sect, number a few hundred and reside in and around Amman. The Bahā'ī—who in the 19th century also split off from Shī'ite Islām and who number around 1,000—live in al-'Adasīyah in the Jordan Valley. The Armenians, Druze, and Bahā'ī are at once religious and ethnic communities. The Shishan (Chechen) are a Circassian Shī'ite Muslim group, numbering around 1,000, who are descended from 19th-century immigrants. With the Cherkess, who are Sunnite, they make up the most important non-Arab minority. Another small non-Arab group consists of some Turkmen.

Demography. The population structure is predominantly young; persons under the age of 15 constitute the largest component of the population. The birth rate is high relative to the death rate, producing a natural rate of increase of about 3 percent annually. Internal migration from rural to urban centres has added an additional burden to the economy, while the number of Jordanians living abroad has increased significantly.

Refugees

The influx of Palestinian refugees has not only altered Jordan's demographic map but also affected its political, social, and economic life. After the 1948–49 Arab-Israeli War and the annexation of the West Bank, Jordanian citizenship was granted to some 400,000 Palestinians who were residents of and remained on the West Bank and to about half a million refugees from the new Israeli state. Many of these refugees settled on the East Bank. From 1949 to 1967 Palestinians continued to move to the East Bank in large numbers. After the 1967 war an estimated 310,000 Palestinians, mostly from the West Bank, sought refuge in Jordan; thereafter immigration from the West Bank continued at a reduced rate.

More than 60 percent of Jordan's population are ethnic Palestinians. Most are employed and hold full Jordanian citizenship; fewer than half are registered with the United Nations Relief and Works Agency for Palestine Refugees in the Near East (UNRWA). Only a small percentage live in refugee camps or receive aid from the UNRWA.

THE ECONOMY

Despite its basic problems, the Jordanian economy before 1967 showed resilience and growth. Economic growth, which was halted by war in the second half of 1967, continued thereafter at a slower pace but was revitalized by the implementation of a series of state economic plans. The West Bank's contribution to domestic income prior to the war was around one-third of the total, but its occupation by Israel in 1967 required that the government apply its social and economic plans to the East Bank only.

The major sources of revenue are community, social, and

Sources of revenue

personal services; mining and industry, trade, communications and transportation, agriculture, and construction. Income from tourism, which has grown dramatically, is mostly in foreign reserves, and tourism has become a major factor in Jordan's efforts to reduce its balance of payments deficit.

There has been a great outflow of skilled labour from Jordan to neighbouring countries, although the problem has eased somewhat. This change is a result both of better employment opportunities within Jordan itself and of a curb on foreign labour by the neighbouring Persian Gulf states.

Resources. There are only some 90,000 acres (36,000 hectares) of forest in Jordan, most of which are on the rocky highlands. Although unprotected, these forests have survived the depredations of villagers and nomads alike, as well as constant overgrazing. The Jordanian government embarked on a reforestation program in 1948. In the higher regions of the uplands, the predominant types of trees are the Aleppo oak, the Kermes oak, the Palestinian pistachio, the Aleppo pine, and the Oriental strawberry tree. Wild olives are also found there, and the Phoenician juniper occurs in the regions with lower rainfall.

Pastureland is so degraded that it can barely support Jordan's livestock; it has, moreover, been reduced by the extension of land devoted to olive and fruit trees. Artesian wells have been dug to increase the pasturage area. Sheep and goats are by far the most important livestock, but there are some cattle, camels, horses, donkeys, and mules. Livestock decreases when droughts occur. There is fishing in the Gulf of Aqaba.

Mineral resources include large deposits of phosphates, potash, limestone, and marble, as well as dolomite, kaolin, and salt. In addition, newly discovered minerals include barite (the principal ore of the metallic element barium), quartzite, gypsum (used as a fertilizer), and feldspar, and there are unexploited deposits of copper and uranium. Cement production is a major mineral-based industry.

Power in Jordan is generated by fossil fuel, mostly oil. There are several generating plants; the two major power stations, at Amman and az-Zarqā', are linked by a transmission system. By the late 20th century the government had nearly completed a program to link the major cities by a countrywide grid, which later is to include rural areas.

Trade. Exports, though growing, do not cover the value of imports; the deficit is financed by foreign grants, loans, and other forms of capital transfers. Although Jordan's trade deficit has been large, it is offset somewhat by earnings from tourism, remittances sent by Jordanians working abroad, earnings from foreign investments made by the Jordan Central Bank, and subsidies from Arab and other governments.

Administration of the economy. The economy is primarily based on private enterprise. To reduce its dependence on outside assistance, raise the standard of living, and reduce the imbalance between imports and exports, the government, in 1964, began the process of planning the economy. All of the government plans have given a high priority to agriculture, mining and industry, tourism, and the service industries, in addition to the strengthening and widening of the social, health, welfare, and education infrastructure. Unemployment among the male population has been reduced while income per capita has increased dramatically.

Aside from a licensing system, a moderate taxation on luxury items, and the establishment of standards and health measures, both internal and international trade are virtually free of restraint. Governmental support takes the form of ensuring the basic framework of internal security, guaranteeing a stable currency, and maintaining the transport and communication routes and other facilities required for economic progress. The government also has participated with private enterprise in establishing the largest mining, industrial, and tourist firms in the country. The government also owns a significant share of the largest companies.

Taxation

Fiscal policy has aimed at increasing revenue by raising various tax rates and by reforming the tax system. Measures applied since 1964 include increases in customs and excise duties and an increase in income taxes. Although the government has placed great effort on reforming the income tax, both to increase revenue and to redistribute income, revenue from indirect taxes continues to exceed that from direct taxes. Tax measures have been adopted to increase the rate of savings necessary for financing investments. Exemptions on foreign investments and transfers of foreign profits and capital have been continued.

The existence of labour unions and employer organizations is recognized by law. The weakness of the trade-union movement is partly compensated for by the government, which has special procedures for settling labour disputes; if governmental efforts fail, the union or employers may resort to the judicial process.

The small size of the Jordanian market, the fluctuations in agricultural production because of irregular rainfall, lack of capital, political instability, and the presence of refugees all combine to make the continuation of outside help a necessity. The economy is thus highly sensitive both to domestic and to international policy.

The road system

Transportation. Jordan has a main, secondary, and rural road network, most of which is hard-surfaced. This roadway system, maintained by the Ministry of Public Works, links the major cities and towns and also links the kingdom with neighbouring countries. Within cities, towns, and villages, however, the local authorities are responsible for road upkeep. One of the main traffic arteries is the Amman–Jarash–ar-Ramthā highway, which links Jordan with Syria. The route from Amman via Maʿān to the port of al-ʿAqabah is the principal route to the sea. From Maʿān the Desert Highway passes through al-Mudawwarah, linking Jordan with Saudi Arabia. The Amman–Jerusalem highway, passing through Nāʿūr, is a major tourist artery. The Hejaz Jordan Railway is government-operated and extends from Darʿā in the north via Amman to Maʿān in the south. A new line operated by the Aqaba Railway Corporation runs to the port of al-ʿAqabah, and another line between Maʿān and Medina in Saudi Arabia is being constructed. Rail connections also run from Darʿā north to Damascus in Syria. The Royal Jordanian Airline links Jordan to Arab, African, Asian, American, and European countries. Queen Alia international airport near al-Jīzah, south of Amman, was opened in 1983. Amman and al-ʿAqabah have smaller airports.

Before 1948 Jordan's Mediterranean Sea trade was through Haifa. The Arab-Israeli conflicts severed that link, and Jordan's outlet was then, for a time, through Beirut, in Lebanon. The expansion of the Jordanian economy, especially the export of phosphate, led to the development of the port of al-ʿAqabah.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The 1952 constitution is the most recent of a series of legislative instruments that, both before and after independence, moved toward increased executive responsibility. The constitution declares Jordan to be a constitutional hereditary monarchy with a parliamentary form of government. Islām is the official religion of the state, and Jordan is declared to be part of the Arab *ummah* ("nation"). The king wields wide powers over the executive, legislative, and judicial branches. Jordan's central government is headed by a prime minister appointed by the king; the prime minister then chooses his Cabinet. According to the constitution, the appointments of both prime minister and Cabinet are subject to parliamentary approval. The Cabinet coordinates the work of the different departments and establishes general policy.

Under the constitution the membership of the upper house of the bicameral legislature, composed of *al-aʿyān* ("notables"), is appointed by the king for four years. Elections for *nuwwāb* ("deputies") of the lower house are to be held at least every four years, although elections have been frequently suspended. The ninth parliament, elected in 1965, was prorogued several times before being replaced in 1978 by the National Consultative Council, an appointed body with reduced power that debates government programs and activities. The parliament was reconvened, however, in a special session called in January 1984.

Persons 18 years of age and over may vote provided they meet the legal requirements and are not members of the royal family. Voting participation has varied and has run as high as 70 percent. Political parties were banned in 1957, however. Before that date several parties—Communist, Arạb Baʿath Socialist, National Socialist, Muslim Brotherhood, Liberation Movement, Arab Constitutional, Nahda (Renaissance), and Ummah (Community)—ran candidates, but none had a mass following. The Muslim Brotherhood was the only party exempted from the 1957 ban, but the group has been kept under close surveillance.

Jordan is divided into administrative *muḥāfaẓāt* (governorates), which in turn are divided into districts and subdistricts, each of which is headed by an official appointed by the minister of the interior. Cities and towns have mayors and elected councils.

Three categories of courts

Justice. The judiciary is constitutionally independent, though judges are appointed and dismissed by royal *irādah* ("decree") following a decision of the Justices Council. There are three categories of courts. The first category consists of regular courts, including magistrates' courts, courts of first instance, and courts of appeals and cassation in Amman, which hear appeals passed on from lower appeals courts. The constitution also provides for the Diwān Khāṣṣ (Special Council), which interprets the laws and passes on their constitutionality. The second category consists of Sharīʿah Muslim courts and other religious courts for non-Muslims; these exercise jurisdiction over matters of personal status. The third category consists of special courts, such as land, government, property, municipal, tax, and customs courts.

The armed forces. The Jordanian armed forces, which include an air force equipped with modern jet aircraft, developed from the Arab Legion, which was originally commanded by British officers. There also is a small navy. The king is commander in chief of the armed forces.

Education. There are three types of schools in Jordan—government schools, private schools, and the UNRWA schools for refugee children. Schooling consists of six years of elementary, three years of preparatory, and three years of secondary education. The Ministry of Education supervises all schools and establishes the curricula, teachers' qualifications, and state examinations; it also distributes free books to students in government schools and enforces compulsory education to the age of 14. The majority of the students attend government schools. In addition to Khadduri Agricultural Training Institute, there are agricultural secondary schools, as well as a number of vocational, labour, and social affairs institutes, a Sharīʿah (Qur'ānic) seminary, and nursing, military, and teachers' colleges. The State University of Jordan was established in Amman in 1962; Yarmūk University was established in 1976. A third university, Mu'tah, opened in 1981.

Health and welfare. Infectious diseases, except for dysentery and eye infections, have been brought under control. The number of physicians has grown rapidly. Comprehensive health facilities are operated by the government. A national health insurance program covers medical, dental, and eye care at a modest cost; service is provided free to the indigent.

Welfare services were private until the Ministry of Social Affairs was established in 1951. Besides supervising and coordinating social and charitable organizations, the ministry administers welfare programs.

Wages are higher in industry than in agriculture and in the cities than elsewhere. There are minimum wages for both skilled and unskilled labourers. The development of local services and resources, the availability of hard currencies from foreign aid, and a competitive import policy have given Jordan a degree of price stability not often encountered among developing nations.

Housing. The housing situation has remained critical despite continuing housing construction. Housing surveys conducted in Amman and the East Jordan Valley showed that the major proportion of households consist of one-room dwellings. The Housing Corporation and the Jordan Valley Authority build units for low-income families. Urban renewal projects in Amman and az-Zarqā' have provided new units and renovated others. The Housing Bank issues home building loans.

CULTURAL LIFE

Culturally, Jordan is an integral part of the Arab world and thus cannot be said to have a separate and distinct culture of its own. As in the rest of the Arab world, the highest form of artistic expression remains oral. Jordan's most famous poet was Muṣṭafā Wahbah aṭ-Ṭāl, whose style and content rank him among the major Arab poets of the 20th century. After World War II a number of important poets and prose writers emerged, though few have achieved an international reputation.

Both private and governmental efforts have been made to foster the arts, and an art gallery has been opened. Modernity has lessened the influence of the traditional Islāmic injunction against the portrayal of animate objects. Thus, in addition to the traditional decorative design, architecture, and the various handicrafts, it is possible to find sophisticated forms of painting and sculpture.

Folk and popular arts

Folk art survives in tapestry work and in the making of leather, pottery, and ceramics, as well as in the manufacturing of wool and goat-hair rugs with varicoloured stripes. Popular culture is manifested in such oral arts as songs, ballads, and storytelling. The villagers have special songs for births, circumcisions, weddings, funerals, planting, plowing, and harvesting. Several types of *debkah* (dances characterized by the pounding of feet on the floor to mark the rhythm) are danced on festive occasions, while the *sahjeh* is a well-known Bedouin dance. The Circassian minority has a sword dance, as well as several other Cossack dances. Government interest in preserving folk arts has resulted in the formation of a national troupe that is regularly featured on state radio and television programs.

Newspapers are privately owned and extensively regulated. There are several literary magazines and scientific and topical periodicals. Most professional groups and government departments issue their own periodicals. Radio and television stations, which are government-owned, feature programs from both Arab and foreign, mostly Western, countries. Most major towns have movie theatres that offer both Arab and foreign films. There is no legitimate theatre in Jordan, but amateur groups perform in institutions of learning, on radio and television, and in the various foreign cultural centres in Amman, Irbid, and Jerusalem.

For statistical data on the land and people of Jordan, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL.

History

Modern Jordan occupies an area rich in archaeological and religious associations. The Jordanian desert appears to have been the home of hunters from Lower Paleolithic times, as their flint tools are found widely distributed. In the southeast of the country, at Jabal aṭ-Ṭubayq, there are rock carvings of all periods, the earliest of which have been attributed to the Paleolithic-Mesolithic. As the lowest of deposits show, mesolithic hunters appear to have visited Jericho (Tall as-Sulṭān). A series of later levels reveal, first, round houses and then plastered rectangular houses built by people who did not know the use of pottery. Following these were two pottery levels and a series of deposits on the site and in the neighbouring tombs, extending to the Late Bronze Age. This site, excavated first by John Garstang and later by K.M. Kenyon, is important also for its massive pre-pottery defensive works, dated by radiocarbon tests to about 7000 BC. In the Chalcolithic period the neighbouring site of Tulaylāt al-Ghassūl (also in the Jordan Valley) shows a well-built village of *c.* 4500–3000 BC with painted plaster walls.

Early Bronze Age deposits

The Early Bronze Age (*c.* 3000–2100 BC) is marked by deposits at the base of Dhībān; but although many sites have been found in the north of the country few have been excavated, and little evidence of settlement in this period is found south of ash-Shawbak. The Early Bronze Age was terminated by a nomadic invasion that destroyed the principal towns and villages and marked the end of a period of apparently peaceful development. Security was not again reestablished until the advent of the Egyptians after 1580 BC. It was once thought that the area was unoccupied between 1900 and 1300 BC, but a systematic archaeological survey has shown that the country had a settled population throughout the period. At Amman a small temple with Egyptian, Mycenaean, and Cypriot imported objects has been found confirming this.

BIBLICAL ASSOCIATIONS

From the Middle Bronze Age onward there are biblical accounts of the area, mentioning kingdoms such as Gilead in the north and those of Moab, in central Jordan, and Midian. At the time of the Exodus the Israelites tried to pass through Edom in southern Jordan but were refused permission. They were at first repulsed by the Amorites, whom they later defeated. The Israelite tribes of Gad and Reuben and half of the tribe of Manasseh settled in the conquered territory of the Ammonites, Amorites, and Bashan and rebuilt many of the towns they had partially destroyed. A nearly contemporary record of this period is the Mesha or Moabite stone found at Dhībān in 1868 and now in the Louvre, Paris. It is inscribed in an eastern form of Canaanite, closely akin to Hebrew.

The next few centuries (1300–1000 BC) were marked by constant raiding from both sides of the Jordan. David attacked Moab and Edom, killing two-thirds of the population of Moab and all the males of Edom. Although Ammon with its capital, Rabbath Ammon (modern Amman), was held for a time, it regained its independence on the death of David (*c.* 960 BC). Solomon had a port on the Gulf of Aqaba at Ezion-geber, later Elat in Israel, where copper ore was smelted from mines in the Wadi al-ʿArabah and trade carried on with the southern Arabian states.

Hostilities remained constant between Judah and Edom. A Hebrew king, Amaziah, even captured Sela (Petra), the capital, and slew 10,000 prisoners. The next invaders were the Assyrians, who under Adadnirari III (811 or 810–782 BC) overran the eastern part of the country as far as Edom. Revolts led to the retaking of the country by Tiglath-pileser III (reigned 745–727 BC) in the first year of his reign and to the division of the country into provinces under Assyrian governors. This policy of direct rule continued until the fall of the Assyrian empire in 612 BC. The Assyrian texts are the first to refer to the Nabataeans, who at this time occupied the land south and east of Edom (ancient Midian). After the fall of Assyria the Moabites and Ammonites continued to raid Judah until the latter was conquered by the Neo-Babylonians under Nebuchadrezzar II. Little is known of the history of Jordan under the Neo-Babylonians and Persians, but during this period the Nabataeans infiltrated into Edom and forced the Edomites out into southern Palestine.

It was not until the Hellenistic rule of the Seleucids and the Ptolemies that the country prospered, trade increased, and new towns were built. Rabbath Ammon was renamed Philadelphia, and Jarash became Antioch-on-the-Chrysorrhoas, or Gerasa. Hostilities between the Seleucids and Ptolemies enabled the Nabataeans to extend their kingdom northward and to increase their prosperity based on the caravan trade with Arabia and Syria. The northern part of Jordan was for a time in Jewish hands, and there were constant struggles between the Jewish Maccabees and the Seleucids. It is to this period that the majority of the Dead Sea Scrolls may be attributed.

During 64–63 BC the kingdom of Nabataea was conquered by the Romans under Pompey, who restored the Hellenistic cities destroyed by the Jews and set up the Decapolis. The country remained independent but paid imperial taxes.

Jordan under Roman rule

Roman policy seems to have been to maintain Nabataea as a buffer state against the desert tribes. In 25–24 BC it served as a starting point for Aelius Gallus' ill-starred expedition in search of Arabia Felix. Nabataea was finally absorbed into the Roman Empire by Trajan in AD 106 as the province of Palaestina Tertia. Under Roman rule Jordan prospered, and many new towns and villages were established. The whole country, except the Decapolis, was made part of the new province called Arabia Petraea, with its capital first at Petra and later at Buṣrá as-Shām in Syria. After AD 313 Christianity became a recognized religion, and a large number of churches were built.

THE LATIN KINGDOM AND MUSLIM DOMINATION

The whole area was devastated in the 6th and 7th centuries AD by the intermittent warfare between Byzantium and Sāsānian Persia. In AD 627 the emperor Heraclius finally defeated the Persians and reestablished order in the area, but Byzantium had been gravely weakened by the long struggle and was left in a state of exhaustion to face the totally unexpected menace of a new power that had arisen in Arabia. In AD 636 the Muslims, led by the famous "Sword of Islām," Khālid ibn al-Walīd, destroyed a Byzantine army at the Battle of the Yarmūk River and brought the greater part of Syria and Palestine under Muslim rule.

The caliphs of the Umayyad dynasty (AD 660–750) established their capital at Damascus and built splendid hunting lodges and palaces in the Jordanian desert. These can still be seen at sites such as Qaṣr ʿAmrah, al-Kharānah, aṭ-Ṭūbah, and Qaṣr al-Mshattā. Many Roman forts were rebuilt. After the seizure of power by the ʿAbbāsids in AD 750, the capital was transferred to Baghdad, and Syria, which had been the Umayyad metropolitan province and from which they had drawn most of their support, was severely repressed. Jordan, now distant from the centre of power, became a backwater and slowly reverted to the old Bedouin way of life. With the capture of Jerusalem by the crusaders in AD 1099, the Latin kingdom of Jerusalem was extended east of the Jordan, and a principality known as Oultre Jourdain was set up. A capital was established at al-Karak. After the crusaders retreated, the history of Jordan remained uneventful. In the 16th century it submitted to Ottoman rule and became part of the *vilayet* of Damascus.

In the 19th century the Turks settled Circassian, Caucasian, and other refugees in Transjordan to protect their communications with Arabia; in 1908 they completed the Hejaz railway linking Damascus and Medina. As a territorial state Jordan is a creation of the 20th century.

TRANSJORDAN, THE HASHEMITE KINGDOM, AND THE PALESTINE WAR

The Conference of San Remo

World War I led in 1916 to the Arab revolt against the Turks and to the cutting of the Hejaz railway, in which T.E. Lawrence played an important role. Al-ʿAqabah was taken in July 1917, and by October 1918 Damascus had fallen into Allied hands with the assistance of the Arab forces in pinning down the Turkish Army in Jordan. In 1920 the Conference of San Remo allotted the Palestine mandate to Great Britain and the Syrian mandate to France, thus effectively separating the areas now covered by Israel and Jordan from Syria, with which they had previously been linked. In November 1920 Abdullah of the Hashemite house arrived in Maʿān, then part of the Hejaz, accompanied by a group of armed supporters and intent on raising the tribes to attack the French, who had forced his brother Fayṣal to relinquish his newly founded kingdom in Syria. Abdullah was persuaded by the British government to refrain and, instead, to take over the government of what became known as Transjordan (1921).

Transjordan, although part of the Palestinian mandate, was expressly excluded from the clauses regarding the establishment of a "Jewish national home." In May 1927, Transjordan was formally recognized as an independent constitutional state, although it remained under British tutelage. In 1939 an elected Cabinet took the place of an Executive Council, and the emir was authorized to raise military forces and to open independent Transjordanian consulates. After World War II, in 1946, a new treaty was signed with Britain, and in May of that year Abdullah became the first king of the Hashemite Kingdom of Transjordan (later Jordan). Two years later the Palestinian mandate ended, and the Transjordanian Arab Legion, commanded by Glubb Pasha (John [later Sir John] Bagot Glubb), joined Egyptian, Syrian, Lebanese, and Iraqi troops in an attempt to prevent the creation of a Jewish state. A considerable enclave of Palestinian territory—including Hebron (al-Khalīl), Bethlehem (Bayt Laḥm), Ram Allah, and Nābulus, as well as part of Jerusalem (the Old City)—totaling about 2,000 square miles remained in Jordanian hands at the time of the signing of the Jordanian-Israeli armistice pact on April 3, 1949. This territory was formally annexed in 1950, but it brought with it the problem of a large refugee population that, on the whole, was hostile to the Hashemite regime.

Assassination of King Abdullah

King Abdullah was assassinated by a follower of Haj Amīn al-Ḥusaynī, the former mufti of Jerusalem, on July 20, 1951. Abdullah's elder son, Talal, succeeded him, but mental illness caused his deposition by the parliament in August 1952. Talal's son Hussein was crowned in Amman at his coming of age (by the Muslim calendar) on May 2, 1953.

KING HUSSEIN

The early years of Hussein's reign were beset by problems and difficulties. In 1955 Jordan refused to join the Baghdad Pact. In 1956 Hussein dismissed his British advisers, including Glubb, and later that year the Jordanian treaty with Great Britain was abrogated. In 1957 Hussein survived the first of several military plots to overthrow his regime; political parties were also dissolved by royal decree. (K.S.A.J./Ed.)

After Egypt's union with Syria in the establishment of the United Arab Republic (U.A.R.) in 1958, fear of the consequences for his regime led Hussein to conclude a federal union with Iraq. The overthrow of the Iraqi monarchy in the same year brought a quick end to this federation; subsequently, British paratroops were flown to Jordan to prevent the overthrow of the monarchy by a Nasserist army and by Palestinians.

Syria's secession from the U.A.R. in 1961 brought temporary relief to the regime in Jordan. During the years that followed, U.S. aid made it possible for Hussein to promote considerable economic development in his country. There was also a boom in tourism, centred on the Old City of Jerusalem. The political climate, however, was less positive for Hussein. His Palestinian subjects, who included some of the best educated and most talented elements in the Arab world, objected to the king's suspension of the constitution, outlawing of political parties, and establishment of a royal dictatorship based largely on tribal and foreign support.

Formation of the Palestine Liberation Organization

Until 1964 the Palestinians had been leaderless and unorganized. They had received United Nations aid supplemented by the provision of cheap labour in the "host" countries. Their cause had been argued for them by others, and such political activity as they evinced had come about in relationship with one or other of the Arab states or leaders. In 1964, however, an Arab summit meeting led to the formation of the Palestine Liberation Organization (PLO). About the same time, Palestinians formed a secret organization called the Palestine National Liberation Movement, known from a reversal of its Arabic initials as al-Fatah. Both the PLO and al-Fatah undertook the training of guerrilla units for raids on Israel. (W.L.O.)

By the end of 1966, relations with Israel had undergone a marked deterioration, and Arab raids emanating mainly from Jordan provoked sharp reprisals. In December there were armed clashes on the border with Syria, as Jordan attempted to prevent Syrian-based Palestinian infiltration through Jordan into Israel. Jordan finally broke off diplomatic relations with Syria in May 1967, but this did not prevent it from proclaiming its solidarity with Syria in the face of the alleged Israeli military buildup near the Syrian frontier and from signing a defense pact with Egypt. When war broke out between Israel and the Arab states on June 5, 1967, Jordan was fully committed, and its forces were engaged almost immediately. Within 48 hours Israeli forces had overrun the whole of the territory west of the river, involving the capture of Bethlehem, Hebron, Jericho, Nābulus, Rām Allāh, and Janīn, as well as the entire city of Jerusalem. Jordan suffered heavy casualties, lost one-third of its most fertile land, and was faced with some 200,000 new refugees for an already overburdened economy to support.

Jordan's losses were in part made good by foreign aid, notably from Saudi Arabia, Kuwait, and Libya, but the progress that had been made in Jordan's economy prior to June 1967 was reversed. Another legacy of the June war was a considerable increase in guerrilla activities aimed at the overthrow of the Hashemite dynasty as a prelude to resuming war with Israel.

CIVIL WAR

Relations between the guerrillas and the government steadily deteriorated. In February 1970 the government published a ban on the carrying of arms in public in towns, the storage of arms and explosives, and activity by political parties. There were clashes with the army, and the government was forced to suspend the decrees. Life in Amman came to a standstill. Eventually a compromise was reached, and tension was reduced. The government agreed to give full support for the war against Israel, while the guerrillas undertook to discipline their members and to respect Jordan's internal security.

Under the terms of a peace plan drafted by U.S. Secretary of State William Rogers, Israel would give up all occupied Arab territory in exchange for a binding peace treaty and other concessions. Following Jordan's adherence to the Rogers peace plan, which was rejected by all the guerrilla organizations, there were demonstrations in Amman, and the guerrillas declared their intention to escalate the struggle. The Rogers plan was abandoned in 1971.

Popular Front for the Liberation of Palestine

In September 1970 a Marxist guerrilla group, the Popular Front for the Liberation of Palestine (PFLP), hijacked three airliners to Dawson's Field, a desert airstrip near Amman, holding the crews and passengers hostage against the release of guerrillas detained in Great Britain, West Germany, Switzerland, and Israel. Meanwhile, the situation in Jordan became chaotic, and outbreaks of fighting in Amman developed rapidly into full-scale civil war. On September 16 a curfew was imposed, and Jordan was virtually cut off from the outside world. A Syrian attempt at invasion on behalf of the guerrillas was repelled, and, after nine days of bitter fighting, a peace agreement was reached. Intensive efforts at mediation by the other Arab states were concluded at a Cairo summit meeting on September 27. An agreement involving substantial concessions to the guerrillas was signed by King Hussein and the guerrilla leader Yasir Arafat, but tension persisted.

During the early months of 1971 there were outbreaks of serious fighting during which the guerrillas were forced out of Amman and back toward the Syrian frontier. Despite their protests, the other Arab governments failed to intervene, and by April Hussein was strong enough to demand the guerrillas' withdrawal from Amman. In July their remaining strongholds in the north were destroyed by the Jordanian Army.

The suppression of the Palestinian resistance movements by the Jordanian government was bitterly resented, and reprisals were threatened. Members of the "Black September" organization shot and fatally wounded Prime Minister Wasfi at-Tal outside a hotel in Cairo on November 28, and an unsuccessful attempt was made on the life of the Jordanian ambassador in London.

JORDAN AFTER CIVIL WAR

On March 15, 1972, in Amman, King Hussein proposed a plan for a federated "United Arab Kingdom," to consist of the Gaza Strip linked to the West and East banks of the Jordan River, each bank to have its own elected parliament under a federal parliament, with Hussein as supreme head of state. The plan met with mixed reactions from Israel and hostility from the rest of the Arab world, excepting Morocco and Saudi Arabia. Egypt severed diplomatic relations with Jordan over the plan in April. (K.S.A.J.)

In February 1973 a group of Palestinian guerrillas, arrested after entering Jordan from Syria, was sentenced to death by a military court for aiming to overthrow the government. The sentences were commuted by King Hussein after mediation efforts by other Arab leaders. In September 1973 Jordan restored diplomatic relations with Egypt and in October with Syria.

In the Arab-Israeli War of October 1973, Jordan sent troops to the Syrian front but did not open a third front along the Jordan River. The war brought into the open the conflicting claims of Hussein and Yasir Arafat, leader of the PLO, to represent the Palestinian Arabs. In November King Hussein suggested to Egypt and Syria that they accept Jordan's claim to sovereignty over the West Bank and East Jerusalem; once this was accepted the West Bank population could determine its government through a plebiscite. After rejection of the proposal, in May 1974 King Hussein said that he would recognize the PLO as the sole legitimate representative of all the Palestinian people if all the other Arab states would also recognize the PLO. Jordan would, however, abandon its responsibility for recovery of lost Palestinian territory. At an Arab summit meeting in Rabat, Mor., on October 26–28, King Hussein, under pressure from the Arab states, signed a resolution naming the PLO as the sole legitimate representative of the Palestinian people and affirming their right to an independent state in their homeland. In return Jordan reportedly was to receive $300,000,000 in aid from Saudi Arabia.

In 1975 and 1976 Jordan made progress in reestablishing closer relations with other Arab states. Jordan's relations with the United States, however, were strained over a major arms deal opposed by the U.S. Congress. Jordan eventually decided to purchase Hawk missiles, despite the conditions set on the sale.

The reluctance of the United States to supply arms and the Egyptian-Israeli Sinai agreement of 1975 led Jordan to pursue closer relations with Syria. A Jordanian-Syrian joint high commission was formed in 1976 to coordinate the countries' foreign policies and armed forces, and a major economic and trade agreement was made. In December 1976 King Hussein and Syrian president Ḥafiz al-Assad announced their intention to unite the countries.

In 1977 steps were taken toward a reconciliation with the PLO. Several meetings were held in early 1977 with PLO representatives. King Hussein rejected Egyptian president Anwar el-Sādāt's proposal of the establishment of a Jordanian-Palestinian link before the Geneva peace conference.

In April 1978 King Hussein formed a 60-member National Consultative Council to replace the parliament, which had been dissolved in 1974 and briefly reconvened in 1976. The power of the council was reduced to debating proposed legislation. In January 1984, however, King Hussein reconvened the National Assembly, which appointed new West Bank representatives. Elections were held in March to fill the empty East Bank seats, with women being included in the electorate for the first time.

Relations with other nations

Relations with the other Arab nations fluctuated during the 1980s in response to the Egyptian-Israeli peace talks, the Iran–Iraq War, and attempts at negotiation between Jordan and the PLO. Jordan had broken off diplomatic relations with Egypt following the Camp David accords of 1979 between Egypt and Israel, but in September 1984 full diplomatic relations were again established. Differences between Jordan and Syria led to a confrontation in November 1980; both countries mobilized troops along their border, but no fighting occurred. When the Iran–Iraq War broke out in 1980, Jordan supported Iraq and became its major supplier of arms. Relations with the United States became strained when the U.S. Congress moved to restrict sales of arms to Jordan in 1986, and Jordan then looked to the Soviet Union and France for additional military and economic aid.

On July 31, 1988, King Hussein announced that Jordan would surrender to the PLO all claim to the Israeli-occupied West Bank and sever all legal and administrative ties to the territory. Earlier Hussein had dissolved the lower house of Jordan's parliament, half of whose members represented West Bank districts, and canceled a West Bank development plan. Jordan stopped paying salaries to Palestinian civil servants, teachers, and health workers on the West Bank, and West Bank residents lost their legal status as Jordanian citizens.

For later developments in the history of Jordan, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 911, 924, 962, 96/11, and 978. (Ed.)

BIBLIOGRAPHY. ABDULLAH, KING OF JORDAN, *My Memoirs Completed* (Eng. trans. from the Arabic, 1954); MUḤAMMAD ABŪ ḤASSĀN, *Turāth al-badw al-qaḍāʾī* (1974), a study in Arabic of the life-styles, customs, and heritage of the Bedouin; NASEER H. ARURI, *Jordan: A Study in Political Development* (1972), a discussion of the political development of Jordan; C.S. COON, *Caravan: The Story of the Middle East,* rev. ed. (1958), an anthropological study of the Middle Eastern peoples and their environment; C.D. CREMEANS, *The Arabs and the World* (1963), a discussion of Arab politics, with a section on each country; H.M. DAVIS (comp.), *Constitutions, Electoral Laws, Treaties of States in the Near and Middle East,* 2nd ed. rev. (1953); FAO MEDITERRANEAN DEVELOPMENT PROJECT, *Jordan* (1967), a study of Jordan's economy, with emphasis on social and agricultural aspects; W.B. FISHER, *The Middle East: A Physical, Social, and Regional Geography,* 6th ed. (1970); SIR JOHN BAGOT GLUBB, *A Soldier with the Arabs* (1959), the author's personal insights; L.G. HARDING, *The Antiquities of Jordan,* rev. ed. (1967), an excellent résumé; G.L. HARRIS, *Jordan: Its People, Its Society, Its Culture* (1958), a general reference book; RICHARD F. NYROP (ed.), *Jordan: A Country Study,* 3rd ed. (1980), a general work; A.H. HOURANI, *Minorities in the Arab World* (1947), a scholarly account of the various minority groups and their backgrounds; J.C. HUREWITZ, *Middle East Politics: The Military Dimension* (1969), a voluminous work with emphasis on the role of the military; HUSSEIN, KING OF JORDAN, *Uneasy Lies the Head* (1962), an account that traces his early life and experiences in Jordan, and *Mihnatī Ka-Malik* (1978), a personal exposé of Jordanian politics; INTERNATIONAL BANK FOR RECONSTRUCTION AND DEVELOPMENT, *The Economic Development of Jordan* (1957); A.M. LABADI, *Land and Water Use in the Hashemite Kingdom of Jordan* (1969); G. MOUNTFORT, *Portrait of a Desert* (1965), an illustrated description of life in the desert of Jordan; NATIONAL PLANNING COUNCIL, *The Jordan Five Year Socio-Economic Development Plan, 1981–1985* (1981), an outline of the social and economic plans for the five-year period; R. PATAI, *The Kingdom of Jordan* (1958), a comprehensive general work; G. SPARROW, *Modern Jordan* (1961), a personal account of the author's travels; CHRISTINE OSBORNE, *An Insight and Guide to Jordan* (1981), an overview of the landscape, society, and culture; Y.T. TONI, *Jordan: A Geographical Introduction* (1968); BAHA-UDDIN TOUKAN, *A Short History of Trans-Jordan* (1945); P.J. VATIKIOTIS, *Politics and the Military in Jordan* (1967), a historical sketch outlining the influence of the military on Jordanian politics; SHELAGH WEIR, *The Bedouin* (1976), an illustrated study of the arts and crafts of the Bedouin of Jordan.

(K.S.A.J.)

Judaism

Judaism, the religion of the Jews, is the complex expression of a religious and ethnic community, a way of life as well as a set of basic beliefs and values, which is discerned in patterns of action, social order, and culture as well as in religious statements and concepts. The first section of this article treats the history of Judaism in the broadest and most complete sense, from the early ancestral beginnings of the Jewish people down to contemporary times. In the second section the beliefs, practices, and culture of Judaism are discussed. Dates are listed throughout as BCE (Before the Common Era = BC) and CE (Common Era = AD).

The article is organized as follows:

THE HISTORY OF JUDAISM

It is history that provides the clue to an understanding of Judaism, for its primal affirmations appear in early historical narratives. Many contemporary scholars agree that although the biblical (Old Testament) tales report contemporary events and activities, they do so for essentially theological reasons. Such a distinction, however, would have been unacceptable to the authors, for their understanding of events was not superadded to but was contemporaneous with their experience or report of them. For them, it was primarily within history that the divine presence was encountered. God's presence was also experienced within the natural realm, but the more immediate or intimate disclosure occurred in human actions. Although other ancient communities saw a divine presence in history, this was taken up in its most consequent fashion within the ancient Israelite community and has remained, through many developments, the focus of its descendants' religious affirmations. It is this particular claim—to have experienced God's presence in human events—and its subsequent development that is the differentiating factor in Jewish thought. As ancient Israel believed itself through its history to be standing in a unique relationship to the divine, this basic belief affected and fashioned its lifestyle and mode of existence in a way markedly different from groups starting with a somewhat similar insight. The response of the people Israel to the divine presence in history was seen as crucial not only for itself but for all mankind. Further, God had—as person—in a particular encounter revealed the pattern and structure of communal and individual life to this people. Claiming sovereignty over the people because of his continuing action in history on its behalf, he had established a *berit* ("covenant") with it and had required from it obedience to his Torah (teaching). This obedience was a further means by which the divine presence was made manifest—expressed in concrete human existence. The corporate life of the chosen community was thus a summons to the rest of mankind to recognize God's presence, sovereignty, and purpose—the establishment of peace and well-being in the universe and in mankind.

God's presence in history

History, moreover, disclosed not only God's purpose but also manifested man's inability to live in accord with it. Even the chosen community failed in its obligation and had, time and again, to be summoned back to its responsibility by divinely called spokesmen—the prophets—who warned of retribution within history and argued and reargued the case of affirmative human response. Israel's role in the divine economy and thus Israel's particular culpability were dominant themes sounded against the motif of fulfillment, the ultimate triumph of the divine purpose, and the establishment of divine sovereignty over all mankind.

General observations

NATURE AND CHARACTERISTICS

In nearly 4,000 years of historical development, the Jewish people and their religion have displayed both a remarkable adaptability and continuity. In their encounter with the great civilizations, from ancient Babylonia and Egypt down to Western Christendom and modern secular culture, they have assimilated foreign elements and integrated them into their own socioreligious system, thus maintaining an unbroken line of ethnic and religious tradition. Furthermore, each period of Jewish history has left behind it a specific element of a Judaic heritage that continued to influence subsequent developments, so that the total Jewish heritage at any time is a combination of all these successive elements along with whatever adjustments and accretions are imperative in each new age.

The fundamental teachings of Judaism have often been grouped around the concept of an ethical (or ethical-historical) monotheism. Belief in the one and only God of Israel has been adhered to by professing Jews of all ages and all shades of sectarian opinion. By its very nature monotheism ultimately postulated religious universalism, although it could be combined with a measure of particularism. In the case of ancient Israel (see below *Biblical Judaism*), particularism took the shape of the doctrine of election; that is, of a people chosen by God as "a kingdom of priests and a holy nation" to set an example for all mankind. Such an arrangement presupposed a covenant between God and the people, the terms of which the chosen people had to live up to or be severely punished. As the 8th-century-BCE prophet Amos expressed it: "You only have I known of all the families of the earth; therefore I will punish you for all your iniquities." Further, it was a concept that combined with the messianic idea, according to which, at the advent of the Redeemer, all nations would see the light, give up war and strife, and follow the guidance of the Torah (divine guidance, teaching, or law) emanating from Zion (a hill in Jerusalem that has a special spiritual significance). With all its variations in detail, messianism has, in one form or another, permeated Jewish thinking throughout the ages and, under various guises, has coloured the outlook of many secular-minded Jews (see also DOCTRINES AND DOGMAS: *Eschatology*).

Law became the major instrumentality by which Judaism was to bring about the reign of God on earth. In this case law meant not only what the Romans called *jus* (human law) but also *fas,* the divine or moral law that embraces practically all domains of life. The ideal, therefore, as expressed in the Ten Commandments, was a religioethical conduct that involved ritualistic observance as well as individual and social ethics, a liturgical–ethical way constantly expatiated on by the prophets and priests, rabbinic sages, and philosophers. Such conduct was to be placed in the service of God, as the transcendent and immanent Ruler of the universe, and as such the Creator and propelling force of the natural world, and also as the One giving guidance to history and thus helping man to overcome the potentially destructive and amoral forces of nature. According to Judaic belief, it is through the historical evolution of man, and particularly of the Jewish people, that the divine guidance of history constantly manifests itself and will ultimately culminate in the messianic age. Judaism, whether in its "normative" form or its sectarian deviations, never completely departed from this basic ethical–historical monotheism.

(S.W.B./L.H.S.)

PERIODIZATION

The division of the millennia of Jewish history into periods—a procedure frequently dependent on individual preferences—has not been devoid of theological or scholarly presuppositions. The Christian world long believed that until the rise of Christianity the history of Judaism was but a "preparation for the Gospel" (*preparatio evangelica*) followed by the "manifestation of the Gospel" (*demonstratio evangelica*) as revealed by Christ and the Apostles. This formulation could be theologically reconciled with the assumption that Christianity had been preordained even before the creation of the world.

On the other hand, 19th-century biblical scholars moved the decisive division back into the period of the Babylonian Exile and restoration of the Jews to Judah (6th–5th centuries BCE). They asserted that after the first fall of Jerusalem (586 BCE) the ancient "Israelitic" religion gave way to a new form of the "Jewish" faith, or Judaism, as formulated by Ezra the Scribe and his school (5th century BCE). A German historian, Eduard Meyer, in 1896 published *Die Entstehung des Judentums* ("The Origin of Judaism"), in which he placed the origins of Judaism in the Persian period (see below *Biblical Judaism*) or the days of Ezra and Nehemiah (5th century BCE) and actually attributed to Persian imperialism an important role in shaping the new emergent Judaism.

These theories have been discarded by most scholars, however, in the light of a more comprehensive knowledge of the ancient Middle East and the abandonment of a theory of gradual evolutionary development that was dominant at the beginning of the 20th century. Most Jews share a long-accepted notion that there never was a real break in continuity and that Mosaic-prophetic-priestly Judaism was continued, with but few modifications, in the work of the Pharisaic and rabbinic sages (see below *Rabbinic Judaism*) well into the modern period. Even today the various Jewish groups, whether Orthodox, Conservative, or Reform, all claim direct spiritual descent from the Pharisees and the rabbinic sages. In actual historical development, however, many deviations have occurred from so-called normative or rabbinic Judaism.

In any case, the history of Judaism here is viewed as falling into the following major periods of development: biblical Judaism (*c.* 20th–4th century BCE), Hellenistic Judaism (4th century BCE–2nd century CE), rabbinic Judaism (2nd–18th century CE), and modern Judaism (*c.* 1750 to the present). (S.W.B.)

Biblical Judaism (20th–4th century BCE)

THE ANCIENT MIDDLE EASTERN SETTING

The family of the Hebrew patriarchs (Abraham, Isaac, and Jacob) is depicted in the Bible as having had its chief seat in the northern Mesopotamian town of Harran—then (mid-2nd millennium BCE) belonging to the Hurrian kingdom of Mitanni. From there Abraham, the founder of the Hebrew people, is said to have migrated to Canaan (comprising roughly the region of modern Israel and Lebanon)—throughout the biblical period and later ages a vortex of west Asian, Egyptian, and east Mediterranean ethnoculture. Thence the Hebrew ancestors of the people of Israel (named after the patriarch Jacob, also called Israel) migrated to Egypt, where they lived in servitude, and a few generations later returned to occupy part of Canaan. The Hebrews were seminomadic herdsmen and occasionally farmers, ranging close to towns and living in houses as well as tents.

Early Israelite culture

The initial level of Israelite culture resembled that of its surroundings; it was neither wholly original nor primitive. The tribal structure resembled that of West Semitic steppe dwellers known from the 18th-century-BCE tablets excavated at the north central Mesopotamian city of Mari; their family customs and law have parallels in Old Babylonian and Hurro-Semite law of the early and middle 2nd millennium. The conception of a messenger of God that underlies biblical prophecy was Amorite (West Semitic) and found in the tablets at Mari. Mesopotamian religious and cultural conceptions are reflected in biblical cosmogony, primeval history (including the Flood story in Gen. 6:9–8:22), and law collections. The Canaanite component of Israelite culture consisted of the Hebrew language and a rich literary heritage—whose Ugaritic form (which flourished in the northern Syrian city of Ugarit from the mid-

15th century to about 1200 BCE) illuminates the Bible's poetry, style, mythological allusions, and religiocultic terms. Egypt provides many analogues for Hebrew hymnody and wisdom literature. All the cultures among which the patriarchs lived had cosmic gods who fashioned the world and preserved its order, including justice; all had a developed ethic expressed in law and moral admonitions; and all had sophisticated religious rites and myths.

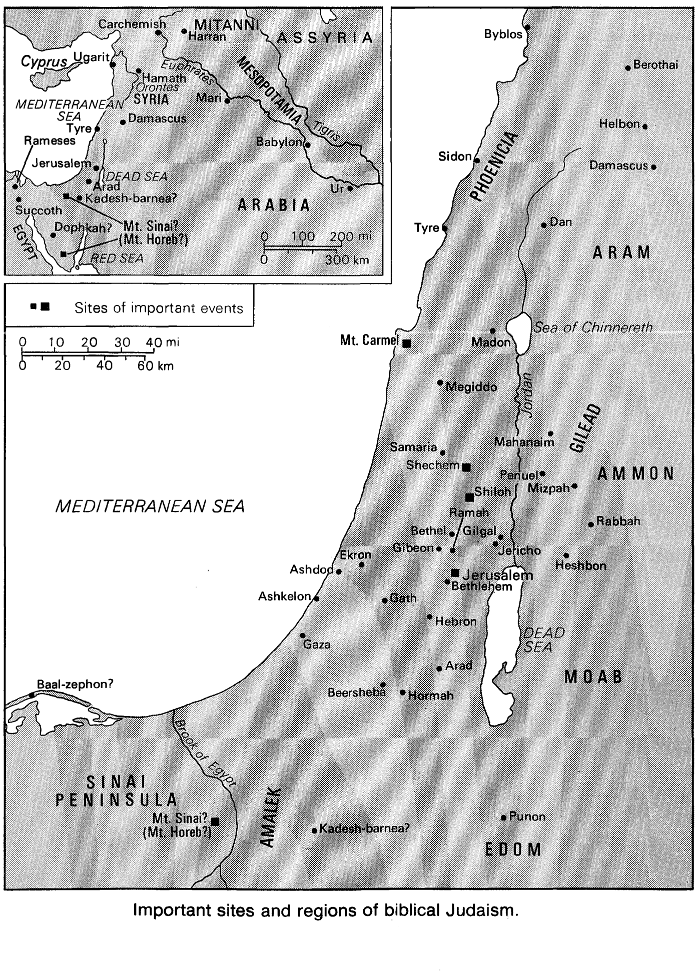
Though plainer when compared with some of the learned literary creations of Mesopotamia, Canaan, and Egypt, the earliest biblical writings are so imbued with contemporary ancient Middle Eastern elements that the once-held assumption that Israelite religion began on a primitive level must be rejected. Late-born amid high civilizations, the Israelite religion had from the start that admixture of high and low features characteristic of all the known religions of the area. Implanted on the land bridge between Africa and Asia, it was exposed to crosscurrents of foreign thought throughout its history.

THE PRE-MOSAIC PERIOD: THE RELIGION OF THE PATRIARCHS

The God of the patriarchs

Israelite tradition identified YHWH (by scholarly convention pronounced Yahweh), the God of Israel, with the Creator of the world, who had been known to and worshipped by men from the beginning of time. Abraham (perhaps 19th or 18th–17th centuries BCE) did not discover this God, but entered into a new covenant relation with him, in which he was promised the land of Canaan and a numerous progeny. God fulfilled that promise through the actions of the 13th-century-BCE Hebrew leader Moses: he liberated the people of Israel from Egypt, imposed Covenant obligations on them at Mt. Sinai, and brought them to the promised land.

Historical and anthropological studies present formidable objections to the continuity of YHWH worship from Adam (the biblical first man) to Moses, and the Hebrew tradition itself, moreover, does not unanimously support even the more modest claim of the continuity of YHWH worship from Abraham to Moses. Against it is a statement in chapter 6, verse 3, of Exodus that God revealed himself to the patriarchs not as YHWH but as El Shaddai—an epithet (of unknown meaning) the distribution of which in patriarchal narratives and Job and other poetical works confirms its archaic and unspecifically Israelite character. Comparable is the distribution of the epithet El Elyon (God Most High). Neither of these epithets appears in postpatriarchal narratives (excepting the Book of Ruth). Other compounds with El are unique to Genesis: El Olam (God the Everlasting One), El Bethel (God Bethel), and El Roʾi (God of Vision). An additional peculiarity of the patriarchal stories is their use of the phrase "God of my [your, his] father." All of these epithets have been taken as evidence that patriarchal religion differed from the worship of YHWH that began with Moses. A relation to a patron god was defined by revelations starting with Abraham (who never refers to the God of his father) and continuing with a succession of "founders" of his worship. Attached to the founder and his family, as befits the patron of wanderers, this unnamed deity (if indeed he was one only) acquired various Canaanite epithets (El, Elyon, Olam, Bethel, *qone eretz* [possessor of the Land]) only after their immigration into Canaan. Whether the name of YHWH was known to the patriarchs is doubtful. It is significant that while the epithets Shaddai and El occur frequently in pre-Mosaic and Mosaic-age names, YHWH appears as an element only in the names of Yehoshuaʿ (Joshua) and perhaps of Jochebed—persons who were closely associated with Moses.



Important sites and regions of biblical Judaism.

The patriarchs are depicted as objects of God's blessing, protection, and providential care. Their response is loyalty and obedience and observance of a cult whose ordinary expression is sacrifice, vow, and prayer at an altar, stone pillar, or sacred tree. Circumcision was a distinctive mark of the cult community. The eschatology (doctrine of ultimate destiny) of their faith was God's promise of land and a great progeny. Any flagrant contradictions between patriarchal and later mores have presumably been censored; yet distinctive features of the post-Mosaic religion are absent. The God of the patriarchs shows nothing of YHWH's "jealousy"; no religious tension or contrast with their neighbours appears, and idolatry is scarcely an issue. The patriarchal covenant differed from the Mosaic Sinaitic Covenant in that it was modelled upon a royal grant to favourites and contained no obligations, the fulfillment of which was to be the condition of their happiness. Evidently not the same as the later religion of Israel, patriarchal religion prepared the way for it in its familial basis, its personal call by the deity, and its response of loyalty and obedience to him.

Little can be said of the relation of the religion of the patriarchs to the religions of Canaan. Known points of contact between the two are the divine epithets mentioned above. Like the God of the fathers, El, the head of the Ugaritic pantheon was depicted both as a judgmental and a compassionate deity. Baal (Lord), the aggressive young agricultural deity of Ugarit, is remarkably absent from Genesis. Yet the socioeconomic situation of the patriarchs was so different from the urban, mercantile, and monarchical background of the Ugaritic myths as to render any comparisons highly questionable.

THE MOSAIC PERIOD: FOUNDATIONS OF THE ISRAELITE RELIGION

The Egyptian sojourn. According to Hebrew tradition, a famine caused the migration to Egypt of the band of 12 Hebrew families that later made up a tribal league in the land of Israel. The schematic character of this tradition does not impair the historicity of a migration to Egypt, an enslavement by Egyptians, and an escape from Egypt under an inspired leader by some component of the later league of Israelite tribes. To disallow these events would make their centrality as articles of faith in the later religious beliefs of Israel inexplicable.

The importance of Moses, the Exodus, and the Sinai Covenant

Tradition gives the following account of the birth of the nation. At the Exodus from Egypt (13th century BCE), YHWH showed his faithfulness and power by liberating Israel from bondage and punishing their oppressors with plagues and drowning at the sea. At Sinai, he made Israel his people and gave them the terms of his Covenant, regulating their conduct toward him and each other so as to make them a holy nation. After sustaining them miraculously during their 40-year wilderness trek, he enabled them to take the land that he had promised to their fathers, the patriarchs. Central to these events is God's apostle, Moses, who was commissioned to lead Israel out of Egypt, mediate God's Covenant to them, and bring them to Canaan.

Behind the legends and the multiform law collections, a historical figure must be posited to whom the legends and the legislative activity could be attached. And it is precisely Moses' unusual combination of roles that makes him credible as a historical figure. Like Muḥammad at the birth of Islām, Moses fills oracular, legislative, executive, and military functions. The main institutions of Israel are his creation: the priesthood and the sacred shrine, the Covenant and its rules, the administrative apparatus of the tribal league. Significantly, though Moses is compared to a prophet in various texts in the Pentateuch (the first five books of the Bible), he is never designated as one—the term being evidently unsuited for so comprehensive and unique a figure.

Mosaic religion. The distinctive features of Israelite religion appear with Moses. The proper name of Israel's God, YHWH, was revealed and interpreted to Moses as meaning *ehye asher ehye*—an enigmatic phrase (literally meaning "I am/shall be what I am/shall be") of infinite suggestiveness. The Covenant, defining Israel's obligations, is ascribed to Moses' mediation. Although it is impossible to determine what rulings go back to Moses, the Decalogue, or Ten Commandments, presented in chapter 20 of Exodus and chapter 5 of Deuteronomy, and the larger and smaller Covenant codes in Ex. 20:22–23:33; 34:11–26) are held by critics to contain early covenant law. From them, the following features may be noted: (1) The rules are formulated as God's utterances—*i.e.,* expressions of his sovereign will. (2) They are directed toward, and often explicitly addressed to, the people at large; Moses merely conveys the sovereign's message to his subjects. (3) Publication being of the essence of the rules, the people as a whole are held responsible for their observance (see also DOCTRINES AND DOGMAS: *Covenant*).

The liberation from Egypt laid upon Israel the obligation of exclusive loyalty to YHWH. This meant eschewing all other gods—including idols venerated as such—and the elimination of all magical recourses. The worship of YHWH was aniconic (without images); even such figures as might serve in his worship were banned—apparently owing to the theurgic overtones (the implication that through them men may influence or control the god by fixing his presence in a particular place and making him accessible). Though a mythological background lies behind some cultic terminology (*e.g.,* "a pleasing odor to YHWH," "my bread"), sacrifice is rationalized as tribute or (in priestly writings) is regarded purely as a sacrament; *i.e.,* as a material means of relating to God. Hebrew festivals also have no mythological basis; they either celebrate God's bounty (*e.g.,* at the ingathering of the harvest) or his saving acts (*e.g.,* the festival of unleavened bread, which is a memorial of the Exodus).

By courtesy of the Societe des Editions Verve, from "La Bible illustree par Chagall," *Verve,* no. 33/34

Moses breaking the tablets of the Law on Mt. Sinai in anger over the Israelites' worship of the golden calf. Etching by Marc Chagall.

Social values and concerns

The values of life and limb, labour, and social solidarity are protected in the rules on relations between man and man. The involuntary perpetual slavery of Hebrews is abolished, and a seven-year limit is set on bondage. The humanity of slaves is defended: one who beats his slave to death is liable to death; if he maims a slave he must set the slave free. A murderer is denied asylum and may not ransom himself from death, while for deliberate and severe bodily injuries the lex talionis ("an eye for an eye" principle) is ordained. Harm to property or theft is punished monetarily, never by death.

Moral exhortations call for solidarity with the poor and the helpless, for brotherly assistance to fellows in need. Institutions are created (*e.g.,* the sabbatical, or seventh, fallow year, in which land is not cultivated) to embody them in practice.

Since the goal of the people was the conquest of a land, their religion had warlike features. Organized as an army (called "the hosts of YHWH" in Ex. 12:41), they encamped in a protective square around their palladium—the tent housing the ark in which the stone "tablets of the Covenant" rested. When journeying, the sacred objects were carried and guarded by the Levites (a tribe serving religious functions), whose rivals, the Aaronites, had a monopoly on the priesthood. God, sometimes called "the warrior," marched with the army; in war, part of the booty was delivered to his ministers.

THE PERIOD OF THE CONQUEST AND SETTLEMENT OF CANAAN

The conquest of Canaan was remembered as a continuation of God's marvels at the Exodus. The Jordan River was split asunder, Jericho's walls fell at Israel's shout; the enemy was seized with divinely inspired terror; the sun stood still in order to enable Israel to exploit its victory. Such stories are not necessarily the work of a later age; they reflect rather the impact of these victories on the actors in the drama, who felt themselves successful by the grace of God.

A complex process of occupation, involving both battles of annihilation and treaty arrangements with the natives, has been simplified in the biblical account of Joshua's wars. Gradually, the unity of the invaders dissolved (most scholars believe that the invading element was only part of the Hebrew settlement in Canaan; other Hebrews, long since settled in Canaan from patriarchal times, then joined the invaders' covenant league). Individual tribes made their way with more or less success against the residue of Canaanite resistance. New enemies, Israel's neighbours to the east and west, appeared, and the period of the judges (leaders, or champions) began.

The religious situation during the time of the Book of Judges

The Book of Judges, the main witness for the period, does not speak with one voice on the religious situation. Its editorial framework describes repeated cycles of apostasy, oppression, appeal to God, and relief through a Godsent champion. The premonarchic troubles (before the kingship of Saul; see below) caused by the weakness of the disunited tribes were thus accounted for by the covenantal sin of apostasy. The individual stories, however, present a different picture. Apostasy does not figure in the exploits of the judges Ehud, Deborah, Jephthah, and Samson; YHWH has no rival, and faith in him is periodically confirmed by the saviours he sends to rescue Israel from their neighbours.

This faith is shared by all the tribes; and it is owing to their common cult that a Levite from Bethlehem could serve first at an Ephraimite and later also at a Danite sanctuary. The religious bond, preserved by the common cult, was enough to enable the tribes to act more or less in concert under the leadership of elders or an inspired champion in time of danger or religious scandal.

To be sure, both written and archaeological testimonies point to the Hebrews' adoption of Canaanite cults—the Baal worship of Gideon's family and neighbours in Ophrah in Judges, chapter 6, is an example. The many cultic figurines (usually female) found in Israelite levels of Palestinian archaeological sites also give colour to the sweeping indictments of the framework of the Book of Judges. But these phenomena belonged to the private, popular religion; the national God, YHWH, remained one—Baal sent no prophets to Israel—though YHWH's claim to exclusive worship was obviously not effectual. Nor did his cult conform with later orthodoxy; Micah's idol in Judges, chapter 17, and Gideon's ephod (priestly or religious garment) were considered apostasies by the editor, in accord with the dogma that other than orthodoxy there is only apostasy—heterodoxy (nonconformity) being unrecognized and simply equated with apostasy.

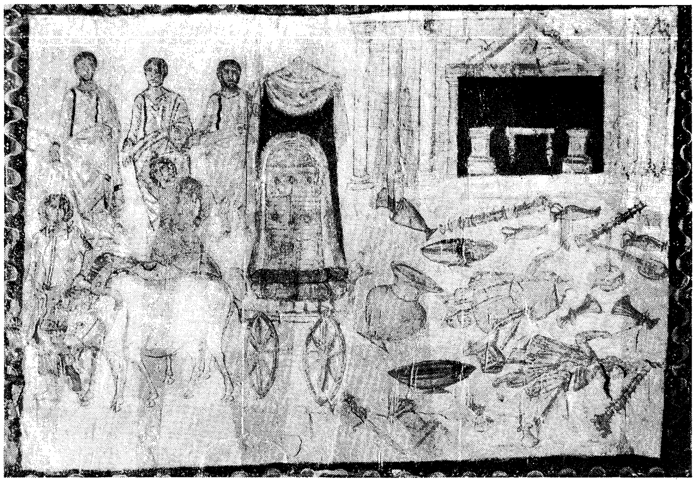
To the earliest sanctuaries and altars honoured as patriarchal foundations—at Shechem, Bethel, Beersheba, and Hebron in Cisjordan (west of the Jordan); at Mahanaim, Penuel, and Mizpah in Transjordan (east of the Jordan)—were now added new ones at Dan, Shiloh, Ramah, Gibeon, and Gilgal, among others. A single priestly family could not operate all these establishments, and Levites rose to the priesthood; at private sanctuaries even non-Levites might be consecrated as priests. The ark of the Covenant was housed in the Shiloh sanctuary, staffed by priests of the house of Eli, who traced their consecration back to Egypt. But the ark remained a portable palladium in wartime; Shiloh was not regarded as its final resting place. The law in Exodus, chapter 20, verses 24–26, authorizing a plurality of altar sites and the simplest forms of construction (earth and rough stone) suited the plain conditions of this period.

THE PERIOD OF THE UNITED MONARCHY

Conflicting views regarding the monarchy

The religiopolitical problem. The loose, decentralized tribal league could not cope with the constant pressure of external enemies—camel-riding desert marauders who pillaged harvests annually or iron-weaponed Philistines (an Aegean people settling coastal Palestine *c.* 12th century BCE) who controlled key points in the hill country occupied by Israelites. In the face of such threats to the Israelites, local, sporadic, God-inspired saviours had to be replaced by a continuous central leadership that could mobilize the forces of the entire league and create a standing army. Two attitudes were distilled in the crisis, one conservative and antimonarchic, the other progressive and promonarchic. The conservative appears first in Gideon's refusal, in Judges, chapter 8, verse 23, to found a dynasty: "I will not rule you," he tells the people, "my son will not rule over you; YHWH will . . . !" This theocratic view pervades one of the two contrasting accounts of the founding of the monarchy fused in chapters 8–12 of the First Book of Samuel: the popular demand for a king was viewed as a rejection of the kingship of God, which was embodied in a series of inspired saviours from Moses and Aaron, through Jerubbaal, Bedan, and Jephthah, to Samuel. The other account depicts the monarchy as a gift of God, designed to rescue his people from the Philistines (I Sam. 9:16). Both accounts represent the seer-judge Samuel as the key figure in the founding of Israel's monarchy, and it is not unlikely that the two attitudes struggled in his breast.

By courtesy of the Bollingen Foundation; photograph, Fred Anderegg



The triumph of the ark of the Covenant over paganism, a representation of the antagonism between Judaism and Hellenistic paganism. It was inspired by the biblical story (I Sam. 5:1–5) of the ark, which was captured in battle by the Philistines and which toppled the cult images of Dagon. Mural painting from the synagogue at Doura-Europus, Syria, 3rd century CE.

The Benjaminite Saul was made king (*c.* 1020 BCE) by divine election and by popular acclamation after his victory over the Ammonites (a Transjordanian Semitic people), but his career was clouded by conflict with Samuel, the major representative of the old order. Saul's kingship was bestowed by Samuel and had to be accommodated to the ongoing authority of that man of God. The two accounts of Saul's rejection by God (through Samuel) involve his usurpation of the prophet's authority. King David, whose forcefulness and religiopolitical genius established the monarchy (*c.* 1000 BCE) on an independent spiritual footing, resolved the conflict.

The Davidic monarchy. The essence of the Davidic innovation was the idea that, in addition to divine election through Samuel and public acclamation, David had God's promise of an eternal dynasty (a conditional, perhaps earlier, and an unconditional, perhaps later, form of this promise exist in Psalms, 132 and II Samuel, chapter 7, respectively). In its developed form, the promise was conceived of as a covenant with David, parallelling the Covenant with Israel and instrumental in the latter's fulfillment; *i.e.,* that God would channel his benefactions to Israel through the chosen dynasty of David. With this new status came the inviolability of the person of God's anointed (a characteristically Davidic idea) and a court rhetoric—adapted from pagan models—in which the king was styled "the [firstborn] son of God." An index of the king's sanctity was his occasional performance of priestly duties. Yet the king's mortality was never forgotten—he was never deified; prayers and hymns might be said on his behalf, but they were never addressed to him as a god.

The chosen dynasty and the chosen city

David captured the Jebusite stronghold of Jerusalem and made it the seat of a national monarchy (Saul had never moved the seat of his government from his home town, the Benjaminite town of Gibeah, about three miles north of Jerusalem). Then, fetching the ark from an obscure retreat, David installed it in his capital, asserting his royal prerogative (and obligation) to build a shrine for the national God—at the same time joining the symbols of the dynastic and the national covenants. This move of political genius linked the God of Israel, the chosen dynasty of David, and the chosen city of Jerusalem in a henceforth indissoluble union.

David planned to erect a temple to house the ark, but the tenacious tradition of the ark's portability in a tent shrine forced postponement of the project to his son Solomon's reign. As part of his extensive building operations, Solomon built the Temple on a Jebusite threshing floor, located on a hill north of the royal city, which David had purchased to mark the spot where a plague had been halted. The ground plan of the Temple—a porch with two free-standing pillars before it, a sanctuary, and an inner sanctum—followed Syrian and Phoenician sanctuary models. A bronze "sea" resting on bulls and placed in the Temple court has a Babylonian analogue. Exteriorly, the Jerusalem Temple resembled Canaanite and other Middle Eastern religious structures, but there were differences; *e.g.,* no god image stood in the inner sanctum, but rather only the ancient ark and the new large cherubim (hybrid creatures with animal bodies, human or animal faces, and wings) whose wings covered it, symbolizing the presence of YHWH who was enthroned upon celestial cherubim.

Alongside a brief, ancient inaugural poem in I Kings, chapter 8, verses 12–13, an extensive (and, in its present form, later) prayer expresses the distinctively biblical view of the temple as a vehicle of God's providing for his people's needs. Since, strangely, no reference to sacrifice is made, not a trace appears of the standard pagan conception of the temple as a vehicle of man's providing for the gods.

That literature flourished under the aegis of the court is to be gathered from the quality of the preserved narrative of the reign of David, which gives every indication of having come from the hand of a contemporary eyewitness. The royally sponsored Temple must have had a library and a school attached to it (in accord with the universally attested practice of the ancient Middle East), among whose products were not only royal psalms but also such liturgical pieces intended for the common man as eventually found their way into the book of Psalms.

The emergence of Torah and wisdom literature

The latest historical allusions in the Torah literature (the Pentateuch) are to the period of the united monarchy; *e.g.,* the defeat and subjugation of the peoples of Amalek, Moab, and Edom by Saul and David, in Numbers, chapter 24, verses 17–20. On the other hand, the polity reflected in the laws is tribal and decentralized, with no bureaucracy. Its economy is agricultural and pastoral, class distinctions apart from slave and free are lacking, and commerce and urban life are rudimentary. A premonarchic background is evident, with only rare explicit reflections of the later monarchy; *e.g.,* in Deuteronomy, chapter 17, verses 14–20. The groundwork of the Torah literature may thus be supposed to have crystallized under the united monarchy.

It was in this period that the traditional wisdom cultivated among the learned in neighbouring cultures came to be prized in Israel. Solomon is represented as the author of an extensive literature comparable to that of other Eastern sages. His wisdom is expressly attributed to YHWH in the account of his night oracle at Gibeon (in which he asked not for power or riches but for wisdom), thus marking the adaptation to biblical thought of this common Middle Eastern genre. As set forth in Proverbs, chapter 2, verse 5, "It is YHWH who grants wisdom; knowledge and understanding are by his command." Patronage of wisdom literature is ascribed to the later Judahite king, Hezekiah, and the connection of wisdom with kings is common in extrabiblical cultures as well.

Domination of all of Palestine entailed the absorption of "the rest of the Amorites"—the pre-Israelite population that lived chiefly in the valleys and on the coast. Their impact on Israelite religion is unknown, though some scholars contend that a "royally sponsored syncretism" arose with the aim of fusing the two populations. That popular religion did not meet the standards of the biblical writers and that it incorporated pagan elements—and that such elements may have increased as a result of intercourse with the newly absorbed "Amorites"—is likely and required no royal sponsorship. On the other hand, the court itself welcomed foreigners—Philistines, Cretans, Hittites, and Ishmaelites are named, among others—and made use of their service. Their effect on the court religion may be surmised from what is recorded concerning Solomon's many diplomatic marriages: foreign princesses whom Solomon married brought along with them the apparatus of their native cults, and the King had shrines to their gods built and maintained on the Mount of Olives. Such private cults, while indeed royally sponsored, did not make the religion of the people syncretistic.

Such compromise with the pagan world, entailed by the widening horizons of the monarchy, violated the sanctity of the holy land of YHWH and turned the king into an idolator in the eyes of zealots. Religious opposition, combined with grievances against the organization of forced labour for state projects, led to the secession of the northern tribes (headed by the Joseph tribes) after Solomon's death.

THE PERIOD OF THE DIVIDED KINGDOM

Religion in the northern kingdom (Israel)

Jeroboam I (10th century BCE), the first king of the north (now called Israel, in contradistinction to Judah, the southern Davidic kingdom), appreciated the inextricable link of Jerusalem and its sanctuary with the Davidic claim to divine election to kingship over all Israel (the whole people, north and south). He therefore founded rival sanctuaries at Dan and at Bethel—ancient cult sites—and manned them with non-Levite priests whose symbol of YHWH's presence was a golden calf—a pedestal of divine images in ancient iconography and the equivalent of the cherubim of Jerusalem's Temple. He also moved the autumn ingathering festival one month ahead so as to foreclose celebration of this most popular of all festivals in common with Judah.

For the evaluation of Jeroboam's innovations and the subsequent official religion of the north down to the mid-8th century, one must rely almost exclusively on the Book of Kings (later divided into two books). This work has severe limitations as a source for religious history. The material of this book, in good part contemporary, is subjugated to a dogmatic historiography that regards the whole enterprise of the north as one long apostasy ending in a deserved disaster. The culmination of Kings' history with the exile of Judah shows its provenience to have been Judahite. Yet the evaluation of Judah's official religion is subject to an equally dogmatic standard, namely, the royal adherence to the Deuteronomic rule of a single cult site. The author considered the Solomonic Temple to be the cult site chosen by God, according to Deuteronomy, chapter 12, the existence of which rendered all other sites illegitimate. Every king of Judah is judged according to whether or not he did away with all extra-Jerusalemite places of worship. (The date of this criterion may be inferred from the indifference toward it of all persons [*e.g.,* the 9th-century-BCE prophets Elijah and Elisha and the Jerusalemite priest Jehoiada] prior to the late-8th-century-BCE Judahite king Hezekiah.) Another serious limitation is the restriction of Kings' purview: excepting the Elijah–Elisha stories, it notices only the royally sponsored cult; notices of the popular religion are very few. From the mid-8th century the writings of the classical prophets, starting with Amos, set in. These take in the people as a whole, in contrast to Kings; on the other hand their interest in theodicy (justification of God) and their polemical tendency to exaggerate and generalize what they deem evil must be taken into consideration before approving their statements as sober history.

Religion in the southern kingdom (Judah)

For a half-century after the north's secession (*c.* 922 BCE), the religious situation in Jerusalem was unchanged. The distaff side of the royal household perpetuated, and even augmented, the pagan cults. King Asa (reigned *c.* 908–867 BCE) is credited with a general purge, including the destruction of an image made for the goddess Asherah by the queen mother, granddaughter of an Aramaean princess. He also purged the *qedeshim* ("consecrated men"—conventionally rendered as "sodomites," or "male sacred prostitutes").

Foreign cults entered the north with the marriage of the 9th-century-BCE king Ahab to the Tyrian princess Jezebel. Jezebel brought with her a large entourage of sacred personnel to staff the temple of Baal and Asherah that Ahab built for her in Samaria, the capital of the northern kingdom of Israel. In all else, Ahab's orthodoxy was irreproachable, though others of his court may have joined the worship of the foreign princess. That fierce opposition to the non-YWHW cults sprang up must be supposed in order to account for Jezebel's persecution of the prophets of YHWH, conduct untypical of a polytheist except in self-defense. Elijah's assertion that the whole country apostatized is a hyperbole based on the view that whoever did not actively fight Jezebel was implicated in her polluted cult. Such must have been the view of the prophets, whose fallen were the first martyrs to die for the glory of God. The quality of their opposition may be gauged by Elijah's summary execution of the foreign Baal cultists after they failed the test at Mt. Carmel, where they vied against him in a contest over whose god was truly God. A three-year drought (attested also in Phoenician sources), declared by Elijah to be punishment for the sin, must have done much to kindle the prophets' zeal.

To judge from the Elisha stories, the Baal worship in the capital city, Samaria, was not felt in the countryside.

There the religious tone was set by the popular prophets and the prophetic companies ("the sons of the prophets") who attached themselves to them. In popular consciousness these men were wonder-workers—healing the sick and reviving the dead, foretelling the future, and helping to find lost objects. To the biblical narrator they witness the working of God in Israel. Elijah's rage at the Israelite king Ahaziah's recourse to the pagan god Baalzebub, Elisha's cure of the Syrian military leader Naaman's leprosy, and anonymous prophets' directives and predictions in matters of peace and war all serve to glorify God. Indeed, the equation of Israel's prosperity with God's interest generated the issue of "true" and "false" prophecy that made its first appearance at this time. That prophecy of success could turn out to be a snare is exemplified in a story of conflict between Micaiah, the lone 9th-century-BCE prophet of doom, and 400 unanimous prophets of victory who lured Ahab to his death. The poignancy of the issue is highlighted by Micaiah's acknowledgment that the 400 were also prophets of YHWH—but inspired by him deliberately with a "lying spirit."

The issue of "true" and "false" prophecy

THE PERIOD OF CLASSICAL PROPHECY AND CULT REFORM

The emergence of the literary prophets. By the mid-8th century a hundred years of chronic warfare between Israel and Aram had finally ended—the Aramaeans having suffered heavy blows from the Assyrians. King Jeroboam II (8th century BCE) was able to undertake to restore the imperial sway of the north over its neighbour, and a prophecy of Jonah that he would extend Israel's borders from the Dead Sea to the entrance to Hamath (Syria) was borne out. The well-to-do expressed their relief in lavish attentions to the institutions of worship and their private mansions. But the strain of the prolonged warfare showed in the polarization of society between the wealthy few who had profited from the war and the masses whom it had ravaged and impoverished. Dismay at the dissolution of Israelite society animated a new breed of prophets who now appeared—the literary or classical prophets, first of whom was Amos, an 8th-century-BCE Judahite who went north to Bethel.

Prophetic denouncements of social injustices

That apostasy would set God against the community was an old conception of early prophecy; that violation of the sociomoral injunctions of the Covenant would have the same result was first proclaimed by Amos. Amos almost ignored idolatry, denouncing instead the corruption and callousness of the oligarchy and rulers. The religious exercises of such villains he proclaimed were loathsome to God; on their account Israel would be oppressed from the entrance to Hamath to the Dead Sea and exiled from its land.

The westward push of the Neo-Assyrian Empire in the mid-8th century BCE soon brought Aram and Israel to their knees. In 733–732 Assyria took Gilead and Galilee from Israel and captured Aramaean Damascus; in 721 Samaria, the Israelite capital, fell. The northern kingdom sought to survive through alliances with Assyria and Egypt; its kings came and went in rapid succession. The troubled society's malaise was interpreted by Hosea, a prophet of the northern kingdom (Israel), as a forgetting of God. As a result, in his view, all authority had evaporated: the king was scoffed at, priests became hypocrites, and pleasure seeking became the order of the day. The monarchy was godless; it put its trust in arms, fortifications, and alliances with the great powers. Salvation, however, lay in none of these, but in repentance and reliance upon God.

Prophecy in the southern kingdom. Judah was subjected to such intense pressure to join an Israelite–Aramaean coalition against Assyria that its 8th-century-BCE king Ahaz chose to submit himself to Assyria in return for relief. Ahaz introduced a new Aramaean-style altar in the Jerusalem Temple and adopted other foreign customs that are counted against him in the book of Kings. It was at this time that Isaiah prophesied in Jerusalem. At first (under Uzziah, Ahaz' prosperous grandfather), his message focussed on the corruption of Judah's society and religion, stressing the new prophetic themes of indifference to God (which went hand in hand with a thriving cult) and the fateful importance of social morality. Under Ahaz, the political crisis evoked Isaiah's appeals for trust in God, with the warning that the "hired razor from across the Euphrates" would shave Judah clean as well. Isaiah interpreted the inexorable advance of Assyria as God's chastisement; Assyria was "the rod of God's wrath." But since Assyria ignored its mere instrumentality and exceeded in an insolent manner its proper function, God, when he finished his purgative work, would break Assyria on Judah's mountains. Then the nations of the world, who had been subjugated by Assyria, would recognize the God of Israel as the lord of history. A renewed Israel would prosper under the reign of an ideal Davidic king, all men would flock to Zion (the hill symbolizing Jerusalem) to learn the ways of YHWH and submit to his adjudication, and universal peace would prevail (see also DOCTRINES AND DOGMAS: *Eschatology*)

The prophecy of Micah (8th century BCE), also a Judahite, was contemporary with that of Isaiah, and touched on similar themes (*e.g.,* the vision of universal peace is found in both their books). Unlike Isaiah however, who believed in the inviolability of Jerusalem, Micah shocked his audience with the announcement that the wickedness of its rulers would cause Zion to become a plowed field, Jerusalem a heap of ruins, and the Temple mount a wooded height. Moreover, from the precedence of social morality over the cult, Micah drew the extreme conclusion that the cult had no ultimate value and that God's requirement of men can be summed up as "to do justice, and to love kindness, and to walk humbly with your God."

Reforms in the southern kingdom. According to Jeremiah (about 100 years later), Micah's prophetic threat to Jerusalem had caused King Hezekiah (reigned *c.* 715–*c.* 686 BCE) to placate God—possibly an allusion to the cult reform instituted by the King in order to cleanse Judah from various pagan practices. A heightened concern over assimilatory trends resulted in his also outlawing certain practices considered legitimate up to his time. Thus, in addition to removing the bronze serpent that had been ascribed to Moses (and that had become a fetish), the reform did away with the local altars and stone pillars, the venerable (patriarchal) antiquity of which did not save them from the taint of imitation of Canaanite practice. Hezekiah's reform, part of a restorational policy that had political, as well as religious, implications, appears as the most significant effect of the fall of the northern kingdom on official religion. The outlook of the reformers is suggested by the catalog in II Kings, chapter 17, of religious offenses that had caused the fall, which the objects of Hezekiah's purge closely resemble. Hezekiah's reform is the first historical evidence for Deuteronomy's doctrine of cult centralization. Similarities between Deuteronomy and the Book of Hosea lend colour to the supposition that the reform movement in Judah, which culminated a century later under King Josiah, was sparked by attitudes inherited from the north.

The reform of King Hezekiah

Hezekiah was the leading figure in a western coalition of states that coordinated a rebellion against the Assyrian king Sennacherib with the Babylonian rebel Merodach-Baladan, shortly after the Assyrian's accession in 705 BCE. When Sennacherib appeared in the west in 701 the rebellion collapsed; Egypt sent a force to aid the rebels, but it was defeated. Hezekiah saw his kingdom overwhelmed and offered tribute to Sennacherib; the Assyrian, however, pressed for the surrender of Jerusalem. In despair, Hezekiah turned to the prophet Isaiah for an oracle. Though the prophet condemned the King's reliance upon Egyptian help, he stood firm in his faith that Jerusalem's destiny precluded its fall into heathen hands. The King held fast, and Sennacherib, for reasons still obscure, suddenly retired from Judah and returned home. This unlooked-for deliverance of the city may have been regarded as a vindication of the prophet's faith and was doubtless an inspiration to the rebels against Babylonia a century later. For the present, while Jerusalem was intact, the country had been devastated and its kingdom turned into a vassal state of Assyria.

During Manasseh's peaceful reign of 55 years in the 7th century BCE, Judah was a submissive ally of Assyria. Manasseh's forces served in the building and military

Foreign influences and the Deuteronomic reform

operations of the Assyrian kings Esarhaddon and Ashurbanipal. Judah benefitted from the upsurge of commerce that resulted from the political unification of the whole Near East. The prophet Zephaniah attests to heavy foreign influence on the mores of Jerusalem—merchants who adopted foreign dress, cynics who lost faith in the efficacy of YHWH to do anything, people who worshipped the pagan host of heaven on their roofs. Manasseh's court was the centre of such influences. The royal sanctuary became the home of a congeries of foreign gods—the sun, astral deities, and Asherah (the female fertility deity) all had their cults there alongside YHWH. The countryside also was provided with pagan altars and priests, alongside the local YHWH altars that were revived. Presumably, at least some of the blood that Manasseh is said to have spilled freely in Jerusalem must have belonged to YHWH's devotees. No prophecy is dated to his long reign.

With Ashurbanipal's death in 627, Assyria's power faded quickly; the young Judahite king Josiah (reigned *c.* 640–609 BCE) had already set in motion a vigorous movement of independence and restoration, a cardinal aspect of which was religious. First came the purge of foreign cults in Jerusalem, under the aegis of the high priest Hilkiah; then the countryside was cleansed. In the course of renovating the Temple, a scroll of Moses' Torah (by scholarly consensus an edition of Deuteronomy) was found. Anxious to abide by its injunctions, Josiah had the local YHWH altars polluted to render them unusable and collected their priests in Jerusalem. The celebration of the Passover that year was concentrated in the Temple, as it had not been "since the days of the judges who judged Israel," according to II Kings 23:22, or since the days of Samuel, according to II Chron. 35:18; both references reflect the unhistorical theory of the Deuteronomic (Josianic) reformers that the Shiloh sanctuary was the precursor of the Jerusalem Temple as the sole legitimate site of worship in Israel (as demanded by Deuteronomy, chapter 12). To seal the reform, the King convoked a representative assembly and had them enter into a covenant with God over the newfound Torah. For the first time, the power of the state was enlisted on behalf of the ancient covenant and in obedience to a covenant document. It was a major step toward the fixation of a sacred canon.

Josiah envisaged the restoration of Davidic authority over the entire domain of ancient Israel, and the retreat of Assyria facilitated his program—until he became fatally embroiled in the struggle of the powers over the dying empire. His death in 609 was doubtless a setback for his religious policy as well as his political aspirations. To be sure, the royally sponsored syncretism of Manasseh's time was not revived, but there is evidence of recrudescence of unofficial local altars. Whether references in Jeremiah and Ezekiel to child sacrifice to YHWH reflect post-Josianic practices is uncertain. There is stronger indication of private recourse to pagan cults in the worsening political situation.

That Assyria's fall should have been followed by the yoke of a harsh new heathen power dismayed the devotees of YHWH who had not been prepared for it by prophecy. Their mood finds expression in the oracles of the prophet Habakkuk in the last years of the 7th century BCE. Confessing perplexity at God's toleration of the success of the wicked in subjugating the righteous, the prophet affirms his faith in the coming salvation of YHWH, tarry though it might. And in the meantime, "the righteous must live in his faith."

Prophecies of Jeremiah and Ezekiel

But the situation in fact grew worse as Judah was caught in the Babylonian–Egyptian rivalry. Some attributed the deterioration to the burden of Manasseh's sin that still rested on the people. For the prophet Jeremiah (active *c.* 626–*c.* 580 BCE), the Josianic era was only an interlude in Israel's career of guilt that went back to its origins. His pre-reform prophecies denounced Israel as a faithless wife and warned of imminent retribution at the hands of a nameless northerner. After Nebuchadrezzar's decisive defeat of Egypt at Carchemish (605 BCE), Jeremiah identified the scourge as Babylon. King Jehoiakim's attempt to be free of Babylonia ended with the exile of his successor, Jehoiachin, along with Judah's elite (597); yet the court of the new king, Zedekiah, persisted in plotting new revolts, relying—against all experience—on Egyptian support. Jeremiah now proclaimed a scandalous doctrine of the duty of all nations, Judah included, to submit to the divinely appointed world ruler, the Babylonian monarch Nebuchadrezzar. In submission lay the only hope of avoiding destruction; a term of 70 years had been set to humiliate all men beneath Babylon. Imprisoned for demoralizing the populace, Jeremiah persisted in what was viewed as his traitorous message; the leaders, on their part, persisted in their policy, confident of Egypt and the saving power of Jerusalem's Temple, to the bitter end.

Jeremiah also had a message of comfort for his hearers. He foresaw the restoration of the entire people—north and south—in the land, under a new David. And since events had shown that man was incapable of achieving a lasting reconciliation with God on his own, he envisioned the penitent of the future being met halfway by God, who would remake their nature so that to do his will would come naturally to them. God's new covenant with Israel would be written on their hearts, so that they should no longer need to teach each other obedience, for young and old would know YHWH.

Among the exiles in Babylonia, the prophet Ezekiel, Jeremiah's contemporary, was haunted by the burden of Israel's sin. He saw the defiled Temple of Manasseh's time as present before his eyes, and described God as abandoning it and Jerusalem to their fates. Though Jeremiah offered hope through submission, Ezekiel prophesied an inexorable, total destruction as the condition of reconciliation with God. The majesty of God was too grossly offended for any lesser satisfaction. The glory of God demanded Israel's ruin, but the same cause required its restoration also. For Israel's fall disgraced YHWH among the nations; to save his reputation he must therefore restore Israel to its land and make it prosper as never before. The dried bones of Israel must revive, that they and all the nations should know that he was YHWH (Ezek. 37). Ezekiel, too, foresaw the remaking of human nature, but as a necessity of God's glorification; the concatenation of Israel's sin, exile, and consequent defamation of God's name must never be repeated. In 586 BCE the doom prophecies of Jeremiah and Ezekiel came true. Rebellious Jerusalem was reduced by Nebuchadrezzar, the Temple was burnt, and much of Judah's population dispersed or deported to Babylonia.

THE EXILIC PERIOD

Postexilic religious forms

The survival of the religious community of exiles in Babylonia demonstrates how rooted and widespread the religion of YHWH was. Abandonment of the national religion as an outcome of the disaster is recorded of a minority only. There were some cries of despair, but the persistence of prophecy among the exiles shows that their religious vitality had not flagged. The Babylonian Jewish community, in which the cream of Judah lived, had no sanctuary or altar (in contrast to the Jewish garrison of Elephantine in Egypt); what developed in their place can be surmised from new postexilic religious forms: fixed prayer; public fasts and confessions; and assembly for the study of the Torah, which may have developed from visits to the prophets for oracular edification. The absence of a local or territorial focus must also have spurred the formation of a literary–ideational centre of communal life—the sacred canon of Covenant documents that came to be the core of the present Pentateuch. Observance of the Sabbath—a peculiarly public feature of communal life—achieved a significance among the exiles virtually equivalent to all the rest of the Covenant rules together. Notwithstanding its political impotence, the spirit of the exiles was so high that foreigners were attracted to their ranks, hopeful of sharing their future glory.

Assurance of that future glory was given not only in the consolations promised by Jeremiah and Ezekiel (the fulfillment of whose prophecies of doom lent credit to their consolations); the great comforter of the exile was the writer or writers of what is known as Deutero-Isaiah (Isa. 40–66), who perceived in the rise and progress (from *c.* 550) of the Persian king Cyrus II the Great the instrument of God's salvation. Going beyond the national hopes of

Ezekiel, animated by the universal spirit of the pre-exilic Isaiah, Deutero-Isaiah saw in the miraculous restoration of Israel a means of converting the whole world to faith in Israel's God. Israel would thus serve as "a light for the nations, that YHWH's salvation may reach to the end of the earth." In his conception of the vicarious suffering of God's servant—through which atonement is made for the ignorant heathen—Deutero-Isaiah found a handle by which to grasp the enigma of faithful Israel's lowly state among the Gentiles. The idea was destined to play a decisive role in the self-understanding of the Jewish martyrs of the Syrian king Antiochus IV Epiphanes' persecution in the 2nd century BCE (in, for example, Daniel) and later again in the Christian appreciation of the death of Jesus.

THE PERIOD OF THE RESTORATION

After conquering Babylon, Cyrus so far justified the hopes put in him that he allowed those Jews who wished to do so to return and rebuild their Temple. Though, in time, some 40,000 made their way back, they were soon disillusioned by the failure of the glories of the restoration to materialize and by the controversy with the Samaritans, and left off building the Temple. (The Samaritans were a judaized mixture of native north Israelites and Gentile deportees settled by the Assyrians in the erstwhile northern kingdom.) A new religious inspiration came under the governorship of Zerubbabel, a member of the Davidic line, who became the centre of messianic expectations during the anarchy attendant upon the accession to the Persian throne of Darius I (522). The prophets Haggai and Zechariah perceived the disturbances as heralds of an imminent overthrow of the heathen Persian Empire and a worldwide manifestation of God and glorification of Zerubbabel. Against that day they urged the people quickly to complete the building of the Temple. The labour was resumed and completed in 516; but the prophecies remained unfulfilled. Zerubbabel disappears from the biblical narrative, and the spirit of the community flagged again.

Role of repentance and the Torah

The one religious constant in the vicissitudes of the restored community was the mood of repentance and the desire to win back God's favour by adherence to his Covenant rules. The anxiety that underlay this mood produced a hostility to strangers, which encouraged a lasting conflict with the Samaritans, who asked permission to take part in rebuilding the Temple of the God they too worshipped. The Jews, however, rejected them on ill-specified grounds—apparently ethnoreligious; *i.e.,* they felt the Samaritans to be alien to their historical community of faith, especially to its messianic hopes. Nonetheless, intermarriage occurred and precipitated a new crisis when, in 458, the priest Ezra arrived from Babylon, intent on enforcing the regimen of the Torah. By construing ancient and obsolete laws excluding Canaanites and others so as to make them apply to their own times and neighbours, the leaders of the Jews brought about the divorce and expulsion of several dozen non-Jewish wives and their children. Tension between the xenophobic (fear of strangers) and xenophilic (love of strangers) in postexilic Judaism was finally resolved some two centuries later with the development of a formality of religious conversion, whereby Gentiles who so wished could be taken into the Jewish community by a single, simple procedure.

The decisive constitutional event of the new community was the covenant subscribed to by its leaders in 444, making the Torah the law of the land: a charter granted by the Persian king Artaxerxes I to Ezra—scholar and priest of the Babylonian Exile—empowered him to enforce the Torah as the imperial law for the Jews of the province Avar-nahra (Beyond the River), in which the district of Judah (now reduced to a small area) was located. The charter required the publication of the Torah and the publication, in turn, entailed its final editing—now plausibly ascribed to Ezra and his circle. Survival in the Torah of patent inconsistencies and disaccords with the postexilic situation indicate that its materials were by then sacrosanct, to be compiled but no longer created. But these survivals made necessary the immediate invention of a harmonizing and creative method of text interpretation to adjust the Torah to the needs of the times. The Levites were trained in the art of interpreting the text to the people; the first product of the creative exegesis later known as Midrash is to be found in the covenant document of Nehemiah, chapter 9—every item of which shows development, not reproduction, of a ruling of the Torah. Thus, with the publication of the Torah as the law of the Jews the basis of the vast edifice of the Oral Law characteristic of Judaism was laid.

Origins of the Oral Law

Concern over observance of the Torah was fed by the gap between messianic expectations and the gray reality of the restoration. The gap signified God's continued displeasure, and the only way to regain his favour was to do his will. Thus it is that Malachi, the last of the prophets, concludes with an admonition to be mindful of the Torah of Moses. God's displeasure, however, had always been signalized by a break in communication with him. As time passed and messianic hopes remained unfulfilled, the sense of a permanent suspension of normal relations with God took hold, and prophecy died out. God, it was believed, would some day be reconciled with his people, and a glorious revival of prophecy would then occur. For the present, however, religious vitality expressed itself in dedication to the development of institutions that would make the Torah effective in life. The course of this development is hidden from view by the dearth of sources from the Persian period. But the community that emerged into the light of history in Hellenistic times is one made over radically by this momentous, quiet process. (Mo.Gr.)

Hellenistic Judaism (4th century BCE–2nd century CE)

THE GREEK PERIOD (332–63 BCE)

Hellenism and Judaism. Actual contact between Greeks and Semites goes back to Minoan and Mycenaean times and is reflected in certain terms in Homer and in other early Greek authors. It is not until the end of the 4th century, however, that Jews are first mentioned by Greek writers, who praise the Jews as brave, self-disciplined, and philosophical.

After being conquered by Alexander the Great (332 BCE), Palestine became part of the Hellenistic kingdom of Ptolemaic Egypt, the policy of which was to permit the Jews considerable cultural and religious freedom.

Hellenizing under Seleucid rule

When in 198 BCE Palestine was conquered by King Antiochus III (247–187 BCE), of the Syrian Seleucid dynasty, the Jews were treated even more liberally, being granted a charter to govern themselves by their own constitution, namely, the Torah. Greek influence, however, was already becoming manifest. Some of the 29 Greek cities of Palestine attained a high level of culture. The mid-3rd century-BCE Zenon papyri—containing the correspondence of a business manager of a high Ptolemaic official—present the picture of a wealthy Jew, Tobiah, who through commercial contact with the Ptolemies acquired a veneer of Hellenism, to judge at least from the pagan and religious expressions in his Greek letters. His son and especially his grandsons became ardent Hellenists. It has been argued that the Hellenic influence was so strong among the Jews of Judaea by the beginning of the 2nd century that if the process had continued without the forcible intervention of the Seleucids in Jewish affairs (see below) Judaean Judaism would have become even more syncretistic than that of Philo, the Hellenistic Jewish philosopher of Alexandria (*c.* 15 BCE–*c.* 40 CE). The apocryphal writer Jesus ben Sirach so bitterly denounced the Hellenizers in Jerusalem (*c.* 180 BCE) that he was forced by the authorities to temper his words.

In the early part of the 2nd century BCE, Hellenizing Jews came into control of the high priesthood itself. Jason as high priest (175–172 BCE) established Jerusalem as a Greek city, Antioch-at-Jerusalem, with Greek educational institutions. His ouster by an even more extreme Hellenizing faction, which established Menelaus (died 162 BCE) as high priest, occasioned a civil war, with the wealthy aristocrats supporting Menelaus and the masses Jason. The Syrian king Antiochus IV Epiphanes, who had initially bestowed exemptions and privileges upon the Jews, intervened upon the request of Menelaus' party. Antiochus' promulgation of decrees against the practice of Judaism and the offensive

The Maccabean revolt

and cruel measures to enforce them led to the revolt of an old priest, Mattathias, and his five sons—the so-called Maccabees or Hasmoneans. It has been conjectured that one of the Dead Sea Scrolls, the *War of the Sons of Light Against the Sons of Darkness,* mirrors the fierceness of this struggle. In any case, the figure of the martyr, as known in Judaism and Christianity—the person who bears witness to the faith through his suffering and death—dates from this event.

The tactics employed both in the countryside and in Jerusalem by the Hasmoneans in their counterattack against Hellenizing Jews, whose children they forcibly circumcised, indicate the inroads that Hellenism had already made. On the whole, however, the chief strength of the Hellenizers lay among the wealthy urban population, while the Maccabees derived their strength from the peasants and urban masses. Yet, there is evidence that the ruthlessness exhibited by the Hasmoneans toward the Greek cities of Palestine had political rather than cultural origins, and that, in fact, they were fighting for personal power no less than for the Torah. In any case, some of those who fought on the side of the Maccabees were idol-worshipping Jews. The Maccabees soon found a modus vivendi with Hellenism: Jonathan (160–142), according to the Jewish historian Josephus (*c.* 38–*c.* 100 CE), negotiated a treaty of friendship with Sparta; Aristobulus (104–103 BCE) actually called himself Philhellene (a lover of Hellenism); Alexander Jannaeus (103–76) hired Greek mercenaries and inscribed his coins with Greek as well as with Hebrew. The Greek influence reached its height under King Herod I of Judaea (37–4 BCE), who built a Greek theatre, amphitheatre, and hippodrome in or near Jerusalem.

Social, political, and religious divisions. During the Hellenistic period the priests were both the wealthiest class and the strongest political group among the Jews of Jerusalem. The wealthiest of all were the Oniad family, who held the hereditary office of high priest until they were replaced by the Hasmoneans; the Temple that they supervised was, in effect, a bank, where the Temple wealth was kept and where private individuals also deposited their money. Hence, from a social and economic point of view, Josephus is justified in calling the government of Judaea a theocracy (rule by those having religious authority). Opposition to the priests' oppression arose among an urban middle class group known as scribes (*soferim*), who were interpreters and instructors of the Torah on the basis of an oral tradition probably going back to the time of the return from the Babylonian Exile (538 BCE and after). A special group of the scribes known as Ḥasidim (Greek, Hasideans), or "Pietists," became the forerunners of the Pharisees (middle-class liberal Jews who reinterpreted the Torah and the prophetic writings to meet the needs of their times) and joined the Hasmoneans in the struggle against the Hellenists, though on religious rather than on political grounds.

Josephus held that the Pharisees and the other Jewish parties were philosophical schools, and some modern scholars have argued that the groupings were primarily along economic and social lines; but the chief distinctions among them were religious and go back well before the Maccabean revolt. The equation of Pharisaic with "normative" Judaism can no longer be supported, at any rate not before the destruction of the Temple in 70 CE. The fact that in 70 CE, according to the Palestinian Talmud (see below *Rabbinic Judaism*), there were 24 types of "heretics" in Palestine indicates that there was, in fact, much divergence among Jews; and this picture is confirmed by Josephus, who notes numerous instances of religious leaders who claimed to be prophets and who obtained considerable followings.

The Pharisees and the Sadducees

Some other modern scholars have sought to interpret the Pharisees' opposition to the Sadducees—wealthy, conservative Jews who accepted the Torah alone as authoritative—as based on an urban–rural dichotomy; but a very large share of Pharisaic concern was with agricultural matters. To associate the rabbis with urbanization seems a distortion. The chief support for the Pharisees came from the lower classes, whether in the country or in the city.

The chief doctrine of the Pharisees (literally "Separatists") was that the Oral Law had been revealed to Moses at the same time as the Written Law. In their exegesis and interpretation of this oral tradition, particularly under the rabbi Hillel at the end of the 1st century BCE, the Pharisees were liberal, and their regard for the public won them considerable support. That the Maccabean ruler John Hyrcanus I broke with them and that Josephus set their number at merely "more than 6,000" at the time of King Herod indicates that they were less numerous and influential than Josephus would have his readers believe. The Pharisees stressed the importance of performing all the commandments, including those that appeared to be of only minor significance; those who were particularly strict in their observance of the Levitical rules were known as *ḥaverim* ("companions"). They believed in the providential guidance of the universe, in angels, in reward and punishment in the world to come, and in resurrection of the dead, in all of which beliefs they were opposed by the Sadducees. In finding a modus vivendi with Hellenism, at least in form and in terminology, however, the Pharisees did not differ greatly from the Sadducees. Indeed, the supreme council of the Great Synagogue (or Great Assembly) of the Pharisees was modelled in its organization on Hellenistic religious and social associations. Because they did not take an active role in fostering the rebellion against Rome in 66–70 CE, they were able, through their leader Johanan ben Zakkai, to obtain Roman permission to establish an academy at Jabneh (Jamnia), where, in effect, they replaced the cult of the Temple with study and prayer.

The Sadducees and their subsidiary group, the Boethusians (Boethosaeans), who were identified with the great landowners and priestly families, were more deeply influenced by Hellenization. The rise of the Pharisees may thus be seen, in a sense, as a reaction against the more profound Hellenization favoured by the Sadducees, who were allied with the philhellenic Hasmoneans. From the time of John Hyrcanus (135–104 BCE) the Sadducees generally held a higher position in comparison with the Pharisees and were in favour with the Jewish rulers. Religiously more conservative than the Pharisees, they rejected the idea of a revealed oral interpretation of the Torah, though, to be sure, they had their own tradition, the *sefer gezerot* ("book of decrees" or "decisions"). They similarly rejected the inspiration of the prophetic books of the Bible, as well as the Pharisaic beliefs in angels, rewards, and punishments in the world to come, providential governance of human events, and resurrection of the dead. For them Judaism centred on the Temple; but about 10 years before the destruction of the Temple in 70 CE, the Sadducees in effect disappeared from Jewish life when the Pharisees excluded them from entering the Temple.

The "people of the land" and converts

Not constituting any particular party were the unlearned rural masses known as *ʿamme ha-aretz* ("people of the land"), who were to be found among both the Pharisees and Sadducees and even among the Samaritans, descendants of the northern Israelites who had their own Torah and their own sanctuary. The *ʿamme ha-aretz* did not give the prescribed tithes, did not observe the laws of purity, and were neglectful of the laws of prayer; and so great was the antagonism between them and the learned Pharisees that to their daughters was applied the biblical verse, "Cursed be he who lies with any kind of beast." The antipathy was reciprocated, for in the same passage in the Babylonian Talmud (*Pesaḥim*) are added the words, "Greater is the hatred wherewith the *ʿamme ha-aretz* hate the scholar than the hatred wherewith the heathens hate Israel." That there was, however, social mobility is clear from the Talmudic dictum, "Heed the sons of the *ʿam ha-aretz,* for they will be the living source of the Torah." That there is little evidence that the early Christian church was particularly successful in converting *ʿamme ha-aretz* suggests that their position was not unbearable.

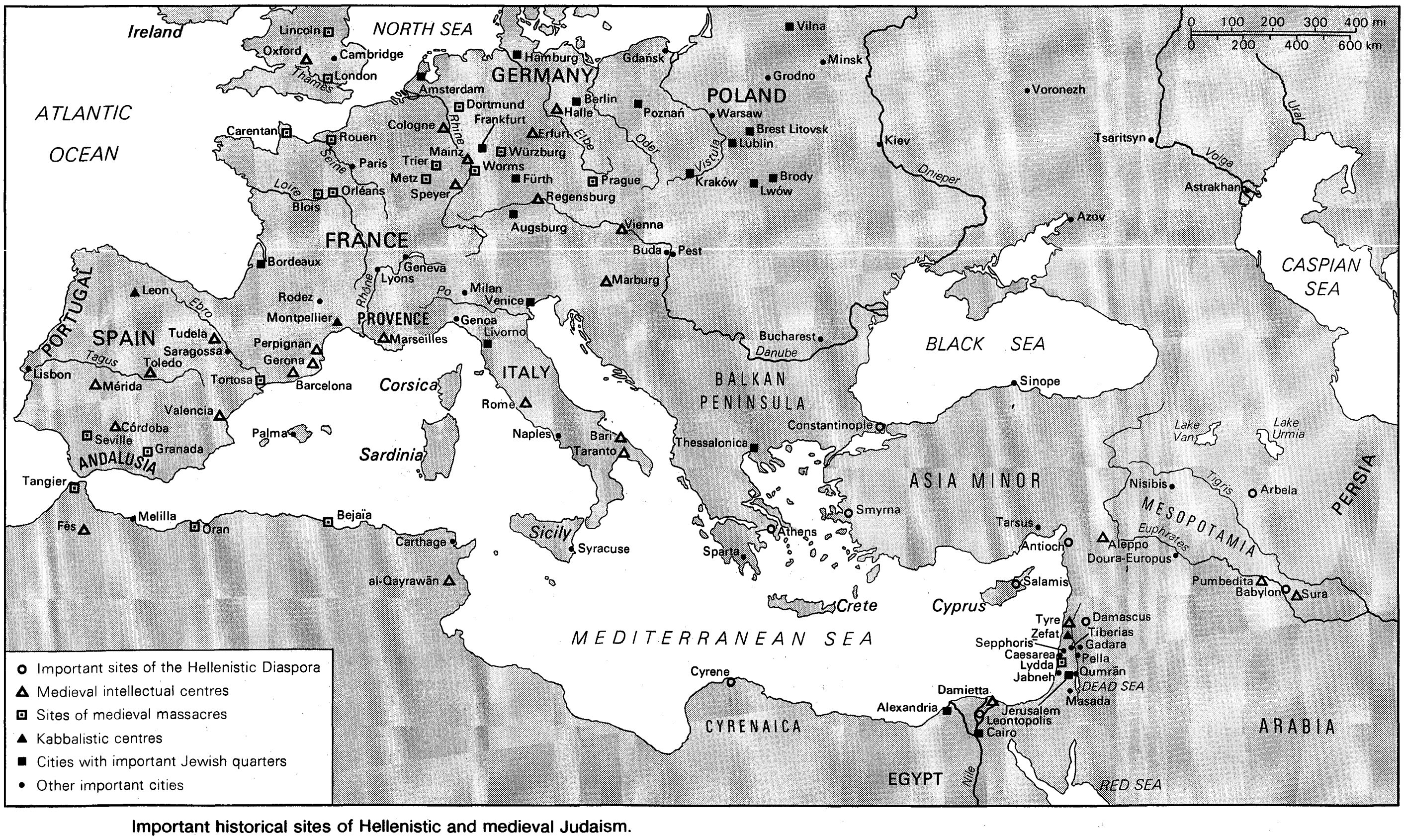
Proselytes (converts) to Judaism, though not constituting a class, became increasingly numerous both in Palestine and especially in the Diaspora (the Jews living beyond Palestine). Scholarly estimates of the Jewish population of this era range from 700,000 to 5,000,000 in Palestine and from 2,000,000 to 5,000,000 in the Diaspora, with the prevailing opinion being that about one-tenth of the population of the Mediterranean world at the beginning of the Christian Era was Jewish. Such numbers represent a considerable increase from previous eras and must have included large numbers of proselytes. Already in 139 BCE the Jews of Rome were charged by the praetor (civil administrator) with attempting to contaminate Roman morals with their religion, presumably an allusion to proselytism. The first large-scale conversions were by John Hyrcanus and Aristobulus, who, in 130 and 103 BCE, respectively, forced the people of Idumaea in southern Palestine and of Ituraea in northern Palestine to become Jews. The eagerness of the Pharisees to win converts is seen in a statement in Matthew that the Pharisees would "traverse sea and land to make a single proselyte." To be sure, some of the proselytes, according to Josephus, did return to their pagan ways, but the majority apparently remained true to their new religion. In addition, there were many "sympathizers" with Judaism who observed one or more Jewish practices without being fully converted.

Outside the pale of Judaism in most, though not all, respects were the Samaritans, who, like the Sadducees, refused to recognize the validity of the Oral Law; and, in fact, the break between the Sadducees and the Samaritans did not occur until the conquest of Shechem by John Hyrcanus (128 BCE). Like the later so-called Qumrān covenanters (the monastic group with whom are associated the Dead Sea Scrolls), they were opposed to the Jewish priesthood and the cult of the Temple, regarded Moses as a messianic figure, and forbade the revelation of esoteric doctrines to outsiders.

Scholars have recently revised an older conception of a "normative" Pharisaic Judaism dominant in Palestine and a deviant Judaism dominant in the Diaspora. On the one hand, the picture of "normative" Judaism is broader than at first believed, and it is clear that there were many differences of emphasis within the Pharisaic party; and, on the other hand, supposed differences between Alexandrian and Palestinian Judaism were not as great as had been formerly thought. In Palestine, no less than in the Diaspora, there were then deviations from Pharisaic standards.

Hellenistic influences on Pharisaic Judaism

Despite the attempts of the Pharisaic leaders to restrain the wave of Greek influence, they themselves showed at least a surface Hellenization. In the first place, as many as 2,500–3,000 words of Greek origin are to be found in the Talmudic corpus, and they supply important terms in the fields of law, government, science, religion, technology, and everyday life, especially in the popular sermons preached by the rabbis. When preaching, the Talmudic rabbis often gave the Greek translation of biblical verses for the benefit of those who understood Greek only. The prevalence of Greek in ossuary (burial) inscriptions and the discovery of Greek papyri in the Dead Sea caves confirm the widespread use of the language, though few Jews, it seems, really mastered Greek. Again, there was a surface Hellenization in the frequent adoption of Greek names, even by the rabbis; and there is evidence (Talmud, *Soṭa*) of a school at the beginning of the 2nd century that had 500 students of "Greek wisdom." Even after 117 CE, when it was prohibited by the rabbis to teach one's son Greek, Rabbi Judah the Prince, the editor of the Mishna (authoritative compilation of the Oral Law) at the end of the 2nd century, remarked, "Why talk Syriac in Palestine? Talk either Hebrew or Greek." Even the synagogues of the period have the form of Hellenistic-Roman basilicas, have frequent inscriptions in Greek, and often have pagan motifs. Many of the anecdotes told about the rabbis have Socratic and Cynic parallels. There is evidence of discussions of rabbis with Athenians, Alexandrians, and Roman philosophers, and even with the emperor Antoninus; but in all of these discussions there is evidence of only one rabbi, Elisha ben Abuyah, who became a Gnostic heretic, accepting certain esoteric religious dualistic views. The rabbis never mention the Greek philosophers Plato or Aristotle or the Hellenistic Jewish philosopher Philo, and they never use any Greek philosophical terms; the only Greek author whom they name is Homer. Again, the parallels between Hellenistic rhetoric and rabbinic hermeneutics are in the realm of terminology rather than of substance, and those between Roman and Talmudic law are inconclusive. Part of the explanation of this may be that, although there were 29 Greek cities in Palestine, none was in Judaea, the real stronghold of the Jews.



Important historical sites of Hellenistic and medieval Judaism.

Religious rites and customs in Palestine: Temple and synagogues. The most important religious institution of the Jews until its destruction in 70 was the Temple in Jerusalem—the Second Temple, erected 538–516 BCE. Though services were interrupted for three years by Antiochus Epiphanes (167–165 BCE) and though the Roman general Pompey desecrated the Temple (63 BCE), Herod lavished great expense in rebuilding it. The high priesthood itself became degraded by the extreme Hellenism of such high priests as Jason and Menelaus; and the institution declined when Herod began the custom of appointing the high priests for political and financial considerations. That not only the multitude of Jews but the priesthood itself suffered from sharp divisions is clear from the bitter class warfare that ultimately erupted in 59 CE between the high priests on the one hand and the ordinary priests and the leaders of the populace of Jerusalem on the other.

Though the Temple remained central in Jewish worship, synagogues may already have emerged during the Babylonian Exile in the 6th century BCE. In any case, in the following century, Ezra stood upon a pulpit of wood and read from the Torah to the people (Nehemiah). According to the interpretation of some scholars, a synagogue existed even within the precincts of the Temple; and certainly by the time of Jesus, to judge from the references to Galilean synagogues in the New Testament, synagogues were common in Palestine. Hence, when the Temple was destroyed in 70, the spiritual vacuum was hardly as great as it had been after the destruction of the First Temple (586 BCE).

The Sanhedrin

The chief legislative, judicial, and educational body of the Palestinian Jews during the period of the Second Temple was the Great Sanhedrin (council court), consisting of 71 members, among whom the Sadducees were an important party. The members shared the government with the king during the early years of the Hasmonean dynasty, but beginning with Herod's reign their authority was restricted to religious matters. In addition, there was another Sanhedrin, set up by the high priest, which served as a court of political council, as well as a kind of grand jury.

Religious and cultural life in the Diaspora. During the Hellenistic–Roman period the chief centres of Jewish population outside Palestine were in Syria, Asia Minor, Babylonia, and Egypt, each of which is estimated to have had at least 1,000,000 Jews. The large Jewish community of Antioch—which, according to Josephus, had been given all the rights of citizenship by the Seleucid founder-king, Seleucus Nicator (died 280 BCE)—attracted a particularly large number of converts to Judaism. It was in Antioch that the apocryphal book of Tobit was probably composed in the 2nd century BCE to encourage wayward Diaspora Jews to return to their Judaism. As for the Jews of Asia Minor, whose large numbers were mentioned by Cicero (1st century BCE), their not joining in the Jewish revolts against the Roman emperors Nero, Trajan, and Hadrian would indicate that they had sunk deep roots into their environment. In Babylonia, in the early part of the 1st century CE, two Jewish brothers, Asinaeus and Anilaeus, were able to establish an independent minor state; their followers were so meticulous in observing the Sabbath that they assumed that it would not be possible to violate the Sabbath even in order to save themselves from a Parthian attack. In the early part of the 1st century CE, according to Josephus, the royal house and many of their entourage in the district of Adiabene in northern Mesopotamia were converted to Judaism; some of the Adiabenian Jews distinguished themselves in the revolt against Rome in 66 (see below).

The Egyptian Diaspora

The largest and most important Jewish settlement in the Diaspora was in Egypt. There is evidence (papyri) of a Jewish military colony at Elephantine (Yeb), Upper Egypt, as early as the 6th century BCE. These papyri reveal the existence of a Jewish temple—which most certainly would be considered heterodox—and some syncretism (mixture) with pagan cults. Alexandria, the most populous and most influential Hellenistic Jewish community in the Diaspora, had its origin when Alexander the Great assigned a quarter of the city to the Jews. Until about the 3rd century BCE the papyri of the Egyptian Jewish community were written in Aramaic; after that, with the exception of the Nash papyrus in Hebrew, all papyri until 400 CE were in Greek. Similarly, of the 116 Jewish inscriptions from Egypt, all but five are written in Greek. The process of Hellenistic acculturation is, thus, obvious.

The most important work of the early Hellenistic period, dating, according to tradition, from the 3rd century BCE, is the Septuagint, a translation of the Pentateuch into Greek. (The translation of the whole Hebrew Bible was completed during the next two centuries.) The fact that, in the *Letter of Aristeas* (see below) and the works of Philo and Josephus, this translation was itself regarded as divinely inspired led to the neglect of the Hebrew original. The translation shows some knowledge of Palestinian exegesis and the tradition of Halakha (the Oral Law); but the rabbis themselves, noting that the translation diverged from the Hebrew text, apparently had ambivalent feelings about it, as is evidenced in their alternate praise and condemnation of it. The fact that such a concept as Torah was translated as *nomos* ("law") and *tzedaqa* as *dikaiosynē* ("justice") opened the way to antilegalism in early Christianity and to Platonic interpretations; and the introduction of such Greek mythological terms as "Titans" and "Sirens" helped to pave the way for the syncretism of Judaism and paganism.

The establishment of a temple at Leontopolis in Egypt (*c.* 145 BCE) by a deposed high priest, Onias IV, indicates that the temple was clearly heterodox; but this temple never really offered a challenge to the one in Jerusalem and was merely the temple of the military colony of Leontopolis. It is significant that the Palestinian rabbis ruled that a sacrifice intended for the temple of Onias might be offered in Jerusalem. That the temple of Onias made little impact upon Egyptian Jewry can be seen from the silence about it on the part of Philo, who often mentions the Temple in Jerusalem. The temple of Onias, however, continued until it was closed by the Roman emperor Vespasian in 73 CE.

The chief religious institutions of the Egyptian Diaspora were synagogues. As early as the 3rd century BCE there were inscriptions mentioning two *proseuchai,* Jewish prayerhouses. In Alexandria there were numerous synagogues throughout the city, of which the largest was so famous that it is said in the Talmud (see below *Rabbinic Judaism*) that he who has not seen it has never seen the glory of Israel.

Egyptian Jewish literature. In Egypt the Jews produced a considerable literature (most of it now lost), intended to inculcate in Greek-speaking Jews a pride in their past and to counteract an inferiority complex that some of them felt about Jewish cultural achievements. In the field of history, Demetrius, near the end of the 3rd century BCE, wrote a work *On the Kings in Judaea*—perhaps intended to refute an anti-Semitic Egyptian priest and author—showing considerable concern for chronology. In the 2nd century BCE a Jew who used the name of Hecataeus wrote *On the Jews.* Another, Eupolemus (*c.* 150 BCE), like Demetrius, wrote *On the Kings in Judaea;* an indication of its apologetic nature may be seen from the fragment asserting that Moses taught the alphabet not only to the Jews but also to the Phoenicians and to the Greeks. Artapanus (*c.* 100 BCE), in his book *On the Jews,* went even further in romanticizing Moses by identifying him with the Greek Musaeus and the Egyptian Hermes-Thoth (god of Egyptian writing and culture) and by asserting that Moses was the real originator of Egyptian civilization and that he even taught the Egyptians the worship of the deity Apis (the sacred bull) and the ibis (sacred bird). In his history, Cleodemus (or Malchus), in an obvious attempt to win for the Jews the regard of the Greeks, asserted that two sons of Abraham had joined Heracles in his expedition in Africa and that the Greek hero had married the daughter of one of them. On the other hand, Jason of Cyrene (*c.* 100 BCE) wrote a history, of which II Maccabees is a summary, glorifying the Temple and violently attacking the Jewish Hellenizers; but his manner of writing history is typically Hellenistic, with emphasis on pathos. III Maccabees (1st century BCE) is a work of propaganda intended to counteract those Jews who sought to win citizenship in Alexandria. The *Letter of Aristeas,* though ascribed to a pagan courtier, Ptolemy II Philadelphus, was probably composed by an Alexandrian

Jew about 100 BCE to defend Judaism and its practices against detractors.

Egyptian Jews also composed poems and plays, now extant only in fragments, to glorify their history. Philo the Elder (*c.* 100 BCE) wrote an epic *On Jerusalem* in Homeric hexameters. Theodotus (*c.* 100 BCE) wrote an epic *On Shechem,* quite clearly apologetic, to judge from the fragment connecting the name of Shechem with Sikimios, the son of the Greek god Hermes. At about the same time, a Jewish poet wrote a didactic poem, ascribing it to the pagan Phocylides, though closely following the Bible in some details; the author disguised his Jewish origin by omitting any attack against idolatry from his moralizing. A collection known as *The Sibylline Oracles,* containing Jewish and Christian prophecies in pagan disguise, includes some material composed by a 2nd-century-BCE Alexandrian Jew who intended to glorify the pious Jews and perhaps to win converts; it is possible that the *Oracles* were known to the Roman poet Virgil when he wrote his fourth *Eclogue.*

A Jewish dramatist of the period, Ezekiel (*c.* 100 BCE), composed tragedies in Greek. Fragments of one of them, *The Exodus,* show how deeply he was influenced by the Greek dramatist Euripides. Whether such plays were actually presented on the stage or not, they edified Jews and showed the pagans that the Jews had as much material for drama as they did.

Alexandrian Jewish philosophical achievements

The greatest achievement of Alexandrian Judaism was in the realm of wisdom literature and philosophy. In a work on the analogical interpretation of the Law of Moses, Aristobulus in the 2nd century BCE anticipated Philo in attempting to harmonize Greek philosophy and the Torah, in using the method of allegory to explain anthropomorphisms in the Bible, and in asserting that the Greek philosophers were indebted to Moses. The Wisdom of Solomon, dating from the 1st century BCE, shows an acquaintance with the Platonic doctrine of the preexistence of the soul and with a method of argument known as *sorites* that was favoured by the Stoics (Greek philosophers). During the same period the author of IV Maccabees showed an intimate knowledge of Greek philosophy, particularly of Stoicism.

By far the greatest figure in Alexandrian Jewish literature is Philo, who has come to be recognized as a major philosopher. His synthesis of Greek philosophy, particularly that of Plato, and of the Torah, and his formulation of the Logos (Word, or Divine Reason) as an intermediary between God and the world, helped lay the groundwork for Neoplatonism (a philosophy dealing with levels of being), Gnosticism (a dualistic religious movement teaching that matter is evil and that spirit is good), and the philosophical framework of the early Church Fathers. Philo was a devotee of Judaism neither as a mystic cult nor as a collateral branch of Pharisaic Judaism; he was a Diaspora Jew with a profound knowledge of Greek literature who, though almost totally ignorant of Hebrew, tried to find a modus vivendi between Judaism and secular culture.

Mention may be made of the Jewish community of Rome. Numbering perhaps 50,000, it was, to judge from the inscriptions in the Jewish catacombs, predominantly Greek-speaking and almost totally ignorant of Hebrew. References in Roman writers, particularly Tacitus and the satirists, have led scholars to conclude that the community—which was influential, to judge from the pagan jibes—observed the Sabbath and the dietary laws and was active in seeking converts.

Modifications of Jewish traditions by Diaspora Jews

The Hellenization of the Diaspora Jews is, however, to be seen not merely in their literature but even more in the papyri and art objects that have recently been studied at great length. As early as 290 BCE, Hecataeus of Abdera, a Greek non-Jew living in Egypt, had remarked that under the Persians and Macedonians the Jews had greatly modified the traditions of their fathers. The fact that—to judge from other papyri—at least three-fourths of the Egyptian Jews had personal names of Greek, rather than Hebrew, origin is significant. That the only schools of which mention is made are Sabbath schools intended for adults and that, on the contrary, Jews were extremely eager to gain admittance for their children to Greek *gymnasia*—where quite obviously they would have to make compromises with their Judaism—indicates their scale of values. Again, there are a number of violations from the norms of Halakha (which precluded the charging of interest for a loan), most notably in the fact that of 11 known extant loan documents only two are without interest. There are often striking similarities between the documents of sale, marriage, and divorce of the Jews and of the Greeks in Egypt, though some of this, as with the documents of the Elephantine Jewish community, may be due to a common origin in the cuneiform law of ancient Mesopotamia. The charms and apotropaic (designed to avert evil) amulets are often syncretistic, and the Jews can hardly have been unaware of the religious significance of symbols that were still very much filled with meaning in pagan cults. The fact that the Jewish community of Alexandria was preoccupied in the 1st century BCE and the 1st century CE with obtaining rights as citizens—which certainly involved compromises with Judaism, including participation in pagan festivals and sacrifices—shows how far they were ready to deviate. Philo mentions Jews who scoffed at the Bible, which they insisted on interpreting literally, and of others who failed to adhere to the biblical laws that they regarded as mere allegory; he writes too of Jews who observed nothing of Judaism except the holiday of Yom Kippur. But despite such deviations, the pagan writers constantly accuse the Diaspora Jews of being "haters of mankind" and of being absurdly superstitious; and Christian writers later similarly attack the Jews for refusing to give up the Torah. At least they were loyal Jews in their contributions of the Temple tax and in pilgrimages to Jerusalem on the three festivals. Actual apostasy and intermarriage were apparently not common, but the virulent anti-Semitism and the pogroms perpetrated by the Egyptian non-Jews must have served as a deterrent.

Palestinian literature. During this period literature was composed in Palestine in Hebrew, Aramaic, and Greek, with the exact language still a subject of dispute among scholars in many cases and with the works often apparently composed by more than one author over a considerable period of time. Most of the works composed in Hebrew, many of them existing only in Greek—Ecclesiasticus, I Maccabees, Judith, *Testaments of the Twelve Patriarchs,* Baruch, *Psalms of Solomon,* Prayer of Manasseh—and many of the Dead Sea Scrolls are generally conscious imitations of biblical books, often reflecting the dramatic events of the Maccabean struggle and often with an apocalyptic tinge (involving the dramatic intervention of God in history). The literature in Aramaic consists of the following: (1) biblical or Bible-like legends or midrashic (interpretive) additions—*Testament of Job, The Martyrdom of Isaiah, Paralipomena of Jeremiah, Life of Adam and Eve,* the Dead Sea *Genesis Apocryphon,* Tobit, Susanna, Bel and the Dragon; and (2) apocalypses—*Enoch* (perhaps originally written in Hebrew), *Assumption of Moses,* the Syriac Baruch, II (IV) Esdras, and *Apocalypse of Abraham.* In Greek the chief works by Palestinians are histories of the Jewish War against Rome and of the Jewish kings by Justus of Tiberias (both are lost) and the history of the Jewish War, originally in Aramaic, and the *Jewish Antiquities* by Josephus (both written in Rome).

Apocalyptic and wisdom literature

Of the wisdom literature composed in Hebrew, the book of the Wisdom of Jesus the Son of Sirach, or Ecclesiasticus (*c.* 180–175 BCE), modelled on the book of Proverbs, identified Wisdom with the observance of the Torah. The *Testaments of the Twelve Patriarchs,* probably written in the latter half of the 2nd century BCE, patterned on Jacob's blessings to his sons, are now thought to belong to eschatological literature related to the Dead Sea Scrolls. The identification of Wisdom and Torah is stressed in the Mishnaic tract *Pirqe Avot* ("Sayings of the Fathers"), which, though edited 200 CE, contains the aphorisms of rabbis dating back to 300 BCE.

Books such as the *Testament of Job,* the Dead Sea Scroll *Genesis Apocryphon,* the *Book of Jubilees* (now known to have been composed in Hebrew, as seen by its appearance among the Dead Sea Scrolls), and *Biblical Antiquities,* falsely attributed to Philo (originally written in Hebrew, then translated into Greek, but now extant only in Latin), as well as the first half of Josephus' *Jewish Antiquities,* often show affinities with rabbinic Midrashim (interpretive works) in their legendary accretions of biblical details. Sometimes, as in *Jubilees* and in the Pseudo-Philo work, these accretions are intended to answer the questions of heretics, but often, particularly in the case of Josephus, they are apologetic in presenting biblical heroes in a guise that would appeal to a Hellenized audience.

Apocalyptic trends, given considerable impetus by the victory of the Maccabees over the Syrian Greeks, were not—as was formerly thought—restricted to Pharisaic circles. They were (as is clear from the Dead Sea Scrolls) found in other groups as well, and are of particular importance for their influence on both Jewish mysticism and early Christianity. These books, which have a close connection with the biblical Book of Daniel, stress the impossibility of a rational solution to the problem of theodicy—how to reconcile the righteousness of God with observable evil. They also stress the imminence of the day of salvation, which is to be preceded by terrible hardships, and presumably reflected the current historical setting. In the book of *Enoch* there is stress on the terrible punishment inflicted upon sinners in the Last Judgment, the imminent coming of the Messiah and of his kingdom, and the role of angels.

The sole Palestinian Jewish author writing in Greek whose works are preserved is Josephus. His account of the war against the Romans in his *Life* and, to a lesser degree, in the *Jewish War* are largely a defense of his own questionable behaviour as the commander of the Jewish forces in Galilee. But these works and more especially *Against Apion* and the *Jewish Antiquities* are largely defenses of Judaism against anti-Semitic attacks. Josephus' *Jewish War* is often quite deliberately parallel to Thucydides' *History of the Peloponnesian War;* and his *Jewish Antiquities* is quite deliberately parallel to Dionysius of Halicarnassus' *Roman Antiquities,* dating from earlier in the same century.

THE ROMAN PERIOD (63 BCE–135 CE)

New parties and sects. Under Roman rule a number of new groups, largely political, emerged in Palestine. Their common aim was to seek an independent Jewish state. All were zealous for, and strict in their observance of, the Torah.

The Herodians and the Zealots

The Herodians were a political group that after the death of Herod—whom they apparently regarded as the Messiah—sought the reestablishment of the rule of Herod's descendants over an independent Palestine as a prerequisite for Jewish preservation. Unlike the Zealots, however (see below), they did not refuse to pay taxes to the Romans.

The Zealots' party, founded *c.* 6–9 CE, refused to pay tribute to the Romans and advocated overthrowing them on the ground that they should acknowledge God alone as their master. A priestly, eschatologically oriented resistance movement, the Zealots were particularly dedicated to keeping the Temple and its cult pure and used guerrilla tactics toward that end. The Sicarii (Assassins), so-called because of the dagger (*sica*) they carried, arose *c.* 54, according to Josephus, as a group of bandits who kidnapped or murdered those who had found a modus vivendi with the Romans. It was they who made a stand at the fortress of Masada near the Dead Sea, committing suicide rather than be captured by the Romans (73).

The Essenes, New Covenanters, and Dead Sea groups

A number of other parties—various types of Essenes, Damascus Covenanters, and the Qumrān Dead Sea groups—were distinguished by their pursuit of an ascetic monastic life, disdain for material goods and sensual gratification, sharing of material possessions, concern for eschatology, strong apocalyptic views in anticipation of the coming of the Messiah, practice of ablutions to attain greater sexual and ritual purity, prayer, contemplation, and study. The Essenes were like the Therapeutae, a Jewish religious group that had flourished in Egypt two centuries earlier, but the latter actively sought "wisdom" whereas the former were anti-intellectual. Only some of the Essenes were celibate. The Essenes have been termed Gnosticizing Pharisees because of their belief, shared with the Gnostics, that the world of matter was evil; some have seen in them the influence of a quasi-monasticism.

The Damascus sect (New Covenanters) were a group of Pharisees who went beyond the letter of the Pharisaic Halakha. Like the Essenes and the Dead Sea sect, they had a monastic type of organization and opposed the way in which sacrifices were offered in the Temple.

The continuing recent discoveries of scrolls in caves of the Dead Sea area have focussed attention on the groups that lived there. On the basis of paleography, carbon-14 testing, and the coins discovered there, most scholars accept a 1st-century date for them. A theoretical relationship of the communities with John the Baptist and the nascent Christian groups remains in dispute, however. The sectaries have been identified variously as Zealots, an unnamed anti-Roman group, and especially Essenes; but a major difference between the Qumrān groups and the Essenes is that the former were militarily activist (the discovery of hymns and a calendar at Masada—a stronghold of the Sicarii—that had previously been found at Qumrān, may indicate a connection between the groups), while the latter were, for the most part, pacifist. That the groups had secret, presumably apocalyptic, teachings is clear from the fact that among the scrolls are some in cryptographic script and reversed writing; and yet, despite their extreme piety and legalistic conservatism, they apparently were not unaware of Hellenism, to judge from the presence of Greek books at Qumrān.

It has long been debated whether the Gnostic systems of the 1st and 2nd centuries go back to the collapse of the apocalyptic strains in Judaism—which expected a final transforming catastrophic event—when the Temple was destroyed in 70. It is doubtful that there is any direct Jewish source for this Gnosticism, though some characteristic Gnostic doctrines are found in certain groups of particularly apocalyptic 1st-century Jews—the dichotomy of body and soul and a disdain for the material world, a notion of esoteric knowledge, and an intense interest in angels and in problems of creation.

Origin of Christianity: the early Christians and the Jewish community. Though it attracted little attention among pagans and Jews at the beginning, the rise of Christianity was by far the most important "sectarian" development of the Roman period. With the revision, largely due to the discoveries at Qumrān, of the view that Pharisaic Judaism was to be considered normative, primitive Christianity, with its apocalyptic and eschatological interests, has come to be viewed by many scholars as no longer "sectarian" or peripheral to Jewish development but, at least initially, as part of a broad spectrum of attitudes within Judaism. Jesus himself, despite his criticisms of Pharisaic legalism, may now be classified as a Pharisee with strong apocalyptic inclinations; he proclaimed that he had no intention of abrogating the Torah, but of fulfilling it. It is possible to envision a direct line between Jewish currents, both in Palestine and the Diaspora in the Hellenistic Age, and Christianity, particularly in the traditions of martyrdom, proselytism, monasticism, mysticism, liturgy, and religious philosophy, especially the doctrine of the Logos (Word) as an intermediary between God and the world and the synthesis of faith and reason. The Septuagint, in particular, played an important role both theoretically, in the transformation of Greek philosophy into the theology of the Church Fathers, and practically, in converting Jews and Jewish "sympathizers" to Christianity. The connection of nascent Christianity with the Qumrān groups may be seen in their dualism and apocalypticism; but there are differences, notably in the conception of the Incarnation, in the relationship of the Son and the Father, and in Jesus' vicarious suffering for sinners as against the direct suffering of the Qumrān Teacher of Righteousness. Again, the Qumrān group constituted an esoteric movement, militant, with enforced community of goods, concerned with strict observance of the Torah, especially with its calendar, whereas Christianity was pacifist, was open to all, and represented a New Covenant, with stress away from the Torah ritual and with voluntary community of possessions. In general, moreover, Christianity was more positively disposed toward Hellenism than was Pharisaism, particularly under the leadership of Paul, a thoroughly Hellenized Jew.

Primitive Christianity a part of the 1st-century Jewish religious spectrum

When Paul proclaimed his antinomianism (against Torah observance as a means of salvation) many Jewish followers of Jesus became Jewish Christians and continued to observe the Torah. Their two main groupings were the Ebionites—probably to be identified with those called *minim,* or "sectaries," in the Talmud—who accepted Jesus as the Messiah but denied his divinity, and the Nazarenes, who regarded Jesus as both Messiah and God, but regarded the Torah as binding upon Jews alone.

The break between Judaism and Christianity

The percentage of Jews converted to any form of Christianity was extremely small, as can be seen from the frequent criticisms of Jews for their stubbornness by Christian writers. In the Diaspora, despite the strong influence of Hellenism, there were relatively few Jewish converts, though the Christian movement had some success in winning Alexandrian Jews.

There were four major stages in the final break between Christianity and Judaism: (1) the flight of the Jewish Christians from Jerusalem to Pella across the Jordan in 70 and their refusal to continue the struggle against the Romans; (2) the institution by the patriarch Gamaliel II of a prayer in the Eighteen Benedictions against such heretics (*c.* 100), and (3 and 4) the failure of the Christians to join the messianic leaders Lukuas-Andreas and Bar Kokhba in the revolts against Trajan (115–117) and Hadrian (132–135), respectively.

Judaism under Roman rule. When Pompey entered the Temple in 63 BCE as an arbiter both in the civil war between Hyrcanus and Aristobulus and in the struggle of the Pharisees against both Jewish rulers, Judaea in effect became a puppet state of the Romans. During the civil war between Pompey and Julius Caesar, the Idumaean Antipater had ingratiated himself with Caesar by aiding him and was rewarded by being made governor of Judaea; the Jews were rewarded through the promulgation of a number of decrees favourable to them, which were reaffirmed by Augustus and later emperors. His son Herod, king of Judaea, an admirer of Greek culture, supported a cult worshipping the Emperor and built temples to Augustus in non-Jewish cities. Since he was by origin an Idumaean, he was regarded by many Jews as a foreigner. (The Idumaeans, or Edomites, were forcibly converted to Judaism by John Hyrcanus; see above.) On several occasions during and after his reign, Pharisaic delegations sought to convince the Romans to end the quasi-independent Jewish government. After the death of Herod's son and successor Archelaus in 6 CE, his realms were ruled by Roman procurators, the most famous or infamous of whom, Pontius Pilate (26–36), attempted to introduce busts of the Roman emperor into Jerusalem and discovered the intense religious zeal of the Jews in opposing this measure. When Caligula ordered the governor of Syria, Petronius, to install a statue of himself in the Temple, a large number of Jews proclaimed they would suffer death rather than to permit such a desecration. Petronius in response succeeded in getting the Emperor to delay. The procurators of Judaea, being of equestrian (knightly) rank and often of Oriental Greek stock, were more anti-Semitic than the governors of Syria, who were of the higher senatorial order. The last procurators in particular were indifferent to Jewish religious sensibilities; and various patriotic groups, to whom nationalism was an integral part of their religion, succeeded in polarizing the Jewish population and bringing on an extremely bloody war with Rome in 66–70. The climax of the war was the destruction of the Temple in 70, though, according to Josephus, the Roman general (and later emperor) Titus sought to spare it. The war was not ended, however, until 73, when the Sicarii at Masada committed suicide rather than submit to the Romans.

Jewish wars against the Romans

The papyri indicate that the war against Trajan (115–117), involving the Jews of Egypt, Cyrenaica, Cyprus, and Mesopotamia (though only to a minor degree those of Palestine), was a widespread revolt under a Cyrenian king-messiah, Lukuas-Andreas, aimed at freeing Palestine from Roman rule. The same spirit of freedom impelled another messiah, Bar Kokhba, who had the support of the greatest rabbi of the time, Akiba, in his spontaneous uprising (132–135). The result was Hadrian's decrees prohibiting circumcision and public instruction in the Torah, though these were soon revoked by Antoninus Pius. Having suffered such tremendous losses on the field of battle, Judaism turned its dynamism to the continued development of the Talmud (see below *Rabbinic Judaism*). (L.H.F.)

Rabbinic Judaism (2nd–18th centuries)

THE AGE OF THE TANNAIM (135–C. 200)

The role of the rabbis. With the defeat of Bar Kokhba and the ensuing collapse of active Jewish resistance to Roman rule (135–136), politically moderate and quietist rabbinic elements remained the only cohesive group within Jewish society. With Jerusalem off limits to the Jews, rabbinic ideology and practice, which were not dependent on Temple, priesthood, or political independence for their vitality, provided a viable program for autonomous community life and thus filled the vacuum created by the suppression of all other Jewish leadership. The Romans, confident that the will for insurrection had been shattered, soon relaxed the Hadrianic prohibitions of Jewish ordination, public assembly, and regulation of the calendar and permitted rabbis who had fled the country to return and reestablish an academy in the town of Usha in Galilee.

The strength of the rabbinate lay in its ability to represent simultaneously the interests of the Jews and the Romans, whose religious and political needs, respectively, now chanced to coincide. The rabbis were regarded favourably by the Romans, as a politically submissive class, which, with its wide influence over the Jewish masses, could translate the Pax Romana (the peace imposed by Roman rule) into Jewish religious precepts. To the Jews, on the other hand, the rabbinic ideology gave the appearance of continuity to Jewish self-rule and freedom from alien interference. The rabbinic program fashioned by Johanan ben Zakkai's circle (see above *Hellenistic Judaism*) had replaced sacrifice and pilgrimage to the Temple with study of Scripture, prayer, and works of piety, thus eliminating the need for a central sanctuary (in Jerusalem) and making of Judaism a religious association capable of fulfillment anywhere. Judaism was now, for all intents and purposes, a Diaspora religion even on its home soil. Any sense of real break with the past was mitigated by continued adherence to purity laws (dietary and bodily) and by assiduous study of Scripture, including those legal sections that historical developments had now made obsolete. The reward held out for scrupulous study and fulfillment was the promise of messianic deliverance; *i.e.,* divine restoration of all those institutions that had become central in Jewish notions of national independence—the Davidic monarchy, Temple service, the ingathering of Diaspora Jewry—and, above all, the assurance of personal reward to the righteous through resurrection and participation in the national rebirth.

The rabbinic way

Apart from the right to teach Scripture publicly, the most pressing need felt by the surviving rabbis was for the reorganization of a recognized body that would reactivate the functions of the former Sanhedrin and pass on disputed questions of law and dogma. A high court was, accordingly, organized under the leadership of Simeon ben Gamaliel (reigned *c.* 135–*c.* 175), the son of the previous patriarch (the Roman term for the head of the Palestinian Jewish community) of the house of Hillel, in association with rabbis representing other schools and interests. In the ensuing struggle for power, the patriarch managed to concentrate all communal authority in his office. The dominating role of the patriarchate reached its zenith in the days of his son and successor, Judah the Prince, whose reign (*c.* 175–*c.* 220) marked the climax of this period of rabbinic activity, otherwise known as the "age of the *tannaim*" (teachers). Armed with wealth, Roman backing, and dynastic legitimacy (which the patriarch now traced to the house of David), Judah sought to standardize Jewish practice through a corpus of legal norms that would reflect recognized views of the rabbinate on every aspect of life. The Mishna (collection of rabbinic law) that soon emerged became the primary source of reference in all rabbinic schools and constituted the core around which the Talmud (commentary on Mishna, literally "teaching") was later compiled. It thus remains the best single introduction to the complex of rabbinic values and practices as they evolved in Roman Palestine.

The Mishna

The making of the Mishna. Although the promulgation of an official corpus represented a break with rabbinic precedent, Judah's Mishna did have antecedents. During the 1st and 2nd centuries CE, rabbinic schools had compiled for their own reference collections in which the results of their exegesis and application of Scripture to problematic situations (Midrash, "investigation" or "interpretation"; plural Midrashim) had been recorded in terse legal form. By 200 CE several such compilations were circulating in Jewish schools and were being utilized by judges. While adhering to the structural form of these earlier collections, Judah compiled a new one in which universally accepted views were recorded alongside those still in dispute, thereby largely reducing the margin for individual discretion in the interpretation of the law. Although his action aroused opposition, and some rabbis continued to invoke their own collections, the authority of his office and the obvious advantages of a unified system of law soon outweighed centrifugal tendencies, and his Mishna attained quasi-canonical status, becoming known as "The Mishna" or "Our Mishna." For all its clarity and comprehensiveness, its phraseology was often obscure or too terse to satisfy all needs, and a companion known as the Tosefta ("Additions") was compiled shortly thereafter in which omitted traditions and explanatory notes were recorded. Since, however, neither compilation elucidated the processes by which their decisions had been elicited, various authorities set about collecting the midrashic discussions of their schools and recording them in the order of the verses of Scripture. During the 3rd and 4th centuries the tannatic Midrashim on the Pentateuch were compiled and introduced as school texts.

Fundamentally legal in character, this literature was designed to regulate every aspect of life—the six divisions of the Mishna on agriculture, festivals, family life, civil law, sacrificial and dietary laws, and purity encompass virtually every area of Jewish experience—and, accordingly, also recorded the principal Pharisaic and rabbinic definitions and goals of the religious life. One tract of the Mishna, *Avot* ("Sayings of the Fathers"), treated the meaning and posture of a life according to Torah, while other passages made reference to the mystical studies into which only the most advanced and religiously worthy were initiated; *e.g.,* the activities of the *Merkava,* or divine "Chariot," and the doctrines of creation (see below, *Jewish mysticism*). The rabbinic program of a life dedicated to study and fulfillment of the will of God was thus a graded structure in which the canons of morality and piety were attainable on various levels, from the popular and practical to the esoteric and metaphysical. Innumerable sermons and homilies preserved in the midrashic collections, liturgical compositions for daily and festival services, and mystical tracts circulated among initiates all testify to the deep spirituality that informed rabbinic Judaism.

THE AGE OF THE AMORAIM: THE MAKING OF THE TALMUDS (3RD–6TH CENTURIES)

Palestine (c. 220–c. 400). The promulgation of the Mishna initiated the period of the *amoraim* (lecturers or interpreters), those teachers who made the Mishna the basic text of legal exegesis. The curriculum now centred on the elucidation of the text of the standard compilation, harmonization of its decisions with extra-Mishnaic traditions recorded in other collections, and the application of its principles to new situations. The records of these amoraic studies have been preserved in the form of two running commentaries on the Mishna known as the Palestinian (or Jerusalem) Talmud ("Teaching") and the Babylonian Talmud, reflecting the study and legislation of the academies of the two principal centres of Jewish concentration in the Roman and Persian empires of that time. (Talmud is also the comprehensive term for the whole collections, Palestinian and Babylonian, containing Mishna, commentaries, and other matter. See below, *The literature of Judaism*)

The rabbinic educational system

The principal agencies mediating the rabbinic way of life and literature to the masses were the schools, ranging from the primary school to the advanced "house of study" and more formal academy (yeshiva), the synagogue, and the Jewish courts, which not only adjudicated litigations but also decided on ritual problems. Primary schools had long been available in the villages and cities of Palestine, and tannaitic law made education of male children a religious duty. Introduced at the age of five or six to Scripture, the student advanced at the age of 10 to Mishna and finally in mid-adolescence to Talmud or the processes of legal reasoning. Regular reading of Scripture in the synagogue on Mondays, Thursdays, the Sabbaths, and festivals, coupled with concurrent translations into the Aramaic vernacular and frequent sermons, provided for lifelong instruction in the literature and the values elicited from it. The amoraic emphasis on the moral and spiritual aims of Scripture and its ritual is reflected in their midrashic collections, which are predominantly homiletical (sermonic) rather than legal in content.

An amoraic sermon conceded that of every thousand beginners in primary school only one would be expected to continue as far as Talmud. In the 4th century, however, there were enough advanced students to warrant academies in Lydda, Caesarea, Sepphoris, and Tiberias (in Palestine), where leading scholars trained disciples for communal service as teachers and judges. In Caesarea, the principal port and seat of the Roman administration of Palestine, where pagans, Christians, and Samaritans maintained renowned cultural institutions, the Jews, too, established an academy that was singularly free of patriarchal control. The outstanding rabbinic scholar there, Abbahu (*c.* 279–320), wielded great influence with the Roman authorities and, because he combined learning with personal wealth and political power, attracted some of the most gifted students of the day to the city. In *c.* 350 the studies and decisions of the authorities in Caesarea were compiled as a tract on the civil law of the Mishna. Half a century later, the academy of Tiberias issued a similar collection on other tracts of the Mishna, and this compilation, in conjunction with the Caesarean material, constituted the Palestinian Talmud.

The role of the patriarchs

Despite increasing tensions between some rabbinic circles and the patriarch, his office was the agency providing a basic unity to the Jews of the Roman Empire. Officially recognized as a Roman prefect, a government official, the patriarch at the same time delegated apostles to Jewish communities to inform them of the Jewish calendar and of other decisions of general concern and to collect an annual tax of a half shekel paid by male Jews for his treasury. As titular head of the Jewish community of the mother country and as a vestigial heir of the Davidic monarchy, the patriarch was a reminder of a glorious past and of a hope for a brighter future. How enduring these hopes were may be seen from the efforts to gain permission to rebuild the Temple in Jerusalem. Although the emperor Julian (reigned 361–363) actually authorized the reconstruction, the project came to naught as a consequence of a disastrous fire on the sacred site and the subsequent death of the Emperor.

The adoption of Christianity as the religion of the empire had no direct effect on the religious freedom of the Jews; *i.e.,* on their freedom to worship and observe their life rules. The ever-mounting hostility between the two religions, however, resulted in severe curtailment of Jewish disciplinary rights over their coreligionists, interference in the collection of patriarchal taxes, restriction of the right to build synagogues, and, finally, upon the death of the patriarch Gamaliel VI in *c.* 425, the abolition of the patriarchate and the diversion of the Jewish tax to the imperial treasury. Though Mediterranean Jewry was now fragmented into disjointed communities and synagogues, the principles of the regulation of the Jewish calendar had been committed to writing *c.* 359 by the patriarch Hillel II, and this, coupled with the widespread presence of rabbis, ensured continuity of Jewish adherence. Even the emperor Justinian's (reigned 527–565) restrictions on synagogal worship and preaching apparently had no devastating effect. A new genre of liturgical poetry, combining ecstatic prayer with didactic motifs, developed in this period of political decline and won acceptance in synagogues in Asia Minor as well as beyond the Euphrates.

Babylonia (200–650). In the increasingly unfriendly climate of Christendom, Jews drew consolation in the knowledge that in nearby Babylonia (then under Persian rule) a vast population of Jews continued to live under a network of effective and autonomous Jewish institutions and officialdom. Steadily worsening conditions in Palestine had drawn many Jews to Persian domains, where economic opportunities and the Jewish communal structure enabled them to gain a better livelihood while living in accordance with their ancestral traditions. To regulate internal Jewish affairs and ensure the steady flow of taxes, the Parthian, or Arsacid, rulers (247 BCE–224 CE) had appointed *c.* 100 CE an exilarch, or "head of the [Jews in] exile"—who claimed more direct Davidic descent than the Palestinian patriarch—to rule over the Jews as a quasi-prince. In *c.* 220 two Babylonian disciples of Judah the Prince, Abba Arika and Samuel bar Abba, began to propagate the Mishna and related tannaitic literature as the yardsticks of normative practice. As heads of the academies at Sura and Nehardea, respectively, Abba and Samuel cultivated a native Babylonian rabbinate, which increasingly provided the manpower for local Jewish courts and other communal services. While the usual tensions between temporal and religious arms frequently erupted in Babylonia, too, the symbiosis of exilarchate and rabbinate endured uninterruptedly until the middle of the 11th century.

The model for Diaspora Judaism

Paradoxically, Babylonian rabbinism derived its ideological strength from its fundamentally unoriginal character. As a transplant of Palestinian Judaism it claimed historic legitimacy to the Sāsānid rulers (224–651), who protected Jewish practices against interference from fanatical Magian priests, and to native Jewish officials, who argued for the validity of indigenous Babylonian deviations from Palestinian norms. But ultimately the historic importance of this transplantation lay in Babylonia's serving as the proving ground for the adaptability of Palestinian Judaism to a Diaspora situation. Legal and theological adaptations generated by needs of the new locale and times inevitably effected changes in the religious tradition. The laws of agriculture, purity, and sacrifices all of necessity fell into disuse. The values embodied in these laws, however, and the core of the legal-theological system—ranging from doctrinal faith in the revelation and election of Israel, to the requirement that the individual live by the canons of Jewish civil and family law, and the establishment of a network of communal institutions modelled on those of the mother country—remained intact, thereby ensuring a basic continuity and uniformity to rabbinically oriented communities everywhere. The real contribution of the Babylonian rabbinate to Jewish religion lay, accordingly, in its demonstration of how Palestinian Judaism was to be implemented on Gentile soil. Since historic circumstances made Babylonia the mediator of this tradition to all Jewish communities in the High Middle Ages (9th–12th centuries), the Babylonian version of Jewish religion became synonymous with normative Judaism and the measure of Judaic authenticity everywhere.

"The law of the [Gentile] government is binding," the principle formulated by Samuel, head of the academy at Nehardea (died 254), summarizes the essential novelty in rabbinic reorientation to life on foreign soil. Whereas Palestinian rabbis had perforce to comply with imperial decrees of taxation de facto—and this was all that Samuel had in mind—Babylonian teachers now rationalized the legitimacy of governmental authority in this respect de jure and thus enjoined upon the Jews political quietism and submissiveness as part of their religious theory. In all other areas of civil law, the Jews were instructed by their rabbis to bring their litigations to Jewish courts and thus to conduct their businesses as well as their family lives by rabbinic law.

While the rabbis could obviously more effectively impose their discipline in matters of public law than in private religious practice, the density of the Jewish population in many areas of Parthia (northeastern Iran) and Babylonia facilitated the application of moral and disciplinary pressures. The most effective vehicle for the dissemination of their teachings was the academies, of which those of Sura and Pumbedita remained preeminent, where judges and communal teachers were trained. Frequent public lectures in the synagogues of the academies on Sabbaths and festivals were capped by public *kalla* (study-course) assemblies for alumni of the schools during the two months, Adar (February–March) and Elul (August–September), when the lull in agricultural work freed many to attend semiannual refresher instruction. These meetings were followed by regular popular lectures during the festival seasons that soon followed. Thus, while rabbis constituted a distinct class within the community, their efforts were oriented toward making as much of the community as possible members of an elite of learning and religious scrupulosity. The harmonious relations that obtained with but few interruptions over the centuries between the Sāsānian rulers and their Jewish subjects gave the Jewish population the air of a quasi-state, which the Jewish leadership frequently extolled as superior to the Jewish community of Palestine.

Redaction of the Babylonian Talmud, *c.* 500–650

The dissemination of the Palestinian Talmud probably stimulated the Babylonians to follow suit by collecting and arranging in similar fashion the records of study and decisions of their own academies and courts. The Babylonian Talmud, which apparently underwent several stages of redaction (*c.* 500–650) on the basis of the proto-Talmuds—the early collections of commentaries on the Mishna—used in the academies, accordingly became the standard of reference for judicial precedent and theological doctrine for all of Babylonian Jewry and all those communities under its influence. As had been the case with the Mishna, the redaction of the Babylonian Talmud was later designated by authorities as marking the end of a period in Jewish history, and the scholars who put the finishing stylistic touches, known as *savoraʾim* ("explicators"), were classified as a transitional stage between the *amoraim* and *geonim* (see below).

The enduring vigour of Jewish faith throughout these centuries is graphically demonstrated by the missionary activity of Jews throughout the ancient Middle East, especially in the Arabian Peninsula. Proud Jewish tribes living in close proximity to each other in the vicinity of Yathrib (later Medina, Muḥammad's home city), engaged in agriculture and commerce and also in proclaiming the superiority of their monotheistic ethos and eschatology (doctrine of last things). In Yemen (southwestern Arabia) the last of the Ḥimyarite rulers (reigned from *c.* 2nd century CE), Dhu Nuwas, proclaimed himself a Jew and finally suffered defeat (*c.* 525) as a consequence of Christian influence on the Abyssinian armies. Jewish missionaries, however, continued to compete with Christian missionaries and thus helped lay the groundwork for the birth of an indigenous Arabic monotheism—Islām—that was to alter the course of world history.

THE AGE OF THE GEONIM (C. 640–1038)

Triumph of the Babylonian rabbinate. The lightning conquests in the Middle East, North Africa, and the Iberian Peninsula by the armies of Islām (7th–8th centuries) provided the environmental framework for the basically uniform (*i.e.,* Babylonian) character of medieval Judaism. As a "people of the Book" (*i.e.,* of the Bible), the Jews were permitted by the Muslims to live under the same autonomous structure that had developed under Arsacid and Sāsānian rule. The heads of the two principal academies were now formally recognized by the exilarch, and through him by the Muslim caliphate (religiopolitical rulers), as the official arbiters of all questions of religious law and as the religious heads of all Jewish communities that came under Muslim sway. Known as *geonim* (plural of *gaon,* "excellency"), and conducting high courts manned by scholars assigned graded ranks, they drew their financial support from Jewish communities assigned to them by the exilarch. Religious questions and contributions were solicited from all Jewish communities, and these along with formal gaonic replies (*responsa*) were regularly publicized at the semiannual *kalla* convocations. Under the strong leadership of Yehudai, *gaon* of Sura (presided 760–763), the Babylonian rabbinate exerted vigorous efforts to replace Palestinian usage wherever it was still in vogue, including the study of Palestinian amoraic legal literature, by Babylonian practice and texts, thus making the Babylonia Talmud the unrivalled standard of Jewish norms everywhere. The success of this campaign is evidenced by the fact that the term Talmud, when unqualified, has ever since meant the Babylonian Talmud. Indeed, even in Palestine the Babylonian corpus displaced its older rival and caused the study of Palestinian Talmudic literature to be confined to circles of legal specialists.

The Karaite movement

Antirabbinic reactions. The firm, and on occasion oppressive, tactics of exilarchs and *geonim* generated antirabbinic reactions, especially in outlying areas where enforcement was difficult, in the form of sectarian and messianic revolts. Inspired in part by ancient Palestinian sectarian doctrines and in part by Muslim usage, the sects were by and large quickly and forcefully suppressed. In the 9th century, however, a moderate group under the leadership of Anan ben David, a disaffected member of the exilarchic family, successfully organized a dissident movement that soon developed into a formidable challenger of Rabbinite (a term first used for the Talmudic adherents by the dissidents) supremacy. Known as Karaites (Scripturalists), the new sect advocated a threefold program of (1) rejection of rabbinic law as a human fabrication and therefore an unwarranted, unauthoritative addition to Scripture, (2) a return to Palestine to hasten the messianic redemption, and (3) a re-examination of Scripture to retrieve authentic law and doctrine. Under the leadership of Daniel al-Qumisi (*c.* 850?), a Karaite settlement prospered in the Holy Land, from which it spread as far as northwestern Africa and Christian Spain. A barrage of Karaite treatises arguing new views of scriptural exegesis stimulated renewed study of the Bible and Hebrew language in Rabbinite circles as well. The most momentous consequence of these new studies was the invention of several systems of vocalization for the text of the Hebrew Bible (Christian Old Testament) in Babylonia and Tiberias in the 9th and 10th centuries. The annotation of the Masoretic (traditional, or authorized) text of the Bible with vocalic, musical, and grammatical accents in the Tiberian schools of the 10th-century scholars Ben Naftali and Ben Asher fixed the Masoretic text permanently and through it the morphology (basic form and structure) of the Hebrew language for Karaites as well as Rabbinites.

In the face of sectarian challenges, the *geonim* intensified their efforts against any deviation from Rabbinite norms and began to issue handbooks of Jewish law that set forth in concise and unequivocal terms the standards for correct practice. A number of these codes, notably the *Halakhot gedolot* ("Great Laws"), *Siddur Rav Amram Gaon* (on liturgical practice), and *She*ʾ*eltot* ("Disquisitions") by Aḥa of Shabḥa (*c.* 680–*c.* 752), attained authoritative status in local schools and further helped give a unitary stamp to medieval Judaism.

Secular culture and philosophy

The *geonim,* however, were powerless to halt several social developments in the 9th century that progressively undermined their hold even over Rabbinite communities. A renascence of Greek philosophy and sciences in Arabic translation, coupled with the progressive urbanization of the upper classes of all religioethnic groups in the centres of political, commercial, and cultural activity, generated a new intelligentsia that cut across religioethnic lines. Widespread skepticism in basic doctrines of faith such as creation, revelation, and retribution was most poignantly represented by latitudinarianism (the tendency to be flexible and tolerant about deviations from orthodox beliefs and doctrines) and by antinomian (anti-Mosaic-law) Gnostic groups that negated divine providence and omniscience. Ḥiwi al-Balkhī, a 9th-century skeptical Jewish pamphleteer, scandalized the faithful by an open attack on the morality of Scripture and by an expurgated edition of the Bible for schools that omitted "offensive" materials (*e.g.,* alleged stories of God acting dishonestly). A mystifying Hebrew tract entitled *Sefer yetzira* ("Book of Creation") posited in terse and enigmatic epigrams a novel theory of creation that betrayed unmistakable Neoplatonic influence. Karaites joined philosophically oriented intellectuals in heaping scorn on popular Rabbinite customs that smacked of superstition and, above all, on Talmudic homilies that referred to God in anthropomorphic terms.

Gaonic difficulties were compounded by the rise in North Africa and Spain of populous and wealthy Jewish communities that, thanks to the development of their own local schools and native talent, ignored the Babylonian academies or favoured one over the other with religious queries and, in consequence, with financial contributions. To the delight of dissidents and the chagrin of the faithful, competition between the Babylonian academies turned to internecine hostility. Occasional revolts against exilarchic taxation and administration in outlying areas of Persia had to be quelled with armed force. The Palestinian Rabbinites had revived their own academies, and their presidents now not only appealed for support in other Diaspora lands but challenged the authority of the Babylonians to serve as final arbiters on such matters of public import as the regulation of the calendar. By 900 the Rabbinite community of Babylonia was in a state of chaos and dissolution.

The gaonate of Saʿadia ben Joseph. In a bold effort to restore discipline and respect for the gaonate, an able exilarch, David ben Zakkai (916/917–940), bypassed the families from whom the *geonim* had traditionally been selected and in 928 appointed Saʿadia ben Joseph al-Fayyumi to head the academy of Sura. Of Egyptian birth, Saʿadia had gained wide acclaim for his scholarly retorts to Karaites, heretics, and Palestinian Rabbinites. Politically, Saʿadia's brief presidency was a fiasco and aggravated the chaos by a communal civil war. His gaonate, however, gave an official stamp to his many works, which responded to the ideological challenges to Rabbinism by restating traditional Judaism in intellectually cogent terms. Saʿadia thus became the pioneer of a Judeo-Arabic culture that was to come to full flower in Andalusian Spain a century later (see below *Sefardic developments*). His translation of the Bible into Arabic and his Arabic commentaries on Scripture made the rabbinic understanding of the Bible accessible to masses of Jews. His poetic compositions for liturgical use provided the stimulus for the revival of Hebrew poetry. Above all, his rationalist commentary on the puzzling "Book of Creation" and his brilliant philosophic treatise on Jewish faith, *Beliefs and Opinions,* synthesized Torah (the divine law in the Five Books of Moses and the rabbinic understanding of this revelation) and "Greek wisdom" in accordance with the dominant Muslim philosophical school of Kalām and thus made Judaism philosophically respectable and the study of philosophy a religiously acceptable pursuit.

Gaonates of Sherira and Hai

Far from tightening the gaonic hold over the Jewish communities of the Arabic world, Saʿadia's works actually provided the wherewithal for ever-greater intellectual and religious self-sufficiency. While economic, political, and military upheavals progressively weakened all institutional fabrics in the Middle East, concurrent prosperity and consolidation in the West stimulated the maturation of indigenous leadership in Egypt, al-Qayrawān (Kairouan; in present-day Tunisia), and Muslim Spain. To be sure, able *geonim* such as Sherira and his son Hai exercised enormous influence over the Judeo-Arabic world through hundreds of legal *responsa* issued in the course of their successive terms (968–1038) at Pumbedita. Circumstances beyond anyone's control, however, were bringing the curtain down on the effectiveness of exilarchate and gaonate. But by 1038, the year of Hai's death, the consequences of four centuries of gaonic activity had become indelible: the Babylonian Talmud had become the agent of basic Jewish uniformity; the synthesis of philosophy and tradition had become the hallmark of the Jewish intelligentsia; and the Hebrew classics of the past had become the texts of study in Jewish schools everywhere.

MEDIEVAL EUROPEAN JUDAISM (950–1750)

The two major branches. Despite the fundamental uniformity of medieval Jewish culture, the cultural–political divisions within the Mediterranean basin, in which Arabic-Muslim and Latin-Christian civilizations coexisted as discrete and self-contained societies, shaped the character of the Jewish subculture of the area. Two major branches of rabbinic civilization developed in Europe, the Ashkenazic, or Franco-German, and the Sefardic, or Andalusian-Spanish. Distinguished most conspicuously by their varying pronunciation of Hebrew, the numerous differences between them in religious orientation and practice derived, in the first instance, from the geographical fountainheads of their culture—the Ashkenazim (plural of Ashkenazi) tracing their cultural filiation to Italy and Palestine and the Sefardim (plural of Sefardi) to Babylonia—and from the influences of their respective immediate milieus. While the Jews of Christian Europe wrote for internal use almost exclusively in Hebrew, those of Muslim areas regularly employed Arabic for prose works and Hebrew for poetic composition. Whereas the literature of Jews in Latin areas was overwhelmingly religious in content, that of the other branch was well endowed with secular poetry and scientific works inspired by the cultural tastes of the Arabic literati. Most significantly, the two forms of European Judaism differed in their approaches to the identical rabbinic base that both had inherited from the East and in their radically different attitudes to Gentile culture and politics.

The Golden Age of Jewish Arabic and Hebrew culture

Sefardic developments. In Muslim Spain, Jews frequently served the government in official capacities and, therefore, not only took an active interest in political affairs but also engaged in considerable social and intellectual intercourse with influential circles of the Muslim population. Since the support of letters and scholarship was part of state policy in Muslim Spain, and since Muslim savants traced the source of Muslim power to the vitality of the Arabic language, scripture, and poetry, Jews looked at Arabic culture with undisguised admiration and unabashedly attempted to adapt themselves to its canons of scholarship and good taste. The hallmark of the cultured Jew accordingly became a polished command of Arabic style and the ability to display the beauty of his own heritage through a philological mastery of the text of the Hebrew Bible and through the composition of new Hebrew verse, now set to an alien Arabic metre. Since Arabic philosophers and scientists promulgated syntheses of Greek philosophy with the revelation to Muḥammad, rationalist study of the Jewish classics and defense of rabbinic faith in philosophic terms became dominant motifs in the Andalusian Jewish schools (in southern Spain).

The atmosphere generated a fever of literary creativity in classical Jewish disciplines as well as in the sciences cultivated by the Arabs that has gained for the period the title of "the Golden Age of Hebrew literature" (*c.* 1000–1148). What distinguished the Jewish culture of this age was not only the supreme literary merit of its Hebrew poetry, the new spirit of relatively free and rationalist examination of hallowed texts and doctrines, and the extension of Jewish cultural perspectives to totally new horizons—mathematics, astronomy, medicine, philosophy, political theory, aesthetics, belles-lettres—but also the frequent overlapping of the Sefardic religious leadership with the new Jewish courtier class. The unprecedented heights which the latter attained—Ḥisdai ibn Shaprut as counsellor to the caliphs of Córdoba, the Ibn Nagrelas as viziers of Granada, the Ibn Ezras, Ibn Megashs, and Ibn Albalias as high officials in Granada and Seville—and the distinctions of these men and of their protégês in Jewish and worldly letters restored the ancient integration of culture and practical life and generated a neoclassicism ("classicism" here meaning biblicism) that expressed the identification of the Jewish elite with the biblical age of Jewish power and artistic creativity. The effort to recapture the vitality and beauty of biblical poetry stimulated comparative philological and fresh exegetical research that yielded new insights into the morphology of the Hebrew language and into the historical soil of biblical prophecy. Judah ibn Ḥayyuj and Abū al-Walīd Marwān ibn Janāḥ produced manuals on biblical grammar that applied the results of Arabic philology to their own tongue and that have, accordingly, provided the principles of Hebrew grammatical study down to modern times. The anticipations of modern higher biblical criticism by Judah ibn Balaʿam and Moses ibn Gikatilla (flourished 11th century) were popularized in Hebrew a few generations later by Abraham ibn Ezra. In the revival of Hebrew poetry, liturgical as well as secular, that translated the new preoccupation with language and beauty into art, Andalusian Jewry saw its greatest achievements. Solomon ibn Gabirol, Moses ibn Ezra, and Judah ha-Levi were but the acknowledged supreme geniuses of a form of expression that became a passion with thousands the length and breadth of Spain. But by far the most enduring consequence of the new temper was their redefinition of religious faith in the light of Greco-Arabic philosophical theories. Solomon ibn Gabirol's exposition of faith in Neoplatonic terms, Abraham ibn Daud's defense of Rabbinism by Aristotelian categories, Judah ha-Levi's attack on philosophy as religiously bankrupt, and Moses Maimonides' epoch-making synthesis of Judaism and medieval Aristotelianism fixed philosophic inquiry as an enduring subject on the agenda of rabbinic concerns. A new class of philosophers that emerged in the 13th century and sponsored the translation of Arabic literature into Hebrew and of Hebrew and Arabic literature into Latin brought Jews and their thought into the mainstream of Western philosophy and gained for them the position of middlemen of culture between East and West.

Influence of Greco-Arabic philosophy

The salient trends of Sefardic Judaism did not imply relegation of the rabbinic class to a second place. Rather they shaped a fresh approach to rabbinic texts that paralleled in many respects those adopted in biblical exegesis. Strict adherence to consistency, systematization, and philological exactitude yielded new codes that often diverged from gaonic judgments. A digest of Talmudic law by Isaac Alfasi placed the Sefardic rabbinate on a self-reliant footing and epitomized its ideal of getting at the essentials of Talmudic law by sidestepping contingent discussions. In this area, too, it was Moses Maimonides who through his code of Jewish law, *Mishne Torah,* brought the Sefardic principles of comprehensiveness, lucidity, and logical arrangement to their apex. Written in Mishnaic Hebrew, the work remains to this day the only comprehensive treatment of all of Jewish law, including those fields that are not applicable in the Diaspora (agriculture, purity, sacrifices, Temple procedure).

With Maimonides, however, the pure Sefardic tradition came to an end, for the Almohad (Berber Muslim reformers) invasion of Spain in 1147–48 wiped out the Jewish communities of Andalusia and drove thousands either to northern Spain and Provence or, as in the case of Maimonides' family, to North Africa and Egypt. Sefardic Jewry suddenly encountered a discrete, mature, Jewish culture that for centuries had been developing independently and along quite different lines.

Ashkenazic developments. The spokesmen of Ashkenazic Jewry, into whose communities the Sefardim had been thrust by political events, regarded their own heritage and the Christian world in which they lived from a perspective shaped exclusively by rabbinic categories. From the world of the Talmud and Midrash they drew their school texts and the values that determined their judgments. Sensing no intellectual challenge in Christian faith, which they regarded with thinly concealed contempt, they constituted for the most part a merchant class that lived in urban centres under the protection of ecclesiastical and temporal rulers but under their own complex of laws and institutions. Except for mercantile relations, Christian society was closed to them, thanks largely to age-old ecclesiastical prohibitions forbidding all social intercourse with them. With the Arab conquest and the rise of the Carolingians (the 8th–10th-century dynasty that ruled France and Germany), the 12-decade interlude of suppression by the Visigoths (589–711) came to an end, and the Roman precedent of toleration and autonomy again became the rule. Merchants and rabbis moved from Italy to France and the Rhineland and infused new energies into the Jewish communities there. A native religious leadership began to emerge at the very time that Andalusian Jewry was entering its Golden Age. The bloody upheavals of the First Crusade (1096–99) in the communities of the Rhineland, although unleashing a tide of hatred, periodic violence, and progressive restrictions on Jewish activities, struck Jewish communities that had attained sufficient resilience to reestablish their communal institutions shortly afterward and continue the cultivation of their deeply ingrained traditions.

Ashkenazic cultural and religious patterns

By 1150 Ashkenazic Jewry had generated a culture pattern of its own with an indigenous literature that ranged from the popular homily to the esoteric tract on the nature of the divine glory. Study of the Bible and Talmud was oriented toward a mystical pietism in which prayer and contemplation of the secrets embedded in the liturgy were to lead to religious experience. Significantly, the fathers of the Ashkenazic tradition were remembered as liturgical poets and initiates into divine mysteries, and the early codes of the Franco-German schools were heavily weighted with discussions of liturgical usage. After the Second Crusade (1147–49), the German Jewish mystics (also called Ḥasidim, or pietists) placed heavy emphasis on the merits of asceticism, martyrdom and lifelong disciplines of penitence, thus adapting to Jewish idiom the features of saintliness celebrated in the universe of discourse of which they were a part. For the masses of Jews the cultural fare consisted principally of biblical tales and instruction, as interpreted by rabbinic Midrash, the lives of scholars and saints, and liturgical poetry reaffirming the election of Israel and faith in messianic redemption. The chief vehicle of popular instruction consisted of anthologies from the Rabbinic writings and commentaries on Scripture, of which the most popular was that of Rabbi Solomon ben Issac of Troyes, known as Rashi, the acronym formed from the initials of his name in Hebrew. For the more advanced student, Rashi composed a succinct commentary on the Talmud that, unmatched for compact thoroughness and lucidity, achieved an authority approaching that of the text itself.

As living sources of law and values, the Bible and Talmud had an impact that was apparent in communal decision and in the bearing of the leadership at home, in the marketplace, and in the synagogue. Taking their cue from Talmudic precedent and from Christian ecclesiastical procedures of their own times, the Ashkenazic rabbis occasionally gathered in regional synods to enact legislation on problems of a general nature for which there was no adequate precedent in the literature. Among the most enduring of these measures were the prohibition of bigamy and arbitrary divorce and severe economic penalties for abandonment of wives. Of far more immediate concern to the average Jew were the circumvention of Talmudic prohibitions against usury, relaxation of prohibitions regarding traffic with Gentiles in wines, and adoption of severe disciplinary measures, such as excommunication, against informers or those appealing, in cases involving Jews, to the Gentile authorities.

The rise and spread of Kabbala

A new religious trend began in Provence (a province of southeastern France) in the 13th century with the introduction into the Talmudic academies of a novel form of mystical study known as Kabbala (literally, "tradition"), which soon spread to northern Spain. Expressing Gnostic-type doctrines in rabbinic guise, the devotees of Kabbala devised an esoteric vocabulary that reinterpreted the Bible and rabbinic law as allegories of the various modes in which God is manifested in a spiritual universe, access to which was reserved for initiates. The most renowned literary product of this new circle was the *Zohar* ("The Book of Splendour"), a vast mystical commentary on the Pentateuch by Moses de León (*c.* 1275), which with later additions became the Bible of Jewish mystics everywhere. Although some of the theological notions of the Kabbalists deviated from basic postulates of Jewish monotheism, the insistence of the mystics on unflagging ritual orthodoxy and on a nominal acceptance of the biblical text as divine revelation helped them avert the suspicions aroused by Jewish Aristotelians and Averroists (followers of the 12th-century Arabic Aristotelian philosopher Averroës) and, in time, even won for them the status of a rabbinic elite. Indeed, some of the mystics lent their support to the antiphilosophic campaign that began in Montpellier, in southern France, *c.* 1200 and condemned the study of philosophy as generating skepticism, latitudinarianism, and disrespect for traditional literature. (For a fuller discussion of Kabbala see below, *Jewish mysticism.*)

Conflicts, disasters, and new movements. Basically, the conflict between "fundamentalist" and philosopher in Provence and northern Spain represented a clash between two mature Jewish subcultures of diverse geographic origins, the Sefardic and Ashkenazic, each of which had in the course of centuries developed different esoteric doctrines to transcend the legalistic formalism and confining dogmas of normative Judaism. Both forms of speculation sought salvation for exceptional individuals through knowledge and thus provided an immediate substitute for messianic deliverance from exile and servitude. Each group charged the other with distortion of tradition, and each issued apologias (defenses or justifications) and excommunications characteristic of medieval doctrinal controversy. While the rifts within communities attained bitter proportions, the common threat posed by ecclesiastical attacks on the Talmud in public disputations and by the expulsion of the Jews from France in 1306 prevented open rupture or resolution of the conflict. Ever since that time, two strands of orthodoxy representing the two forms of medieval metaphysical speculation have lived side by side in an uneasy truce.

Jews, longing for a return to the Holy Land, point to a visionary Jerusalem, which is depicted in the Gothic style of Christian church architecture. The Jews are shown with the pointed hats they were required to wear to distinguish them from Christians, and are represented with birds' heads, since they felt it was sacrilegious to depict the human form in sacred objects. Illustration from the Birds' Head Haggada, an illuminated manuscript from southern Germany, *c.* 1300.

By courtesy of the Israel Museum, Jerusalem

Most rabbinic circles of the 14th and 15th centuries displayed a progressive dogmatism and insistence on uniformity of practice. The great legal code of Jacob ben Asher of Toledo, *Arbaʿa ṭurim* (*c.* 1335; "Four Rows"), which sought to level differences in usage between Ashkenazim and Sefardim, bespoke the dominant trend of the rabbinate. The increasing hardening of ideological lines, however, did not eliminate independent thinking. Gersonides (Levi ben Gershom) gave Jewish Aristotelianism a new and comprehensive formulation, while Isaac Albalag propounded an Averroist (rationalistic) interpretation of the Bible predicated on a theory of double truth (of reason and revelation). In Muslim areas, the Maimonidean regimen of philosophic contemplation was extended by Maimonides' son Abraham to a quest for pietist ecstasy that betrayed many features of Ṣūfism (Islāmic mysticism).

Anti-Jewish riots and massacres of 1391 and a wave of apostasy in the wake of the disputation of Tortosa (1411–14)—which ended with a papal bull forbidding Talmud study, compelling attendance at Christian sermons, and other onerous measures—struck catastrophic blows in the Spanish communities and fed the anti-intellectualism of

Adversity and response: mysticism, messianism, rationalism

the rabbinate. Hasdai Crescas, while conceding the philosophic untenability of traditional belief in freedom of the will, launched a scathing attack on Aristotelian approaches to religion, and his disciple Joseph Albo issued a compendium on dogma that reaffirmed the traditional postulates of divine creation, revelation, and retribution as axioms of Judaism. But these reassertions of traditional faith could not overcome the ideological and social fragmentation that had split the Spanish communities into congealed strata that were often in open conflict with each other. Widespread marranism (ostensible conversion to Christianity) polarized the community and left deposits of bitterness that extended to those returning to the fold. The expulsions from Spain (1492) and Portugal (1497 and 1506) dealt the final blow and drove the escaping leadership into intensified pursuits of mystical escape from, and rationalization of, the endless calamities that befell their flocks. In Italy and the Ottoman Empire (Asia Minor, northeastern Africa, and southeastern Europe), the two principal centres of refuge for the exiles of the Iberian Peninsula, legalistic Kabbalism, which insisted on strict observance of the law as precondition of mystical practice and study, became the dominant spirit of a rabbinic leadership that in the face of terribly adverse circumstances continued to produce works of encyclopaedic proportions and staggering erudition in every field of Jewish learning.

Inspired by the Jewish tradition that the messianic era—when the messiah would come to bring in the rule of God—would be preceded by horrendous catastrophes, a group of single-minded rabbis established a community in Ẕefat (Safed), Palestine, where in anticipation of the new dawn all of life was to be conducted on principles of saintliness and mystical contemplation. Under the leadership of one Jacob Berab, the ancient practice of ordination was reinstituted in 1538 to form the nucleus of a revived Sanhedrin so as to administer ritual procedures requiring ordained authorities. While the effort failed because of rabbinic opposition, it reflected a widespread temper and further fanned messianic hopes sparked shortly before by the campaigns of tragic consequences by David Reubeni and Solomon Molkho in Italy, which ended in their being burned at the stake by the Christian authorities. In Ẕefat itself Kabbalism soon entered a new phase under the inspiration of Isaac Luria and Ḥayyim Vital, who confided to their disciples that the calamities of Israel were but a mirror of the captivity into which many sparks of the Godhead itself had fallen. Liturgical innovations and a novel mystical theology were formulated to redeem the imprisoned elements of divinity and thus restore creation to the harmony intended for it (see also below, *Jewish mysticism*).

That the Almighty himself was not quite omnipotent, at least with respect to the fate of his chosen people, was cautiously hinted in a Hebrew work of history (1550) by Solomon ibn Verga, who saw the Jewish problem as a sociopolitical one to which theological answers were futile. Such guarded rationalism was entertained by a number of courageous thinkers in 16th-century Italy, where, despite the policy of ghettoization (the segregation of the Jewish community in a restricted quarter) begun by Venice in 1516 and soon extended to all major Italian cities, the spirit of the Renaissance and the passion for historical criticism had captivated many Jews. Catholic scholars and prelates occasionally employed rabbis to instruct them in the Hebrew language and in the secrets of the Kabbala, which some Christians believed actually verified the postulates of their own faith. Contacts with Christian scholars in turn introduced Jews like Azariah dei Rossi, whose *Meor ʿenayim* ("Enlightenment of the Eyes") inaugurated critical textual study of rabbinical texts, to new bodies of literature that had been lost to the Jewish community, such as the works of Philo and Josephus (see above *Hellenistic Judaism*).

The Shabbetaian debacle

Such phenomena, however, were decidedly in the minority and contrary to the dominant trend. Dogmatic Kabbalism spread progressively and finally came to social expression in 1666 with the widespread acceptance of the views of the pseudo-messiah Shabbetai Tzevi (Sabbatai Zevi). Most of European and Ottoman Jewry was swept into a hysterical pitch in the belief that the end was now finally at hand. When the pseudo-messiah converted to Islām after being apprehended by the Ottoman government, mass despondency took the form of crypto-Shabbetaianism in which the apostasy of the messiah was explained as a form of voluntary crucifixion for the sake of the Jews. A witch-hunt on the part of traditionalists to uncover the cells of heresy unsettled Jewish communities everywhere by an emphasis on greater rigidity than before.

The following century (to *c.* 1750) was the darkest in the history of rabbinic Judaism. Scholarship reached an ebb of quality and popular religion a mechanical state such as Jews had never before experienced. The massacres and impoverishment of Polish Jewry after 1648 brought a pall over the growing eastern European centres of Jewish life. Antinomian eruptions of extreme Shabbetaians under the leadership of the self-proclaimed messiah and later Catholic convert Jacob Frank (1726–91) alarmed Gentile authorities almost as much as they did Jews. But the fossilization referred to above was only apparent. Beneath the surface many were restlessly searching for new avenues of faith, and the 18th century saw fresh responses that set the history of the Jews and of Judaism on new directions and spelled the beginnings of a new era. (G.D.C.)

Modern Judaism (c. 1750 to the present)

THE NEW SITUATION

The various criteria used to mark off dividing points in the history of the Jews and Judaism (see above *General observations*) are especially notable when it comes to setting a starting date for the modern period. Historians of thought put it in the late 17th century with the appearance of men, such as the philosopher Benedict de Spinoza, who ceased, in part or in toto, to believe in the inherited faith without at the same time ceasing to be Jews (*i.e.,* to consider themselves and be considered as Jews). Some Israeli scholars set it at about 1700 with the first stirrings of that new and continuing emigration from the Diaspora to the Holy Land that culminated in the mid-20th century in the creation of the State of Israel. Political and social historians set it in the mid- and late-18th-century processes that led to the American and French revolutions and to the results that flowed from these two epochal events, among them the emancipation of Jews from discriminatory and segregative laws and customs, the attainment of legal status as citizens, and the freedom of individual Jews to pursue careers appropriate to their talents. These varying approaches appear to have one thing in common—the view that these postmedieval forms of Jewish experience assume the end of the doctrine of the Exile, whereby Jews saw themselves as a people waiting out centuries of woe in alien lands until the moment of divine redemption. Jewish modernity for most scholars, then, is marked by the end of a passive waiting on the Messiah and the beginning of an active pursuit of personal or national fulfillment on this earth and preferably in one's lifetime.

Although the 18th century Haskala (Enlightenment) among the Ashkenazim of central and eastern Europe is often taken as the starting point of Jewish modernity, the process of Westernization had begun a good deal earlier among the Sefardim in western Europe and Italy. The Marranos who went to such communities as Amsterdam and Venice in the 17th century to declare themselves as Jews carried with them the Western education that they had acquired while living as Christians in the Iberian Peninsula, and the habits of criticism that had kept them from assimilating into the majority during their Marrano years and that some, such as Spinoza, a son of Marranos, used in analyzing all of the biblical tradition, including especially their own religion. In Italy there was an older Jewish community that had never been sealed off culturally from the influence of its environment; some of its figures were influenced by, and participated in, the main currents of the Renaissance (see above *Rabbinic Judaism*).

Increased contact with Western languages, manners, and modes of living came to the Ashkenazim only in the 18th century when new economic opportunities created such possibilities and needs. Jewish bankers and factors in various German principalities, army provisioners in most of the European countries, capitalists who were permitted to live in such places as Berlin because they opened new factories or were otherwise helpful to the expansion of the economy—all were in increasing contact with Gentile society, and most of them began to look upon the goal of their lives as the winning of full acceptance. Around this wealthy element there arose a number of intellectuals who agitated for the end of ghetto ways as the necessary preamble to the emancipation of the Jews.

THE HASKALA, OR ENLIGHTENMENT

The role of Moses Mendelssohn

In central Europe. By far the most outstanding figure of the 18th-century Jewish Enlightenment was the philosopher Moses Mendelssohn, who, while remaining a devoted adherent of Orthodox Judaism, turned away from the traditional Jewish preoccupation with the Talmud and its literature to the intellectual world of the European Enlightenment, of which he became the foremost Jewish representative. Mendelssohn did not attempt a philosophical defense of Judaism until pressed to do so by Christians who questioned how he could remain faithful to what they saw as an unenlightened religion. In his response, *Jerusalem,* published in 1783, Mendelssohn defended the validity of Judaism as the inherited faith of the Jews by defining it as revealed divine legislation and declared himself at the same time to be a believer in the universal religion of reason, of which Judaism was but one historical manifestation. Aware that he was accepted by Gentile society as an "exceptional Jew" who had embraced Western culture, Mendelssohn's message to his own community was to become Westerners, to seek out the culture of the Enlightenment. To that end he joined with a poet, Naphtali Herz (Hartwig) Wessely, in translating the Torah into German, combining Hebrew characters with modern German phonetics in an effort to displace Yiddish, and wrote a modern biblical commentary in Hebrew, the *Beʾur* ("Commentary"). Within a generation, Mendelssohn's Bible was to be found in almost every literate Jewish home in central Europe and had served to introduce its readers to German culture. Through his personal example and his life's work Mendelssohn made it possible for his fellow Jews to join the Western world without sacrificing their Judaism; he had convinced them that their intellectual processes were those of universal reason, with which Judaism accorded.

Educational reform

Mendelssohn's work was carried on by a group of Jewish intellectuals who had gathered around him in his lifetime, forming the nucleus of the Berlin Haskala, which was most active in the 20 years following their mentor's death. In the pages of their Hebrew-language periodical, *ha-meʾassef* ("The Collector"), they preached the virtues of secular culture and used Hebrew as a vehicle by which to introduce that culture. To achieve their goal of an enlightened Judaism, the leaders of the Berlin Haskala publicized the need for secular education. In response to the Holy Roman emperor Joseph of Austria's Edict of Toleration of 1781, Naphtali Wessely welcomed the efforts and issued an urgent call for the reform of Jewish education as a prelude to full emancipation. Purely secular subjects—mathematics, German, and world history and literature—were to take precedence over the traditional Jewish studies. The study of the Bible, since it was generally acknowledged to be a fundamental part of Western culture, was to be emphasized at the expense of the more traditional learning in the Talmud. Following this model, modern Jewish schools were established by Jewish intellectuals and businessmen in several German cities, among them Frankfurt and Hamburg. As its educational activities began to bear fruit in the wide dissemination of secular culture, the Berlin Haskala abandoned the use of Hebrew for German and gradually disintegrated. Unlike Mendelssohn himself, the immediate descendants of his circle and his own children were unable to strike a balance between Jewish and secular culture; their Western education undermined their religious faith and they perceived their identity as Europeans rather than as Jews.

One of Mendelssohn's disciples, David Friedlaender, offered to convert to Christianity without accepting Christian dogma or Christian rites; he felt that both Judaism and Christianity shared the same religious truth but that there was no relation at all between Judaism's ceremonial law and that truth. The offer was refused unless Friedlaender would acknowledge the superiority of Christianity and make an unconditional commitment to it, which he was not prepared to do. Unlike Friedlaender, many others who began by following Mendelssohn chose to leave the Jewish faith as the only way to win full acceptance in the European community of which they felt themselves a part.

The modern Hebrew Renaissance

In eastern Europe. The Haskala, thus, was quickly played out in central Europe; as an idea its further career was to continue in eastern Europe, particularly in the Russian Empire, where it flourished in the middle third of the 19th century until, as a result of the pogroms of 1881, Jews lost faith in the goodwill of Russians to accept "enlightened" Jews. It was a tenet of the Russian Haskala that the tsar was a benevolent leader who would bestow emancipation upon his Jewish subjects as soon as they proved themselves worthy of it; and that it was the task of the Jews, then, to transform themselves into model citizens, enlightened, unsuperstitious, devoted to secular learning and productive occupations. Following the example of the Berlin Haskala, a Russian Hebrew-language writer, Isaac Baer Levinsohn, published a pamphlet, *Teʿuda be-Yisrael* ("Testimony in Israel") citing the benefits of secular education. At the same time, such writers as Joseph Perl and Isaac Erter, though Orthodox Jews themselves, in virulent satire attacked the superstitious folk customs of the masses and opened the way to the anticlericalism which was to become characteristic of the Russian Haskala. In the 1840s and 1850s the emphasis shifted from satire and attack on the cultural parochialism of the Pale of Settlement (the regions to which the Jews were restricted) to romanticization of life outside the Pale, including periods of the Jewish past. Thus, Hebrew poets and novelists, such as Michal Levensohn and Abraham Mapu, arose on Russian soil to contribute their talents to the creation of a modern Hebrew literature. With the climate of government reforms in the 1860s, the Russian Haskala entered a "positivist" phase, calling for practical social and economic reforms. Hebrew-language journals were established and the Hebrew essay and didactic poetry, calling for religious and cultural reforms, came into their own, particularly at the hands of such stylists as the poet Y.L. Gordon and the essayist Moses Leib Lilienblum. Abandoning the original Hebrew and German orientation of the Russian Haskala, a number of Jewish intellectuals, the most prominent of whom were Yoachim Tarnopol, Osip Rabinovich, and Lev Levanda, became Russifiers, founding Russian-language Jewish weeklies devoted to "patriotism, emancipation, modernism." Like their contemporary fellow Jews in western Europe, they declared themselves to be Russians by nationality and Jews by religious belief alone. In 1863 a group of wealthy Jews in St. Petersburg and Odessa created the Society for the Promotion of Culture among the Jews of Russia for the purpose of educating Jewry into "readiness for citizenship." The goal of all segments of the Russian Haskala in the 1860s and 1870s was to turn Jews into good Russians and to make their Jewishness a matter of personal idiosyncrasy alone. The period of reaction that set in with the pogroms (massacres) of 1881 was to prove how deluded the hopes of the Haskala had been.

RELIGIOUS REFORM MOVEMENTS

One element of Westernization that the Haskala had championed was the reform of religion. It began in western Europe during the Napoleonic period (1800–15) when certain aspects of Jewish belief and observance were seen as incompatible with the new position of the Jew in Western society. Napoleon convoked a Sanhedrin (Jewish legislative council) in 1807 to create a new, modern definition of Judaism in its renunciation of Jewish nationhood and national aspirations, its protestations that rabbinic authority was purely spiritual, and its recognition of the priority of civil over religious authorities even in the matters of intermarriage. In areas other than France, the rationale for reform, at least in its early years, was more aesthetic than doctrinal. The external aspects of worship—

German Reform Judaism

i.e., the form of the service—appeared unacceptable to the newly Westernized members of the Jewish bourgeoisie in both Germany and the United States, whose standards of cultural acceptability had been shaped by the surrounding society, and who desired above all to resemble their Gentile peers. Thus, the short-lived Reform temple established in Seesen, by the pioneer German reformer Israel Jacobson, in 1810 enshrined order and dignity of a Protestant type in the service and introduced an organ, sermon, and prayers in German, in place of Hebrew, to create an uplifting spiritual experience. The more radical temple in Hamburg (established 1818) adopted all of Jacobson's reforms and published its own much-abridged prayer book, which deleted almost all the references to the long-awaited restoration of Zion. Reformers in Charleston, South Carolina, introduced similar changes in the synagogue ritual in 1824, for they sought a non-national Judaism similar in form to Protestantism and adapted to the surrounding culture. It was apparent to the reformers that in Western society Judaism would have to divest itself of its alien customs and conform to the cultural and intellectual standards of the new "age of reason."

German Reform in the 1840s became institutionalized, a matter of organized, formal belief and practice, and, at a series of synods held at Brunswick (1844), Frankfurt (1845), and Breslau (1846), it created the first theological rationalization for changes introduced in the previous generation. Judaism, it was declared, had always been a developmental religion that conformed to the demands of the times, and, since the Jews were not now a nation, they were no longer bound by their entire religious–political code of law, but only by the dictates of moral law. Thus, those rituals which stood in the way of full Jewish participation in German social and political life were no longer considered valid expressions of Jewish religious truth. The use of Hebrew in religious services was limited; practices such as the dietary laws and circumcision and all national messianic hopes were discarded upon the altar of the "spirit of the times." Messianism in Reform Judaism was transmuted into active concern for social welfare in the present, and the Jewish role in history became Diaspora-centred, a mission to the Gentiles.

Reform Judaism in the United States

Although Reform was initiated in Europe, it did not enjoy a successful career there, for many central European governments that regulated the existence of religious communities would not recognize more than one form of Judaism in any one locale. Reform did not achieve its greatest success until it was imported into the United States along with the massive German-Jewish immigration of the 1840s and coalesced with earlier American trends toward reform. By 1880 almost all of the 200 synagogues in the United States had become Reform, amalgamating in the Union of American Hebrew Congregations (formed 1873). In 1885 the Reform philosophy was given its most comprehensive formulation in the so-called Pittsburgh Platform, drawn up by a conference of Reform rabbis. This manifesto declared that Judaism was an evolutionary faith, and no longer a national faith, and that it was now to be de-orientalized. While the preservation of historical identity was considered to be beneficial, the maintenance of continuity of tradition was not; the Talmud was to be considered merely as religious literature, and not as legislation. The rationalist principles of the Pittsburgh Platform remained the official philosophy of the American Reform movement until a later generation, seeking to meet different emotional and intellectual needs, reintroduced the concept of Jewish peoplehood into the Columbus Platform of 1937, which also reemphasized Hebrew and traditional liturgy and practices. Classical (19th-century) Reform was very much a late child of the Enlightenment, and by mid-20th century its Enlightenment philosophy appeared antiquated to many Jewish thinkers.

Conservative Judaism

If Reform was a child of Enlightenment rationalism, Conservative Judaism was a child of historical romanticism. It began in 1845 when Zacharias Frankel and a group of followers seceded from a second Reform synod at Frankfurt over the issue of the limitation of the use of Hebrew to a small core of prayers. For Frankel, Hebrew represented the spirit of Judaism and the Jewish people, and Judaism itself was not merely a theology of ethics but the historical expression of the Jewish experience; this definition he called "positive-historical Judaism." Although Conservative Judaism conceived of Judaism as a developmental religion, it charted its course through close study of the tradition and the will of the people, and thus came to largely traditional conclusions about religious observance.

ORTHODOX DEVELOPMENTS

In western and central Europe. The bulk of the official Jewish establishment in western and central Europe, though affected by the efforts at religious reform, remained Orthodox (a term first used by Reform leaders to designate their traditionalist opponents). Under the leadership of Samson Raphael Hirsch, a more modern and militant form of Neo-Orthodoxy arose, based in Frankfurt am Main, which asserted its right to break with any Jewish community that contained Reform elements and to form an independent community. The thinking of this group was profoundly influential, for it indicated the possibility of living a ritually and religiously full life while being totally integrated into Western society. It accomplished this by positing a theoretical division between religion and culture; the Jews were to remain Orthodox in religion (although deferring their messianic aspirations to the unforeseeable future) while becoming Western in manners and culture. This form of Orthodoxy, which became the intellectual model for Western Orthodoxy, continued into the late 20th century in the United States in a variety of religious and academic institutions (such as the Yeshiva University and the bulk of English-speaking Orthodox synagogues), coexisting in substantial tension with a number of Orthodox groups, most notably the Lubavitcher and Satmar Ḥasidim (for Ḥasidism, see below), and some Talmudic academies that saw the Western world itself as the enemy and chose to recreate the ghetto.

Ḥasidism

In eastern Europe. By the mid-18th century Orthodoxy in eastern Europe, having been convulsed by frantic messianism and stifled by the sterility of purely legalistic scholarship, was ripe for revival. The experience of Shabbetaianism (the first messianic movement to excite virtually all of world Jewry) had revealed in the mid-17th century the pervasiveness of Jewish exhaustion with the Exile and fervent longing for messianic redemption, while the nihilistic sect of Frankists (the followers of Jacob Frank, see above *Rabbinic Judaism*) in the 18th century had transmuted that messianism into a this-worldly hysteria. Talmudic piety and study, sunk in excessive *pilpul* (acute logical distinctions that often became mere hairsplitting), was refreshed by the new critical methods of Elijah ben Solomon, the *gaon* of Vilna. Although essentially a legal rigorist, he was open to Western scientific learning insofar as it helped him to elucidate Talmudic texts. Orthodox religious expression also was raised to a new level with the development of Ḥasidism (Pietism) by Israel Baal Shem Tov in the mid-18th century. Although Ḥasidism contained elements of social protest, being at least in part a movement of the poor against the wealthy communal leadership and of the unlearned against the learned—though many of its leaders, among them Rabbi Baer, the *maggid* ("preacher") of Mezhirich and Rabbi Levi Isaac of Berdichev, were well-versed in Talmudic learning—it was essentially a non-messianic outcry in the name of religious emotion, emphasizing prayer and personal religious devotion here and now. Contemporary scholarship is investigating the linkage between Ḥasidism and eastern European Christian pietistic movements. The major innovation that Ḥasidism introduced into Jewish religious life was the charismatic leader, the *rebbe* who served as teacher, confessor, wonder-worker, God's vicar on earth, and occasionally as atoning sacrifice. Although the earliest *rebbes* were democratically chosen, the position of leadership passed to their descendants on the presumption that they had inherited their fathers' charisma and thus created spiritual dynasties. Ḥasidism spread throughout eastern Europe and enjoyed its greatest success in Poland.

Ḥasidism was notably unsuccessful, however, in Lithuania, where the traditional rabbinic class, under the leadership of Elijah, the Vilna *gaon,* was able to stave off its influence by issuing a ban of excommunication (*ḥerem,* "anathema") against the new movement. The tactic (a complete boycott and cutting off of communication) was widely embraced by non-Ḥasidic rabbis, who earned for themselves among the Ḥasidim the title of Mitnaggedim (Opponents), but it proved largely ineffective in areas where the rabbis had lost the respect of the masses, and it called forth a round of counter-excommunications by the Ḥasidic *rebbes.* With the passage of time, Ḥasidim and Mitnaggedim abandoned their conflict and came to see each other as allies against the threat to all Orthodox Jewish religion of Haskala and secularization. The impact of Ḥasidism on eastern European Jewry cannot be overemphasized; even in Lithuania, where it did not take firm hold, it stimulated the growth of a home-grown pietism in the Musar (ethicist) movement of the mid-19th century, and it renewed the Talmudic energies of its opponents.

DEVELOPMENTS IN SCHOLARSHIP

The Science of Judaism

As Jews moved into Western society in central Europe, there arose a group of young Jewish intellectuals who devoted themselves to Jewish scholarship of a far different type from traditional Talmudic learning or medieval philosophy. In 1819 Leopold Zunz and Moses Moser founded the Society for Jewish Culture and Learning for the study of Jewish history and literature. Although the original group quickly dissolved, Zunz became the unofficial leader of a generation of scholars dedicated to the Wissenschaft des Judentums (Science of Judaism). Under its carefully objective and scholarly facade, the Wissenschaft movement embraced a variety of nonacademic motives and goals. All of its members sought to prove that the Jewish past was intellectually respectable and worthy of study, and hence that the Jews deserved an equal place within European societies. Jewish scholarship was enlisted as a weapon in the battles for change. Thus, Isaac M. Jost wrote a general history of the Jews to promote Reform; Zunz's *Gottesdienstliche Vorträge der Juden, historisch entwickelt* (1832; "The Worship Sermons of the Jews, Historically Developed") served to legitimize the modern innovation of the sermon in the vernacular; and Abraham Geiger, the outstanding leader of German Reform in the 1840s and 1850s, interpreted the Pharisees as the forerunners of the reformers of his own day. In their work these intellectuals presented archetypes of what modern Jews should become. To support their claims of academic respectability, the Wissenschaft figures highlighted those aspects of the Jewish past that were closely integrated with general fields of study. In particular, Moritz Steinschneider, who owes his fame to towering achievements in bibliography, was concerned above all with the contribution of Jews to science, medicine, and mathematics. Nineteenth-century Jewish scholarship set out to praise Judaism as one of the cofounders of the Western tradition, and thus to argue that whenever the Jews were not excluded from European society they have produced great culture, and that they would repeat such accomplishments under conditions of social and political equality.

The Wissenschaft movement stimulated the critical study of the Jewish past, and great works of synthesis, written from a variety of perspectives, began to appear: Heinrich Graetz's multivolumed *Geschichte der Juden von den ältesten Zeiten bis auf die Gegenwart* (1853–75; *History of the Jews*), written from a romantic-national point of view; Isaac Halevy's *Dorot ha-rishonim* (1897–1932; "The First Generations") and Ze'ev Jawitz's *Toldot Yisrael* (1894; "History of Israel"), from an orthodox standpoint; and Simon Dubnow's *Weltgeschichte des jüdischen Volkes* (1925–29; "World History of the Jewish People"), reflecting his belief in secular, nationalistic communal autonomy. Since the 1920s this tradition of great synthesis has been carried on in the United States by Salo W. Baron, who by the early 1970s had produced 14 volumes of his *Social and Religious History of the Jews* (1952), and in Israel by Benzion Dinur, whose chief work was *Yisrael ba-gola* (3rd ed., 5 vol., 1961–66; "Israel in the Exile"). Many other first-rank scholars in Europe, Israel, and the United States have made notable contributions to the study of Jewish history, rabbinics, and mysticism. This great emphasis on historical research and knowledge from a wide variety of perspectives tended to propose veneration for the Jewish past as a substitute for waning religious faith.

JEWISH–CHRISTIAN RELATIONS

Christianity and Russian anti-Semitism

Jewish–Christian relations in the 19th century, strained at best, often erupted into open conflict. Established Christianity, and Roman Catholicism in particular, staunch upholders of the old order, identified the Jews as the major beneficiaries of the French Revolution and as the bearers of a liberal, secular, anticlerical, and often revolutionary threat. Clerical anti-Semitism was thus in France allied with the anti-Semitism of the traditional right, and these movements contended with those who affirmed the results of the French Revolution in the great convulsion of the "Dreyfus affair" in the last years of the 19th century. In Russia the conflict of the Jews and the Orthodox Church released the most open and virulent manifestation of religious anti-Semitism. To the church, the Jews were the enemy seeking to undermine Russian Orthodoxy and the tsar, the very foundations of Russian tradition. The church and the tsarist authorities went so far as to condone, and even encourage, the violent pogroms that were perpetrated against the Jews in 1881–82 and again in 1905. Russian Orthodoxy was active as well in spreading the so-called blood libel, a superstitious belief in Jewish ritual murder which had reemerged even in the 19th century, in Damascus in 1840 (in which instance the French Consul in Syria initiated the accusation) and in Hungary in the Tiszaeszlár affair in 1882. In both cases torture was used to obtain false confessions but the accused were ultimately cleared. The most infamous recurrence of the blood libel in modern times, however, was the Beilis case of 1911–13, in which the tsarist government, with church complicity, sought unsuccessfully to convict a Jewish bookkeeper in Odessa of ritual murder. From Russian Orthodox circles, too, arose the *Protocols of the Learned Elders of Zion,* a fraudulent documentation of an alleged international Jewish conspiracy to conquer the world by subverting the social order through Liberalism, Freemasonry, and other modern movements; the concoction appeared around the turn of the century and enjoyed a phenomenal success in anti-Semitic propaganda. In spite of the fact that much of modern anti-Semitism was not Christian but racial, pagan, and often left-wing, Jews have attributed even secular anti-Semitism to older Christian teachings, which they assert persisted as a case of time lag. In the 20th century Jews and Christians have moved toward mutual understanding. In the early decades of the century, some liberal Christian voices were raised against anti-Semitism; in the United States the National Conference of Christians and Jews was founded (1928) as a response to anti-Semitism propagated in Henry Ford's *Dearborn Independent.* Elements of the Church spoke out during the 1930s against the Nazi persecution of the Jews, but the majority of Christian religious figures in Europe remained silent, even during the Holocaust (near extermination of European Jews). In response to the Holocaust, however, the World Council of Churches denounced anti-Semitism in 1946, and in 1965 the Roman Catholic Church's Schema on the Jews and other non-Christian religions, adopted by the Second Vatican Council, revised its traditional attitude toward the Jews as the killers of Christ. A growing sense of ecumenism (of fellowship and common concerns) has been shared by Jews and Christians alike. Although there remain many difficulties related to the question of the place that Zionism and the State of Israel hold within Judaism, the older forms of official church anti-Semitism have radically lessened.

Christianity and the Nazi persecution

ZIONISM

The most striking of the new phenomena in Jewish life is Zionism, which, insofar as it has focussed on the return to Zion (the poetic term for the Holy Land), is a re-echo of older religious themes. Insofar as it has stressed the national concentration of the Jews in a secular state, however, it is yet another example of the secularization of Jewish life and of Jewish messianism. In its secular aspects Zionism attempted to complete the emancipation of the

Jews by transforming them into a nation like all other nations. Although it drew upon the general currents of 19th-century European nationalism, its major impetus came from the revival of a virulent form of racist anti-Semitism in the last decades of the 19th century, as noted above. Zionism reacted to anti-Semitic contentions that the Jews were aliens in European society and could never hope to be integrated into it in any numbers, and transformed this charge into a basic premise of a program of national regeneration and resettlement. Zionism has come to occupy roughly the same place in Jewish life as the "social gospel"—according to which the Kingdom of God is to be achieved in economic and social life—for Christians; the involvement in Israel as the new centre of Jewish energies, creativity, and renewal serves as the secular religion of many Diaspora Jews.

AMERICAN JUDAISM

The central and eastern European immigrant styles of Judaism

The story of Judaism in the United States is the story of several fresh beginnings. In the colonial period the style of the tiny American Jewish community was shaped by the earliest Sefardic immigrants; the community was officially Orthodox but, unlike European Jewish communities, was voluntaristic, and by the early 19th century there was a significant drift of the younger generation from Judaism. By the mid-19th century a new wave of central European immigrants revived the declining American Jewish community and remade it to serve its own needs. Primarily petty shopkeepers and traders, the new immigrants migrated westward, founding new Jewish centres which were almost entirely controlled by laymen. The exigencies of life on the frontier within an open society created a predisposition for religious reform, and it is significant that the greatest American Reform Jewish leader of the 19th century, Isaac Mayer Wise, was based in Cincinnati, Ohio. Wise sought to unite all of American Jewry in the new nontraditional institutions that he founded, Hebrew Union College (1875), the Union of American Hebrew Congregations (1873), and the Central Conference of American Rabbis (1889); but his ever more radical reforming spirit ultimately drove the traditionalist elements within the American Jewish community into opposition. The head of the traditionalist circles had been Isaac Leeser, a native of Germany, who had attempted to create an indigenous American community on the lines of a modernized traditionalism. Conservative forces, after his death, became disorganized, but in reaction to Reform they defined themselves by their attachment to the sabbath, the dietary laws, and especially to Hebrew as the language of prayer. Under the leadership of Sabato Morais, an Orthodox Jew of Italian birth, Conservative circles in 1886 founded a rabbinic seminary of their own, the Jewish Theological Seminary of America.

American Orthodoxy

The eastern European immigrants who moved in large numbers to the American shores in the years from 1881 to 1914 were profoundly different in culture and manners from the older elements of the American Jewish community, and it is they and their descendants who made American Judaism as it is today. The bridge between the existing Jewish community led by German Jews of Reform persuasion and the new immigrant masses was the traditionalist element among the older settlers. Cyrus Adler, traditionalist himself, cooperated with a German Reform circle of Jacob Schiff in reorganizing the Jewish Theological Seminary (1902) and other institutions for the purpose of Americanizing the eastern European immigrants. Enough eastern European rabbis and scholars had immigrated, however, to create their own synagogues, which reproduced the customs of the old world. Whereas in 1880 almost all of the 200 Jewish congregations in the United States were Reform, by 1890 there were 533 synagogues, and most of the new ones founded by immigrant groups were Orthodox. The Union of Orthodox Jewish Congregations, which was established in 1898 by elements associated with the Jewish Theological Seminary, was soon taken over by Yiddish-speaking recent immigrants for whom the seminary was much too liberal. In 1902 immigrant rabbis also formed their own body, the Union of Orthodox Rabbis of the United States and Canada (the Agudath ha-Rabbanim), which fostered the creation of yeshivas (rabbinic academies) of the old type. In 1915 two small yeshivas, Etz Chaim and Rabbi Isaac Elhanan Theological Seminary, added Yeshiva College of secular studies in 1928, and became Yeshiva University in 1945. The eastern European Orthodox elements concentrated primarily on Jewish education and it was they who introduced the movement for Jewish day schools, analogous to Christian parochial schools. Gradually an American version of Orthodoxy developed on the Neo-Orthodox model of Samson Raphael Hirsch, which combined institutional separatism and cooperation with other Jewish groups in umbrella organizations.

The immigrants and their children had three desires; to upgrade themselves socially by joining older congregations or forming their own in an Americanized image; to affirm an unideological commitment to Jewish life; and to maintain their ties to the overseas Jewish communities of their origin. With their strong sense of Jewish peoplehood, they introduced Zionism into American Jewish life, and accepted the basic ideas of Mordecai Kaplan's Reconstructionism, which was committed to Zionism. A small group of anti-Zionists remained a significant force in the 1930s and 1940s, but their central organization, the American Council for Judaism, represented the descendants of earlier German-Jewish immigrants. The later immigrants took over all the earlier institutions of the Jewish community and imbued them with their own spirit.

The religious continuum

American Jewish religious life is a continuum—from the most traditional Orthodoxy to the most radical Reconstructionism. In theory, all of the Orthodox groups agree on the revealed nature of all of Jewish law; for the Reform groups, the moral doctrine of Judaism is divine and its ritual law is man made; the Conservatives see Judaism as the working out in both areas of a divine revelation that is incarnate in a slowly changing human history; and the Reconstructionists (who include both Conservative and Reform Jews) view Judaism as the evolving civilization created by the Jewish people in the light of its highest conscience. What really marks the various bodies in the mind of the Jewish community is their difference in ritual practices, but the ritual variations shade from one group into the other. The role of the rabbi is substantially the same in all three groups; he is no longer a Talmudic scholar but a preacher, pastor, and administrator, a cross between a parish priest and the leader of an ethnic group. Although there was some cooperation among the three major Jewish denominations—Orthodoxy, Reform, and Conservatism—the real effort of organized Jewish religion in America in the late 20th century revolved around the individual synagogue and the denomination to which it belonged. As religious identification became increasingly respectable in American life, the Jews followed the American norm, affiliating in greater numbers with synagogues, though often for ethnic or social, rather than religious reasons.

JUDAISM IN OTHER LANDS

Modernity came first to the Jewish people in Europe and it was, therefore, within the European context that representatives of important non-Ashkenazi communities such as the proto-Zionist Sefardi Judah ben Solomon Ḥai Alkalai of Sarajevo, the Luzzatto family in Italy, and Elijah ben Abraham Benamozegh in France, participated in variations of Jewish modernity. In England and France, more than in Germany or Russia, Wissenschaft des Judentums (see above *Developments in scholarship*), with its enlightenment ideology, was the central focus of Jewish experience; there the "republic of scholarship" became the synagogue of the Jewish intelligentsia. In neither country did Reform Judaism gain a major foothold, for the Orthodox establishment, which remained the official synagogue, liberalized its synagogue practice while retaining its essentially conservative outlook. In Anglo-Jewish life in the last decades of the 19th century the two most pronounced modernist tendencies were Solomon Schechter's moderate romantic traditionalism and the "renewed Karaism" of Claude Joseph Goldsmid Montefiore, whose version of religious reform was "back to the Bible."

Outside of Europe, in such places as South America and Canada, Jewish modernity appeared late, for European Jewry arrived in those places even later than in the United States, attaining a significant number only in the 20th century. These communities have been dependent on immigrant scholars and intellectuals for serious Jewish thought. Jews in the Arab lands, in North Africa and the Middle East, living within traditional societies, entered modernity even later than those on the peripheries of Europe. Many of them received their first introduction to the Western world in widespread schools set up by the Alliance Israélite Universelle (a Jewish defense organization centred in Paris), which combined Jewish education with the language and values of French civilization. Yet most of these communities remained traditionalist almost to the moment when they were expelled or felt compelled to relocate, since 1948, when the state of Israel was created. The ferment of modernity in all its forms is now being felt in their ranks. In Israel, which has received a large segment of Sefardic Jewry, the attention of these communities has turned to attaining equality with the more advanced Ashkenazim rather than developing some forms of modern Jewish thought.

Other groups that may be described as regional or ethnic include the Bene-Israel, descendants of Jewish settlers in the Bombay region of India, whose deviation in some Halakhic matters from the present Orthodox consensus has raised problems for those among them who have migrated to Israel; the Falashas, the Jewish community of Ethiopia whose development has been almost entirely outside the mainstream described in this article; and the Black Jews of the United States whose place in, and relation to, the rest of the community remains unclear.

CONTEMPORARY JUDAISM

As a result of the Holocaust, Judaism has become a non-European religion; its three major centres, which together include more than three-fourths of world Jewry, are Israel, the Soviet Union, and the United States. Although Jews constitute only a small fraction of the population of the United States, Judaism occupies a role far surpassing its numerical importance and is regarded with Roman Catholicism and Protestantism as one of the major American faiths. Similarly, in the international realm of Western religion, Judaism has been welcomed as a partner able to deal with other major religions as an equal on such issues as anti-Semitism, human rights, and world peace.

The rights and needs of the world Jewish community, including Israel, have triggered deep conflicts with which Judaism has been involved with the Arab and Communist worlds. Friction between Israel and the Arab states has created tension with Islām, while the political stand of Israel and the treatment of the Jewish minority within the Soviet Union have led to open clashes with the Communist leadership. Some of the diatribes and charges that have issued from the Arabs and Communists in this struggle have at times re-evoked older forms of anti-Semitism. In the long-range view, the problems of Judaism and Islām seem more soluble than those of Judaism and such secular ideologies as Communism, for the major religions of the world are increasingly seeking accommodation with each other, as all are confronted by hostile secularist ideologies which have retained their conversionary élan.

Secularization and Jewish identity

Within its own community Jewry is faced with the increasing secularization of Jewish identity in its three major centres, each in its own way. In the United States the open society and the melting pot ideologies of past generations have fostered among many Jews a sense of Jewish identity increasingly devoid of concrete religious, national, or historical content; in the Soviet Union government policy since the 1930s has banned the teaching of Judaism and Jewish culture to the young and has severely discouraged any manifestation of Jewish identity as a sign of the disloyalty of "rootless cosmopolitans" to the Russian state; and in Israel a secular nationalism has taken root, raising questions as to the role that Judaism plays in the identity of the average Israeli. Nonetheless, underneath the external secularization there are signs of a persisting deep Jewish religious fervour, in which the sense of history, community, and personal authenticity figure as the intertwined strands of Jewish religious life, especially as it has been affected by the State of Israel. Some of the rituals of the Jewish tradition, especially the rites of passage at the crucial stages of individual existence, are almost universally observed, except in the Soviet Union; in the United States, for example, more than 80 percent of Jewish children receive some formal religious training. Among Jewish youth there is, in some circles, a quest for tradition. In the United States, Jewish communes have been established that seek new forms of Jewish expression; in Israel, groups such as Mevaqshe Derekh (Seekers of the Way) have tried to bridge secular Israeli culture and Jewish tradition and to maintain traditional Jewish ethical standards even in wartime; in the Soviet Union thousands of young people gather on several occasions of the year to dance and sing and express solidarity in front of the synagogues in Leningrad and Moscow. Still, signs of major weaknesses persist. The rate of intermarriage among Jews in the Diaspora increased, while regular synagogue attendance, at the very highest 20 percent in the United States, remained far below church attendance. Despite their lack of traditional piety, there is a general sense among Jews that they remain Jews not because of the force of anti-Semitism but because of the attractiveness of their tradition and their sense of a common history and destiny.

The enduring existence of the Jews and Judaism

If in 1945, the world Jewish community, decimated and horrified by the Holocaust, felt in danger of disappearing, there appeared to be no such despair in the last quarter of the century when there was an expectation that Jewish communal feeling would remain strong, especially, for many or most Jews, in the light of the existence of Israel. Judaism enjoyed a heightened dignity in the eyes of the world as well, not only as a result of the creation of the State of Israel, but also because of its close relations with other world religions. Although the recurring phenomenon of the alienation of young Jews from their tradition was a troubling prospect, it is no more so than in recent past generations. Along with other major religions, Judaism's most disturbing problem yet to be solved was how to deal with secular ideologies and with the growth of secularism within its own ranks. Thus, looking forward to the last decades of the 20th century, it appeared that Judaism would have to contend with as many problems as other major religions, but that it faced them with no less confidence than these, and with more confidence than it had felt earlier in the century. (Ar.H.)

THE JUDAIC TRADITION

The literature of Judaism

A paradigmatic statement is made in the narrative that begins with Genesis and concludes with Joshua. In the early chapters of Genesis the divine is described as Creator of the natural order, including mankind. In the Eden, Flood, and Tower of Babel stories, man is recognized as rebellious and disobedient. In the patriarchal stories (about Abraham, Isaac, Jacob, and Joseph) a particular family is called out of humankind to restore the thwarted relationship through personal and communal responsiveness. The subsequent history of the community thus formed is recounted so that the divinely sought restoration may be recognized and the nature of the obedient community may be observed: the Egyptian servitude, the going out from Egypt, the revelation of the "teaching," the wandering years, and finally fulfillment through entrance into the "land" (Canaan). The prophetic books (in the Hebrew Bible these include the historical narratives up to the Babylonian Exile—*i.e.,* Joshua, Judges, Samuel, and Kings) continue to deal with the rebellion-obedience tension, interpreting it within the changing historical context and adding new levels of meaning to the fulfillment-redemption motif. (The literature of the Old Testament is treated in the article BIBLICAL LITERATURE.)

Philosophical, mystical, and exegetical additions

It is from this "narrative theology," as it has been recited throughout the centuries, that new formulations of the primal affirmations have been drawn. These have been clothed in a number of vocabularies: philosophical, mystical, ethnic, political, and others. The emphases have been various, the disagreements often profound. No single exposition has exhausted the possibilities of the affirmations or of the relationship between them. Philosophers have expounded them on the highest level of abstraction, using the language of the available philosophic systems. Mystics have enveloped them in the extravagant prose of speculative systems and in simple folktales. Attempts have been made to encompass them in theoretical ethical statements and express them through practical ethical behaviour. Yet, in each instance, the proposed interpretations have had to come to terms with the biblical affirmations and with the particular mode of understanding them required by the spiritual and intellectual demands arising out of the community's experience. The biblical texts, themselves the products of a long period of transmission and embodying more than a single outlook, were subjected to extensive study and interpretation over many centuries and, when required, were translated into other languages. The whole literature continued to provide the basis of further developments, so that any attempt to formulate a statement of the affirmations of Judaism must, however contemporary it seeks to be, give heed to the scope and variety of speculation and formulation in the past.

SOURCES AND SCOPE OF THE TORAH

The concept "Giver of Torah" played a central role in the understanding of God, for it is Torah, or "teaching," that confirms the events recognized by the community as the act of God. In its written form, Torah was considered to be especially present in the first five books of the Bible (the Pentateuch), which therefore came to be called Torah. In addition to this written Torah, or "Law," there were also unwritten laws or customs and interpretations of them, carried down in an oral tradition over many generations, which acquired the status of oral Torah.

The Talmud ("study" or "learning") is the literary culmination of this oral tradition, which, according to the rabbis who created the Talmud, originated at Mt. Sinai as part of the divine revelation vouchsafed to Moses, along with the material recorded in the Pentateuch. In its broadest sense, the Talmud is a set of books consisting of the Mishna ("repeated study"), the Gemara ("completion"), and certain auxiliary materials. The Mishna is a collection of originally oral laws supplementing scriptural laws. The Gemara is a collection of commentaries on and elaborations of the Mishna, which in "the Talmud" is reproduced in juxtaposition to the Gemara. For present-day scholarship, however, Talmud in the precise sense refers only to the materials customarily called Gemara—an Aramaic term prevalent in medieval rabbinic literature that was used by the church censor to replace the term Talmud within the Talmudic discourse in the Basel edition of the Talmud, published 1578–81. This practice continued in all later editions.

Mishna and Gemara

The term Midrash ("exposition" or "investigation"; plural, Midrashim) is also used in two senses. On the one hand, it refers to a mode of biblical interpretation prominent in the Talmudic literature; on the other, it refers to a separate body of commentaries on Scripture using this interpretative mode.

The oral tradition interpreted the written Torah, adapted its precepts to ever-changing political and social circumstances, and supplemented it with new legislation. Thus the oral tradition added a dynamic dimension to the written code, making it a self-regenerating, endless source of guidance, a perpetual process rather than a closed system. The vitality of this tradition is fully demonstrated in the way the ancient laws were adapted after the destruction of the Temple in 70 CE and by the role the Talmud played in the survival of the Jewish people in exile. By the 11th century, Diaspora Jews lived within a Talmudic culture that unified them and that superseded geographical boundaries and language differences. Jewish communities governed themselves according to Talmudic law, and individuals regulated the smallest details of their lives by it.

Central to this vast structure was, of course, the Jewish community's concern to live in accordance with the divine will embodied and expressed in Torah in the widest sense. Scripture, Halakhic and Haggadic Midrash, Mishna, and Gemara were the sources from which the leaders of the communities drew in order to provide both stability and flexibility. The dispersion of the Jews outside Palestine confronted communities and individuals with novel and unexpected situations that had to be dealt with in such a way as to provide continuity while at the same time making it possible to exist with the unprecedented.

Prophecy and religious experience. Torah in the broad sense includes the whole Hebrew Bible, including the prophetic books. In biblical prophecy, God is seen as continuing to be disclosed in the nexus of historical events and as making ethical demands upon the community. According to rabbinic Judaism, this source of Torah—the charismatic person—dried up in the period of Ezra (*i.e.,* about the time of the return from the Babylonian Exile in the 5th century BCE). This opinion may have been a defensive reaction to the luxuriant growth of apocalyptic speculation about the end of the world and the kingdoms of this world, a development that was considered dangerous and unsettling in the period after the Bar Kokhba revolt (132–135 CE). Indeed, there appears to have developed an ongoing suspicion that unrestrained individual experience as the source of Torah was inimical to the welfare of the community. Such an attitude was by no means new. Deuteronomy (13:2–19) had already warned against such "misleaders." The culmination of this attitude is to be found in a Talmudic narrative in which even the *bat qol,* the divine "echo" that announces God's will, is ignored on a particular occasion. Related to this is the reluctance on the part of teachers in the early Christian centuries to point to wonders and miracles in their own time. Far from expressing an ossification of religious experience—the development of the *Siddur* and the Talmudic reports on the devotional life of the rabbis contradict such an interpretation—the attitude seems to be a response to the development of such religious enthusiasm as was exhibited, for example, in the behaviour of the Christian Church in Corinth (I Cor.) and among Gnostic sects and sectarians. Thus, even among the speculative mystics of the Middle Ages, where allegorization of Scripture abounds, the structure of the community and the obligations of the individual are not displaced by the deepening of personal religious life through mystical experience. The decisive instance of this is Joseph Karo (16th century), who was thought to be in touch with a supernal guide, but who was, at the same time, the author of an important codification of Jewish law, the *Shulḥan ʿarukh.*

Anticharismatic trends in post-biblical Judaism

Admittedly, there have been occasions when Torah, even in the wide sense, has been rigidly viewed and applied. In certain historical situations, the dynamic process of rabbinic Judaism has been treated as a static structure. What is of greater significance, however, is the way in which this tendency toward inflexibility has been checked and reversed by the inherent dynamism of the rabbinic tradition.

Modern views of Torah. In modern times—since the end of the 18th century—the traditional position has been challenged both in detail and in principle. The rise of biblical criticism has raised a host of questions about the origins and development of Scripture and thus about the very concept of Torah, in the senses in which it has functioned in Judaism. Naturalistic views of God have required a reinterpretation of Torah in sociological terms as the ideals and sancta (holy things) of the Jewish people. Other and varying positions of many sorts have been and undoubtedly will be forthcoming. What is crucial, however, is the concern of all these positions to retain—with whatever modifications are required—the concept of Torah as one of the central and continuing affirmations of Judaism.

Opposition to the Talmud. Despite the central place of the Talmud in traditional Jewish life and thought, sig-

Anti-Talmudic positions

nificant Jewish groups and individuals have opposed it vigorously. The Karaite sect in Babylonia, beginning in the 8th century, refuted the oral tradition and denounced the Talmud as a rabbinic fabrication. Medieval Jewish mystics declared the Talmud a mere shell covering the concealed meaning of the written Torah, and heretical messianic sects in the 17th and 18th centuries totally rejected it. The decisive blow to Talmudic authority came in the 18th and 19th centuries when the Haskala (the Jewish Enlightenment movement) and its aftermath, Reform Judaism, secularized Jewish life and, in doing so, shattered the Talmudic wall that had surrounded the Jews. Thereafter, modernized Jews usually rejected the Talmud as a medieval anachronism, denouncing it as legalistic, casuistic, devitalized, and unspiritual.

There is also a long-standing anti-Talmudic tradition among Christians. The Talmud was frequently attacked by the church, particularly during the Middle Ages, and accused of falsifying biblical meaning, thus preventing Jews from becoming Christians. The church held that the Talmud contained blasphemous remarks against Jesus and Christianity and that it preached moral and social bias toward non-Jews. On numerous occasions the Talmud was publicly burned, and permanent Talmudic censorship was established.

On the other hand, since the Renaissance there has been a positive response and great interest in rabbinic literature by eminent non-Jewish scholars, writers, and thinkers in the West. As a result, rabbinic ideas, images, and lore, embodied in the Talmud, have permeated Western thought and culture.

Content, style, and form. The Talmud is first and foremost a legal compilation. At the same time it contains materials that encompass virtually the entire scope of subject matter explored in antiquity. Included are topics as diverse as agriculture, architecture, astrology, astronomy, dream interpretation, ethics, fables, folklore, geography, history, legend, magic, mathematics, medicine, metaphysics, natural sciences, proverbs, theology, and theosophy.

This encyclopaedic array is presented in a unique dialectic style that faithfully reflects the spirit of free give-and-take prevalent in the Talmudic academies, where study was focussed upon a Talmudic text. All present participated in an effort to exhaust the meaning and ramifications of the text, debating and arguing together. The mention of a name, situation, or idea often led to the introduction of a story or legend that lightened the mood of a complex argument and carried discussion further.

This text-centred approach profoundly affected the thinking and literary style of the rabbis. Study became synonymous with active interpretation rather than with passive absorption. Thinking was stimulated by textual examination. Even original ideas were expressed in the form of textual interpretations.

Halakha and Haggada; Midrash and Mishna

The subject matter of the oral Torah is classified according to its content into Halakha and Haggada and according to its literary form into Midrash and Mishna. Halakha ("law") deals with the legal, ritual, and doctrinal parts of Scripture, showing how the laws of the written Torah should be applied in life. Haggada ("narrative") expounds on the nonlegal parts of Scripture, illustrating biblical narrative, supplementing its stories, and exploring its ideas. The term Midrash denotes the exegetical method by which the oral tradition interprets and elaborates scriptural text. It refers also to the large collections of Halakhic and Haggadic materials that take the form of a running commentary on the Bible and that were deduced from Scripture by this exegetical method. In short, it also refers to a body of writings. Mishna is the comprehensive compendium that presents the legal content of the oral tradition independently of scriptural text.

Modes of interpretation and thought. Midrash was initially a philogical method of interpreting the literal meaning of biblical texts. In time it developed into a sophisticated interpretive system that reconciled apparent biblical contradictions, established the scriptural basis of new laws, and enriched biblical content with new meaning. Midrashic creativity reached its peak in the schools of Rabbi Ishmael and Akiba, where two different hermeneutic methods were applied. The first was primarily logically oriented, making inferences based upon similarity of content and analogy. The second rested largely upon textual scrutiny, assuming that words and letters that seem superfluous teach something not openly stated in the text.

The Talmud (*i.e.,* the Gemara) quotes abundantly from all Midrashic collections and concurrently uses all rules employed by both the logical and textual schools; moreover, the Talmud's interpretation of Mishna is itself an adaptation of the Midrashic method. The Talmud treats the Mishna in the same way that Midrash treats Scripture. Contradictions are explained through reinterpretation. New problems are solved logically by analogy or textually by careful scrutiny of verbal superfluity.

Midrashic-Talmudic thinking

The strong involvement with hermeneutic exegesis—interpretation according to systematic rules or principles—helped develop the analytic skill and inductive reasoning of the rabbis but inhibited the growth of independent abstract thinking. Bound to a text, they never attempted to formulate their ideas into the type of unified system characteristic of Greek philosophy. Unlike the philosophers, they approached the abstract only by way of the concrete. Events or texts stimulated them to form concepts. These concepts were not defined but, once brought to life, continued to grow and change meaning with usage and in different contexts. This process of conceptual development has been described by some as "organic thinking." Others use this term in a wider sense, pointing out that, although rabbinic concepts are not hierarchically ordered, they have a pattern-like organic coherence. The meaning of each concept is dependent upon the total pattern of concepts, for the idea content of each grows richer as it interweaves with the others.

EARLY COMPILATIONS

Ezra the scribe who, according to the Book of Ezra, reestablished and reformed the Jewish religion in the 5th century BCE, began the "search in the Law . . . to teach in Israel statutes and ordinances."

His work was continued by *soferim* (scribes), who preserved, taught, and interpreted the Bible. They linked the oral tradition to Scripture, transmitting it as a running commentary on the Bible. For almost 300 years they applied the Torah to changing circumstances, making it a living law. They also introduced numerous laws that were designated "words of the *soferim*" by Talmudic sources. By the end of this period, rabbinic Judaism—the religious system constructed by the scribes and rabbis—was strong enough to withstand pressure from without and mature enough to permit internal diversity of opinion.

At the beginning of the 2nd century BCE, a judicial body headed by the *zugot*—pairs of scholars—assumed Halakhic authority. There were five pairs in all, between *c.* 150 and 30 BCE. The first of the *zugot* also introduced the Mishnaic style of transmitting the oral tradition.

The *tannaim*

The making of the Mishna: 2nd–3rd centuries. Hillel and Shammai, the last of the *zugot,* ushered in the period of the *tannaim*—"teachers" of the Mishna—at the end of the 1st century BCE. This era, distinguished by a continuous attempt to consolidate the fragmentary Midrashic and Mishnaic material, culminated in the compilation of the Mishna at the beginning of the 3rd century CE. The work was carried out in the academies of Hillel and Shammai and in others founded later. Most scholars believe that Halakhic collections existed prior to the fall of Jerusalem, in 70 CE. Other compilations were made at Yavne, a Palestinian town near the Mediterranean, as part of the effort to revitalize Judaism after the disaster of 70 CE. By the beginning of the 2nd century there were many such collections. Tradition has it that Rabbi Akiba organized much of this material into separate collections of Midrash, Mishna, and Haggada and introduced the formal divisions in tannaitic literature. His students and other scholars organized new compilations that were studied in the different academies.

After the rebellion of the Jews against Roman rule led by Simeon bar Kokhba in 132–135, when the Sanhedrin (the Jewish supreme court and highest academy) was revived, the Mishnaic compilation adopted by the Sanhedrin president became the official Mishna. The Sanhedrin reached its highest stature under the leadership of Judah ha-Nasi (Judah the Prince, or President); he was also called Rabbi, as the preeminent teacher.

It seems certain that the official Mishna studied during his presidency was the Mishna we know and that he was its editor. Judah aimed to include the entire content of the oral tradition. He drew heavily from the collections of Akiba's pupils but also incorporated material from other compilations, including early ones. Nevertheless, the accumulation was such that selection was necessary. Thus almost no Midrash or Haggada was included. Colleagues and pupils of Judah not only made minor additions to the Mishna but tried to preserve the excluded material, the Baraitot ("Exclusions"), in separate collections. One of these was the Tosefta ("Addition"). Midrashic material was gathered in separate compilations, and later revisions of some of these are still extant. The language of all of the tannaitic literature is the new Hebrew developed during the period of the Second Temple (*c.* 6th century BCE–1st century CE).

The *amoraim*

The making of the Talmuds: 3rd–6th century. The expounders of the Mishna were the *amoraim* ("interpreter"), and the two Talmuds—the Palestinian (or Jerusalem) and the Babylonian—consist of their explanations, discussions, and decisions. Both take the form of a running commentary on the Mishna.

The foundations for these two monumental works were begun by three disciples of Judah ha-Nasi: Joḥanan bar Nappaḥa, Rav (Abba Arika), and Samuel bar Abba, in their academies at Tiberias, in Palestine, and at Sura and Nehardea in Babylonia, respectively. Centres of learning where the Mishna was expounded existed also at Sepphoris, Caesarea, and Lydda in Palestine. In time new academies were established in Babylonia, the best known being those at Pumbedita, Mahoza and Naresh, founded by Judah bar Ezekiel, Rava, and Rav Pappa, respectively. The enrollment of these centres often numbered in the thousands, and students spent many years there. Those who no longer lived on the academy grounds returned twice annually for the *kalla,* a month of study in the spring and fall.

Academies differed in their methods of study. Pumbedita, for example, stressed casuistry, while Sura emphasized breadth of knowledge. Students often moved from one academy to another and even from Palestine to Babylonia or from Babylonia to Palestine. This kept open the channels of communication between the various academies and resulted in the inclusion of much Babylonian material in the Palestinian Talmud, and vice versa.

Despite the overwhelming similarity of the two Talmuds, however, they do differ in some ways. The Palestinian Talmud is written in the Western Aramaic dialect, the Babylonian in the Eastern. The former is invariably shorter, and, not having been subject to final redaction, its discussions are often incomplete. Its explanations tend to remain closer to the literal meaning of the Mishna, preferring textual emendation to casuistic interpretation. Finally, some of the legal concepts in the Babylonian Talmud reflect the influence of Persian law, for Babylonia was under Persian rule at the time.

The main endeavour of the *amoraim* was to thoroughly explain and exhaust the meaning of the Mishna and the Baraitot. Apparent contradictions were reconciled by such means as explaining that conflicting statements referred to different situations or by asserting that they stemmed from the Mishnayot (Mishnas) of different *tannaim.* The same techniques were used when amoraic statements contradicted the Mishna. These discussions took place for hundreds of years, and their content was passed on from generation to generation, until the compilation of the Talmud.

The compilation of the Palestinian and Babylonian Talmuds

The portion of the Palestinian Talmud dealing with the three Bavot ("gates")—*i.e.,* the first three tractates of the fourth order of the Mishna (for orders and tractates, see *Talmudic and Midrashic literature,* below)—was compiled in Caesarea in the middle of the 4th century and is distinguished from the rest by its brevity and terminology. The remainder was completed in Tiberias some 50 years later. It seems likely that its compilation was a rescue operation designed to preserve as much of the Halakhic material collected in Palestinian academies as possible, for by that time the deterioration of the political situation had forced most Palestinian scholars to emigrate to Babylonia.

The Babylonian Talmud was compiled up to the 6th century. Some scholars suggest that the organization of the Talmud began early and that successive generations of *amoraim* added layer upon layer to previously arranged material. Others suggest that at the beginning a stratum called Gemara, consisting only of Halakhic decisions or short comments, was set forth. Still others theorize that no overall arrangement of Talmudic material was made until the end of the 4th century.

The statement in the tractate *Bava metzia* that "Rabina and Rav Ashi were the end of instruction" is most often understood as referring to the final redaction of the Talmud. Since at least two generations of scholars following Rav Ashi (died 427) are mentioned in the Talmud, most scholars suggest that "Rabina" refers to Rabina bar Huna (died 499) and that the redaction was a slow process lasting about 75 years to the end of the 5th century.

According to the tradition of the *geonim*—the heads of the academies at Sura and Pumbedita from the 6th to the 11th centuries—the Babylonian Talmud was completed by the 6th-century *savoraim* ("expositors"). But the extent of their contribution is not precisely known. Some attribute to them only short additions. Others credit them with creating the terminology linking the phases of Talmudic discussions. According to another view, they added comments and often decided between conflicting opinions. The proponents of the so-called Gemara theory noted above ascribe to them the entire dialectic portion of Talmudic discourse.

TALMUDIC AND MIDRASHIC LITERATURE

Mishna. The Mishna is divided into six orders (*sedarim*), each order into tractates (*massekhtot*), and each tractate into chapters (*peraqim*). The six orders are *Zeraʿim, Moʿed, Nashim, Neziqin, Qodashim,* and *Ṭohorot.*

Zeraʿim, Moʿed, and *Nashim*

1. *Zeraʿim* ("Seeds") consists of 11 tractates: *Berakhot, Pea, Demai, Kilayim, Sheviʿit, Terumot, Maʿaserot, Maʿaser sheni, Ḥalla, ʿOrla,* and *Bikkurim.* Except for *Berakhot* ("Blessings"), which treats of daily prayers and grace, this order deals with laws related to agriculture in Palestine. It includes prohibitions against mixtures in plants (hybridization), legislation relating to the sabbatical year (when land lies fallow and debts are remitted), and regulations concerning the portions of harvest given to the poor, the Levites, and the priests.

2. *Moʿed* ("Season" or "Festival") consists of 12 tractates: *Shabbat, ʿEruvin, Pesaḥim, Sheqalim, Yoma, Sukka, Betza, Rosh Hashana, Taʿanit, Megilla, Moʿed qaṭan,* and *Ḥagiga.* This order deals with ceremonies, rituals, observances, and prohibitions relating to special days of the year, including the Sabbath, holidays, and fast days. Since the half-shekel Temple contribution was collected on specified days, tractate *Sheqalim,* regarding this practice, is included.

3. *Nashim* ("Women") consists of seven tractates: *Yevamot, Ketubbot, Nedarim, Nazir, Soṭa, Giṭṭin,* and *Qiddushin.* This order deals with laws concerning betrothal, marriage, sexual and financial relations between husband and wife, adultery, and divorce. Since Nazirite (ascetic) and other vows may affect marital relations, *Nedarim* ("Vows") and *Nazir* ("Nazirite") are included here.

Neziqin, Qodashim, and *Ṭohorot*

4. *Neziqin* ("Damages") consists of 10 tractates, the first three of which were originally considered one (the *Bavot*): *Bava qamma, Bava metzia, Bava batra, Sanhedrin, Makkot, Shevuʿot, ʿEduyyot, ʿAvoda zara, Avot,* and *Horayot.* This order deals with civil and criminal law concerning damages, theft, labour relations, usury, real estate, partnerships, tenant relations, inheritance, court composition, jurisdiction and testimony, erroneous decisions of the Sanhedrin, and capital and other physical punishments. Since idolatry, in the literal sense of worship or veneration of material images, is punishable by death, *ʿAvoda zara* ("Idolatry") is included. *Avot* ("Fathers"), commonly called "Ethics of the Fathers" in English, seems to have

been included to teach a moral way of life that precludes the transgression of law.

5. *Qodashim* ("Sacred Things") consists of 11 tractates: *Zevaḥim, Menaḥot, Ḥullin, Bekhorot, ʿArakhin, Temura, Keretot, Meʿila, Tamid, Middot,* and *Qinnim.* This order incorporates some of the oldest Mishnaic portions. It treats of the Temple and includes regulations concerning sacrifices, offerings, and donations. It also contains a detailed description of the Temple complex.

6. *Ṭohorot* ("Purifications") consists of 12 tractates: *Kelim, Ohalot, Negaʿim, Para, Ṭohorot, Miqwaʾot, Nidda, Makhshirin, Zavim, Ṭevul yom, Yadayim,* and *ʿUqtzin.* This order deals with laws governing the ritual impurity of vessels, dwellings, foods, and persons, and with purification processes.

Tosefta. The Tosefta ("Addition") closely resembles the Mishna in content and order. In its present form it at times supplements the Mishna, at other times comments on it, and often also opposes it. There is no Tosefta on the tractates *Avot, Tamid, Middot,* and *Qinnim.* The Talmud quotes from many other collections of Mishnaiot and Baraitot: some are attributed to *tannaim,* and predate the established Mishna; and others, to *amoraim.* The original material is lost.

Incompleteness of the Talmuds

Talmud (Gemara). Although the entire Mishna was studied at the Palestinian and Babylonian academies, the Palestinian Talmud (Gemara) covers only the first four orders (except chapters 21–24 of *Shabbat* and chapter 3 of *Makkot*) and the first three chapters of *Nidda* in the sixth order. Most scholars agree that the Palestinian Talmud was never completed to the fifth and sixth orders of the Mishna and that the missing parts of the other orders were lost. A manuscript of chapter 3 of *Makkot* was, in fact, found and was published in 1946.

The Babylonian Talmud does not cover orders *Zeraʿim* (except *Berakhot*) and *Ṭohorot* (except *Nidda*) and tractates *Tamid* (except chapters 1,2,4), *Sheqalim, Middot, Qinnim, Avot,* and *ʿEduyyot.* Scholars concur that the Talmud for these parts was never completed, possibly because their content was not relevant in Babylonia.

Midrashim. *Halakhic.* Halakhic Midrashim are exegetic commentaries on the legal content of Exodus, Leviticus, Numbers, and Deuteronomy. The five extant collections are *Mekhilta,* on Exodus; *Mekhilta deRabbi Shimʿon ben Yoḥai,* on Exodus; *Sifra,* on Leviticus; *Sifre,* on Numbers and Deuteronomy; *Sifre zuṭa,* on Numbers. (*Mekhilta* means "measure," a norm or rule; *Sifra,* plural *Sifre,* means "writing" or "book.") Critical analysis reveals that *Mekhilta* and *Sifre* on Numbers differ from the others in terminology and method. Most scholars agree that these two originated in the school of Ishmael and the others in that of Akiba. In their present form they also include later additions. Mention should also be made of *Midrash tannaim* on Deuteronomy, consisting of fragments recovered from the Yemenite anthology *Midrash ha-gadol.*

Haggadic. Haggadic Midrashim originated with the weekly synagogue readings and their accompanying explanations. Although Haggadic collections existed in tannaitic times, extant collections date from the 4th–11th centuries. Midrashic compilations were not authoritatively edited and tend to be coincidental and fragmentary.

Biblical and nonbiblical Haggadot

Most notable among biblical collections is *Midrash rabba* ("Great Midrash"), a composite of commentaries on the Pentateuch and five Megillot (Song of Songs, Ruth, Ecclesiastes, Esther, Lamentations) differing in nature and age. Its oldest portion, the 5th-century *Genesis rabba,* is largely a verse-by-verse commentary, while the 6th-century *Leviticus rabba* consists of homilies and *Lamentations rabba* (end of 6th century) is mainly narrative. The remaining portions of *Midrash rabba* were compiled at later dates.

The *Tanḥuma* (after the late-4th-century Palestinian *amora* Tanḥuma bar Abba), of which two versions are extant, is another important Pentateuchal Midrash. Additional Midrashic compilations include those to the books of Samuel, Psalms, and Proverbs. Mention should also be made of *Pesiqta* ("Section" or "Cycles") *deRab Kahana* (after a Babylonian *amora*) and *Pesiqta rabbati* ("The Great Cycle"), consisting of homilies on the Torah (Pentateuch) readings that occur on festivals and special Sabbaths.

Haggadic compilations independent of biblical text include *Avot deRabbi Natan, Tanna deve Eliyyahu, Pirqe* ("Chapters") *deRabbi Eliezer,* and tractates *Derekh eretz* ("Correct Conduct"). These primarily deal with ethics, moral teachings, and biblical narrative.

Among the medieval anthologies are the *Yalquṭ* ("Compilation") *Shimoni* (13th century), *Yalquṭ ha-makhiri* (14th century), and *ʿEn Yaʿaqov* ("Eye of Jacob," 16th century). The two most important modern Haggadic anthologies are those of Wilhelm Bacher and Louis Ginzberg.

Codes. The Talmud's dialectic style and organization are not those of a code of laws. Accordingly, codification efforts began shortly after the Talmud's completion. The first known attempt was *Halakhot pesuqot* ("Decided Laws"), ascribed to Yehudai Gaon (8th century). *Halakhot gedolot* ("Great Laws"), by Simeon Kiyyara, followed 100 years later. Both summarize Talmudic Halakhic material, omitting dialectics but preserving Talmudic order and language. The later *geonim* concentrated on particular subjects, such as divorce or vows, introducing the monographic style of codification.

Codification literature gained impetus by the beginning of the 11th century. During the next centuries many compilations appeared in Europe and North Africa. The most notable, following Talmudic order, were the *Hilkhot Harif,* by Isaac Alfasi (11th century), and *Hilkhot Harosh,* by Asher ben Jehiel (13th–14th centuries). Though modelled after *Halakhot gedolot,* the *Hilkhot Harif* encompasses only laws applicable after the destruction of the Temple but includes more particulars. The *Hilkhot Harosh* closely follows Alfasi's code but often also includes the reasoning underlying decisions.

The *Mishne Torah, Sefer ha-ṭurim,* and *Shulḥan ʿarukh*

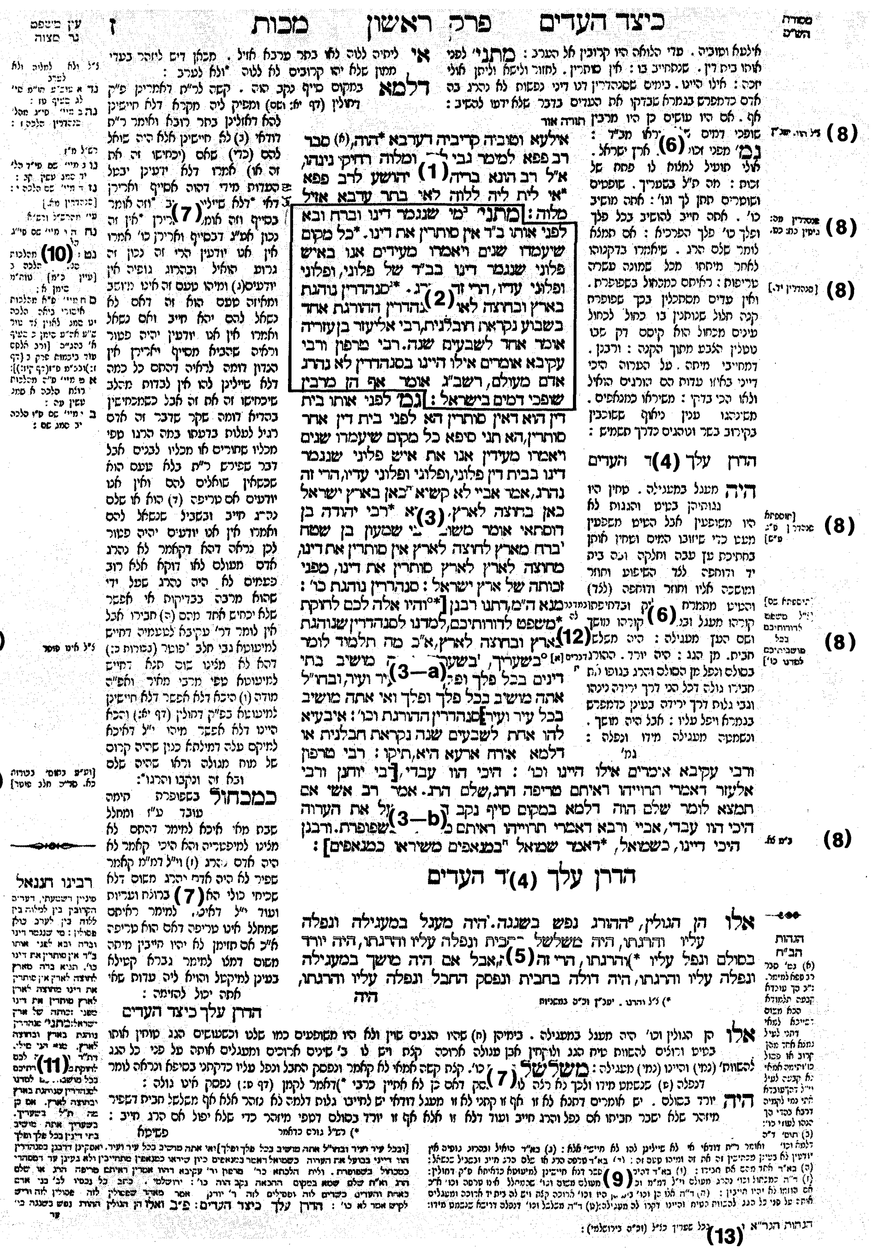
The most important of the topically arranged codifications were: the *Mishne Torah, Sefer ha-ṭurim,* and *Shulḥan ʿarukh.* (1) The *Mishne Torah* ("The Torah Reviewed") by Maimonides (12th century), is a monumental work, original in plan, language, and order; it encompasses all religious subject matter under 14 headings and includes theosophy, theology, and religion. (2) The *Sefer ha-ṭurim* ("Book of Rows," or " Parts"), by Jacob ben Asher (14th century), the son of Asher ben Jehiel, introduced new groupings, dividing subject matter into four major categories (*ṭurim*) reminiscent of the Mishnaic orders; it includes only laws applicable after the destruction of the Temple. (3) The *Shulḥan ʿarukh* ("The Prepared Table") by Joseph Karo (16th century), the last of the great codifiers, is structured after the *Sefer ha-ṭurim,* but presents the Sefardic (Middle Eastern and North African) rather than the Ashkenazic (Franco-German and eastern European) tradition, with decisions largely following those of Alfasi, Maimonides, and Rabbi Asher. When the 16th-century Ashkenazic codifier Moses Isserles added his notes, this became the standard Halakhic code for all Jewry.

Commentaries. The interpretive literature on the Talmud began with the rise of academies in Europe and North Africa. The earliest known European commentary, though ascribed to Gershom ben Judah (10th–11th centuries), is actually an eclectic compilation of notes recorded by students of the Mayence (Mainz) Academy. Compilations of this kind, known as *quntresim* ("notebooks"), also developed in other academies. Their content was masterfully reshaped and reformulated in the renowned 11th-century commentary of Rashi (acronym of *Ra*bbi *Sh*lomo *Yitzḥaqi*), in which difficulties likely to be encountered by students are anticipated and detail after detail is clarified until a synthesized, comprehensible whole emerges.

The commentaries of Ḥananel ben Ḥushiel and Nissim ben Jacob ben Nissim, the first to appear in North Africa (11th century), are introductory in nature. They summarize the content of Talmudic discussions, assuming that details will be understood once the general idea becomes comprehensible. This style was later followed by the Spanish school, including Joseph ibn Migash and Maimonides. However, as Rashi's work became known, it displaced all other commentaries. (Note its predominant role in the sample page of Talmud.)

Tosafot, novellae, and *responsa*

A new phase in Talmudic literature was initiated by Rashi's grandchildren, Rabbis Isaac, Samuel, and Jacob, the sons of Meir, who established the school of *tosafot.*



Sample page (7a) of the tractate *Makkot* (of the fourth order, *Neziqin*) of the Vilna edition of the Babylonian Talmud, first printed in 1880–86. It discusses the fate of a man who was convicted and escaped and how he is to be judged. Code numbers, a box surrounding the Mishna, and brackets indicating the extent of comments (3-a) and (3-b) have been superimposed onto the original page. (1) End of the Gemara to the previous Mishna. (2) Mishna. (3) Gemara. (3-a) Halakhic Midrash supporting the Mishna. (3-b) Three short comments from Palestine, Sura, and Pumbedita. (4) Mark indicating the end of the chapter. (5) Mishna of the next chapter. (6) Commentary of Rashi (1040–1105). (7) *Tosafot,* discussing special points in the Gemara and Rashi. (8) Cross-reference notes to other Talmudic and Rabbinic sources and textual variants. (9) Notes by Joel Sirkes (1561–1640). (10) References to the codes of Maimonides, Moses of Concy, the *Tur,* and the Shulḥan ʿarukh. (11) Commentary of Hananel of North Africa, early 11th century. (12) References to scripture. (13) Notes by Elijah Gaon of Vilna (1720–97).

By courtesy of the Library of the Jewish Theological Seminary of America, New York; Frank J. Darmstaedter

(These medieval "additions" are not to be confused with the tannaitic Tosefta discussed above.) Reviving Talmudic dialectic, they treated the Talmud in the same way that it had treated the Mishna. They linked apparently unrelated statements from different Talmudic discourses and pointed out the fine distinctions between seemingly interdependent statements. This dialectic style was soon adopted in all European academies. Even the writings of Ravad (Abraham ben David), Zerahiah ha-Levi, and Yeshaya deTrani, three of the most original Talmudists (12th century), reflect the impact of Tosafist dialectic.

The works of Meir Abulafia and Menaḥem Meiri, although of the North African genre, include a strong dialectic element. In Spain such dialectic works were known as *ḥiddushim* or *novellae* (since they sought "new insights"), the most famous being those written by four generations (13th–14th centuries) of teacher and pupil: Ramban (Naḥmanides, or Moses ben Naḥman), Rashba (Solomon ben Adret), Ritba (Yomtov ben Abraham), and Ran (Nissim ben Reuben Gerondi).

A major role in establishing Talmudic authority was also played by the *responsa* literature, replies (*responsa*) to legal and religious questions. Beginning in the 7th century, when the Babylonian *geonim* responded in writing to questions concerning the Talmud, it developed into a branch of Talmudic literature that continued to the present. Then, as now, Talmudic authorities were approached for explanations and decisions. Among the *geonim* the best known were Sherira (10th century) and his son Hai. In the Middle Ages the most important were Alfasi, Ibn Migash (Joseph ibn Migash), Maimonides, Ravad (Abraham ben David of Posquières), Ramban, Rashba, Rosh (Asher ben Jehiel), Ran, and Ribash (Isaac ben Sheshet Perfet).

Writing and printing of the Talmuds. Study in the academies was always oral; hence the question of when the Mishna and Talmud were first committed to writing has been the subject of much discussion. According to some scholars, the process of writing began with Judah ha-Nasi. Others attribute it to the *savoraim.*

The Palestinian Talmud was first printed in Venice (1523–24). All later editions followed this one. Printing of the Babylonian Talmud was begun in Spain about 1482, and there have been more than 100 different editions since. The oldest extant full edition appeared in Venice (1520–23). This became the prototype for later printings, setting the type of page and pagination (a total of close to 5,500 folios). The standard edition was printed in Vilna beginning in 1886. It carries many commentaries and commentaries upon commentaries. In the sample page reproduced here, the Mishna and the Gemara are placed in the centre column of the page and are printed in the heavy type. The commentary of Rashi is always located in the inner column of the page and the *tosafot* in the outer column. Other commentaries and references to legal codes and to scriptural verses surround the major commentaries, in smaller type. Talmudic citations are made by tractate name, folio number, and side of the folio (a or b).

NONLEGAL SUBJECT MATTER

God, Torah, and Israel

Main religious doctrines. While the Talmudic rabbis never formally systematized their beliefs, their underlying religious concepts are clearly reflected in their decisions, ideas, and attitudes. Preeminent in rabbinic thinking were the concepts of God, Torah, and Israel.

God. The rabbinic God was primarily the biblical God who acted in history, the creator and source of life who was experienced through the senses rather than intellect. In reaction to sectarian teachings (*i.e.,* Gnosticism and early Christianity), however, the rabbis stressed God's universality, absolute unity, and direct involvement with the world. His immanence and transcendence (being present in and beyond the universe) were emphasized, and biblical anthropomorphisms (ascribing human attributes to God) were explained metaphorically. The rabbis also stressed an intimacy into the relationship between God and man. God became the father to whom each individual could turn in direct prayer for his needs. To the names YHWH and Elohim, which traditionally were identified with God's mercy and judgment, respectively, the rabbis added new terms reflecting his other attributes—*e.g.,* Shekhina ("Presence"), representing his omnipresence, or immanence; and Maqom ("Place"), his transcendence.

Torah. Torah, in the Talmudic sense, refers to all religious and ethical teachings handed down by tradition. According to the rabbis, God created the Torah long before the world. It contained the eternal divine formula for the world's future workings and thus the answers to all problems for all times and all people. God himself is depicted as studying the Torah, for even he cannot make decisions concerning the world that contradict it.

Israel. The people Israel, according to the rabbis, were chosen by God to be the guardian of his Torah, and, just as God chose Israel, Israel chose God. Thus, the concept of Israel as a nation bound together by an irrevocable commitment to bring the Torah to the world, and bearing corporate responsibility for this mission, was formed. No Jew can free himself from this commitment, but anyone accepting it, regardless of race, becomes a full-fledged Jew with obligations binding him and his descendants.

With this in mind, the rabbis repeatedly emphasized the importance of studying Torah. They pointed out that the Torah is not a declaration of religious beliefs. Rather it is a statement of a discipline regulating each detail of life. Any transgression of this discipline hampers the divine plan of establishing God's way of life in this world.

Sanctification of everyday life

Worship. The intensive rabbinic religious involvement led to the growth of a new concept of worship. While in the Bible worship was usually centred in the sanctuary of the Temple in Jerusalem, the rabbis, particularly after the destruction of the Second Temple (70 CE), attempted to sanctify all of life. Thus, they said that one must bless God upon arising in the morning, before dressing, before and after meals, and in all ordinary daily actions or routines. Each move in life should be an act of worship glorifying God's name.

Messianic kingdom. In rabbinic thinking the establishment of God's kingdom was tied to the Messiah, who was to be a descendant of King David, wise, just, a great scholar, a moral leader, and courageous king. He would redeem the Jews from exile and reestablish their independence in the land of Israel. With this the world would be ushered into a new era of righteousness and universal peace. The rabbis referred to this era as "the world to come," portraying it as an immense academy in which the righteous would study Torah without interruption. They refrained from describing it further, saying that human language and fantasy are inadequate to its wonders.

The nature of the Messiah and the time of his arrival raised much speculation. Following the defeat of Bar Kokhba, leader of the revolt against Roman rule (135 CE), the Messiah's coming, in rabbinic thought, faded into the mysterious and distant future, and descriptions concerning his personality assumed supernatural overtones.

For a fuller discussion of major religious doctrines, see below, *Basic beliefs and doctrines.*

Doctrine of man. The fate of man, his achievements and failures, his being and nothingness, occupy an important place in Talmudic literature. The rabbis' concept of man was a universal one. While they assumed that Jews are bound by greater religious duties than others, they considered all men equal, all created in the image of God. "Therefore, but a single man was created . . . That none should say to his fellow, 'My father was greater than thy father'" (tractate *Sanhedrin*).

The world, according to the Talmud, was created for the sake of man, and it is incumbent upon him to keep it in order. His responsibility begins at home. Man must care for his health, marry, build a family, provide for and educate children, honour parents, friends, and elders. He also carries social responsibilities and has to be part of the community. He must learn a trade and work so that he does not become a burden to the community.

The *yetzer ha-ṭov* and *yetzer ha-ra*

The uniqueness of man in this world, likened by the Talmud to the uniqueness of God in the universe, lies in his freedom of choice. Nature follows its laws and angels their missions, but man is his own master. In contrast to St. Paul's doctrine that the original sin of Adam made sin an integral part of human nature, the rabbis considered man a wondrous and harmonious being. The duality of his nature was explained by the existence of a good and bad impulse, personified by two angels, *yetzer ha-ṭov* (the good inclination) and *yetzer ha-raʿ* (the evil inclination), which enter each man after birth. It is the duty of man to overcome his evil inclination, and it is for this that he is rewarded. Moreover, since there is corporate responsibility, not only is the sinner punished but the community at large also suffers. Here again, however, man is his own master. He can reverse the course of sin and punishment by repentance. Although repentance may be accompanied by formal and ceremonial acts, such as fasting, its basic principle is the renunciation of the sin and the wholehearted decision not to repeat it. When a man transgresses against God, his sin is forgiven by repentance alone, but, when he transgresses against his fellow man, he must make good his wrongdoing as well as repent.

Medicine and science. The Talmud devoted considerable attention to the maintenance of good health, regarding it a religious duty. A keen understanding of the importance of hygiene in preventing illness was reflected in an emphasis upon bodily cleanliness. The rabbis also stressed the necessity for moderation in eating and drinking and the importance of a proper diet. The Talmud prescribed remedies for illnesses and mentioned surgical techniques, such as cesarean section.

Religious concerns surrounding the calendar, prohibitions against planting seeds of different kinds together, dietary laws, and Sabbath-walking limits resulted in an intense rabbinical interest in astronomy, zoology, mathematics, and geometry.

The challenge of pagan myths, legend, and folklore

Legend and folklore. Side by side with the Midrashic Haggada, which was the outgrowth of Bible exegesis and developed in the academies, the Talmuds and Midrashic collections contain a large quantity of Haggadic material with mythological rudiments, allusions to pagan beliefs and customs, and folkloristic elements of a world strange to the rabbis. Folktales and legends, animal lore, and adventure narratives, containing pagan ideas and beliefs, that were told by their Gentile neighbours were no doubt a major attraction to the common Jews, especially those in the countryside (the *ʿam ha-aretz,* or "people of the land"). The rabbis realized the great danger involved in this situation and developed their own folk material. They adopted the dramatic and artistic parts of these stories but rejected the unwanted elements, replacing them with their own ideas. Thus the animals and birds in fables quote the Bible and discuss it in the same manner that the rabbis do.

Ancient mythology seems to have been well known and liked by the Jewish masses. Again, in order to fight its influence, the rabbis reworked its content in their own spirit. They retained the mythological suspense—the sea tries to drown the earth—but there is no mythological struggle between equal powers; angels try to prevent the creation of man, but they do not possess titanic power. All are subdued by the command of God. Thus, the rabbis transformed the ancient myths into dramatic evidence against polytheism. (See also below, *Jewish myth and legend.*)

Astrology, magic, and divination. Astrology was a recognized science in the ancient world. The rabbis could not reject it entirely, and some concluded that the power of the stars is confined to Gentiles. Others made it part of God's order, saying that stars influence this world in the same way that climate influences plants. The rabbis strenuously objected to omens and other forms of divination because they considered them magic. Dreams were considered by some rabbis as meaningless, while others saw in them an element of prophecy.

The rabbis believed in the efficacy of magic but strenuously objected to its practice. They permitted only magic that had been proved effective in healing. They also permitted the use of incantations for the purpose of counteracting the hold of magic. Because of their supposedly protective nature, the use of amulets was also countenanced.

Demons (evil spirits) and the evil eye

The existence of a demonic kingdom was accepted by the rabbis without question. Evil spirits are invisible and fill the nether world. They avoid sunlight and concentrate in waters and deserted places. They also mingle with people, trouble them, and help them. They have passions and are born and die like people. However, they also have some of the traits and powers of angels. The evil eye was considered as dangerous as evil spirits. It was thought that for mysterious reasons some people have the power to injure others by looking at them and that it is generally jealousy that triggers this effect. The rabbis, however, repeatedly emphasized that all of these strange powers are under the divine government and, moreover, that they cannot hurt the pious.

TALMUDIC LAW AND JURISPRUDENCE

Sacral basis of Talmudic law

Unlike the Romans, who considered ritual law (*fas*) God-given and social law (*lex*) man-made, the rabbis believed all Jewish law to be of divine origin. Thus, for example, unfairness in labour relations was considered a religious sin and caring for the sick a religious obligation. Though familiar with the concept of natural law (ethical principles inherent in the nature of things and apprehensible through human reason), the rabbis objected to making nature the basis of law. Even rabbinic ordinances were regarded as having validity only because the authority of the rabbis is sanctioned by the Torah.

Methods of arriving at legal principle and decisions. Ancient Halakha knew no controversy. The earliest controversy dates to the pre-tannaitic *zugot.* Hillel and Shammai differed on significant issues, and, with the rise of their schools, Halakhic uniformity began to crumble. Halakha became a scholastic discipline that developed in academic rather than judicial settings, more and more issues remaining unresolved. Over 300 controversies between the schools of Hillel and Shammai (called the House of Hillel and the House of Shammai, respectively) are reported in Talmudic sources. As time passed, disputes proliferated even more and were considered legitimate provided they conformed to the rule of Halakhic discipline.

No attempt was made to restore Halakhic uniformity until the beginning of the 2nd century CE. Controversies were sometimes resolved by citing old traditions, by establishing precedents, or, when the sages could convene, by vote taking.

At Yavne, Gamaliel II, the president of the revived Sanhedrin (*c.* 80–*c.* 115 CE), attempted to suppress diversity of opinion, but failed. The right to differ was already established. Moreover, in the Halakhic collection compiled at Yavne (tractate *ʿEduyyot*), the views of individual scholars were preserved. The sages at Yavne, however, did take a major step toward restoring Halakhic consistency by upholding the generally more lenient views of the House of Hillel over those of the House of Shammai, thus establishing the Hillelite tradition as the main trend of rabbinic Judaism.

Guidelines for settling Halakhic controversies

The principle that differing opinions should be recorded was followed by Judah ha-Nasi in his Mishna. Modern scholars differ as to whether he meant to compile a code of law or merely a Halakhic collection. The *amoraim,* however, accepted his Mishna as the definitive code and introduced a set of guidelines according to which disputes were decided. Thus, for example, collective (". . . the sages said") and individual opinions stated anonymously were taken as law; Akiba's decisions were upheld over those of his colleagues. Similar guidelines developed also with regard to amoraic controversies.

With the completion of the Talmud, a new phase in Halakhic development began. Not only were there two different Talmuds and a large Haggadic literature but even within each of the Talmuds diversified opinions were reported. The *geonim* laid down rules governing the use of this enormous literature for lawmaking. They designated the Babylonian Talmud the highest authority, taking the Palestinian Talmud into consideration only when it did not disagree with the Babylonian or when the latter expressed no opinion on a subject. They also deprived the Haggadic literature of Halakhic authority and set guidelines for the precedence of opinion among *amoraim.* These geonic rules served as the basis of all future codifications.

After the geonic period two methods of decision making were applied. The first of these relied primarily upon the authoritative codes. The Mediterranean rabbis, for example, made the code of Maimonides the source of all of their lawmaking. The second method relied on the original Talmudic sources for decision making. This method was applied by the Tosafists and their followers, who, though they consulted the older codes, did not accept them as the final authorities. The *responsa* literature represents a synthesis of these two methods. Although it makes use of codes as the main source of law, its decisions are always accompanied by a discussion and analysis of earlier relevant literature. This approach has been used by rabbis to the present day.

In addition to the above, in particular instances throughout the ages rabbinic authorities promulgated ordinances (*taqqanot*) and edicts (*gezerot*). These were made in response to pressing needs of time and circumstance, and this form of lawmaking was most frequently used by rabbinic synods in the Middle Ages.

The Great and Lesser Sanhedrins

Administration of justice. *Courts.* A comprehensive judicial system is described in Talmudic sources. The highest court was the Great Sanhedrin. It consisted of 71 members and convened daily in one of the Temple halls. It was the highest legal and religious authority in the country and had exclusive jurisdiction over matters of a national and public nature. It also functioned as the court of appeals, dealing with cases that were not resolved by the lower courts.

Next in line of judicial authority was the Lesser Sanhedrin. Each town with a population of 120 or more had a court of this kind. These courts each consisted of 23 members and dealt with cases involving capital punishment.

The members of the Sanhedrins had to be ordained, pious, mature in age, sound in mind and body, of wide knowledge, and of pure Jewish descent. Persons who were too old or who had never had children were ineligible, for it was thought that they might not be merciful.

The lower courts dealt with all remaining cases. Each consisted of three members and convened on Mondays and Thursdays. In cases involving a penalty the three judges had to be ordained, but in those involving ordinary monetary litigation ordination was not required. In the latter type of case, concerned parties were allowed the alternative of setting up ad hoc arbitration bodies.

Rules of evidence. Jewish law was extremely strict regarding evidence acceptable in court. In cases entailing physical punishment, no circumstantial evidence, confession, or self-incrimination was recognized. The testimony of two eyewitnesses who confronted the defendant was required. In monetary cases documentary evidence and, at times, oaths were acceptable. Any mental or moral defects or self-interest in the case disqualified witnesses. Relatives could not serve as judges or witnesses.

Trial procedure. Jewish law knows of no lawyers. After the facts were presented, the court investigated, deliberated, and made its decision by voting. Both sides had to be treated equally, even to the point of seeing to it that neither should be dressed more richly than the other. Each side could be heard only in the presence of the other.

In the trial procedure of capital cases, there was a clear tendency toward bias in favour of the defendant. Thus, only the judges could argue for conviction, but all present could argue for acquittal. The most junior judges voted first so that they would not be unduly influenced by their seniors. A majority of one was sufficient for acquittal, but a majority of two was necessary for conviction. A verdict of acquittal could be reached on the same day but one of conviction only on the following day. When the court erred, only its convictions, and not its acquittals, were reversed.

Death penalties

Criminal law. In Jewish law, ritual and nonritual transgressions were crimes punishable by court. Each of the 36 most severe transgressions (*e.g.,* adultery, sodomy, idolatry, sorcery, or murder) carried one of four types of death penalty (stoning, burning, beheading, and strangling). Rabbinic law, however, tended to minimize the practice of capital punishment. The rigorous cross-examination of witnesses and the warning of impending punishment that the transgressor had to receive immediately before committing his crime made it almost impossible to reach a death verdict.

If despite all of this a death verdict was reached, every legal effort was made to allow for a last-minute reversal. Execution was expedited and carried out in the most humane manner possible, the accused being given an opiate before dying. To show their compassion the judges fasted on the day of execution. According to tradition the death penalty was abolished 40 years before the destruction of the Temple, when the Great Sanhedrin was exiled from the Temple complex.

The punishment for 207 other transgressions (*e.g.,* perjury, some forms of incest, the eating of forbidden foods) was flagellation. Here, too, the rabbis tended to be lenient. As in capital cases, a rigorous cross-examination and a warning were required. The maximum number of stripes administered was 39. Prior to flagellation the transgressor was examined medically to determine the number of stripes he could withstand.

Side by side with the above penalties, the courts also inflicted *makkat mardut* (disciplinary stripes) and excommunication in cases where regular flagellation could not legally be applied. These two punishments were generally used in Babylonia, where ordained courts did not exist. It should be mentioned also that the Mishna includes a few obscure references to a form of imprisonment used instead of capital punishment.

Civil and social law. Although the rabbis considered both ritual and nonritual law sacred, they demonstrated great independence in supplementing the relatively brief relevant scriptural comments and regulations with a comprehensive system of civil and social law. In response to variations in social and economic circumstances, certain differences in Palestinian and Babylonian Talmudic law emerged. The Babylonian rabbis, for example, recognized the law of the state as binding in monetary matters, while the Palestinian rabbis did not. In general, however, Jewish civil law developed relatively autonomously. In instances where the rabbis did adopt alien legal concepts, they elaborated upon them until they could be fully integrated into the spirit and structure of Jewish law.

Social welfare, torts, family law, financial law

The following are some of the areas covered: (1) Social welfare: a comprehensive social welfare system was worked out, including obligations to provide for children, educate them, and train them for a profession. Regulations of charity, medical assistance, and burial of the dead were established. (2) Torts: included were all damages caused by a person directly or indirectly via his property. The main aim was to compensate for damages. Consequently, no torts were classified as criminal. Even "an eye for an eye" was interpreted to mean financial compensation. (3) Family law: included were regulations concerning marriage and divorce procedures and the innovation of the *ketubba* (marriage contract), which spells out the mutual obligations of husband and wife in the areas of finance, medical care, clothing, housework, sexual relations, and child care. According to biblical law, the right to inherit belongs to sons first. To protect the rights of wives and daughters, rabbinic law obligated the sons to maintain the widows and unmarried daughters. (4) Financial law: except for Gen. 23:9 ff., Jer. 32:10, and Ruth 4:8, Scripture makes no reference to transaction procedures. The growth of finances, industry, and land estates led the rabbis to develop laws concerning contracts, partnerships, and legal arrangements to circumvent the biblical prohibition against usury. A series of modes of transaction effecting the transfer and acquisition of property evolved. Labour relations, rents, and leases were also carefully regulated.

THE TALMUD TODAY

Role in the Jewish community. With the rebirth of a Jewish national state (since 1948) and the concomitant revival of Jewish culture, the Talmud has achieved renewed importance. Orthodox Jewry has always focussed upon its study and has believed it to be the absolute Halakhic authority. This belief has now become even further intensified. While rabbinic courts in Israel have jurisdiction only in the area of family life, it has become one of the aims of religious (Orthodox) Jewry there to establish Talmudic law as the general law of the state.

It should also be noted that, aside from the special case of Israel, the legal system described above has continued to function down to the present day in Jewish communities all over the world. The jurisdiction of rabbinic courts is voluntarily accepted by Orthodox Jews. These courts continue to exert authority, especially in the areas of family and dietary law, the synagogue, and the organization of charity and social activity.

Conservative Jewry, too, has always been committed to rabbinic tradition. It has, however, conceptualized this tradition as an evolutionary process in which Halakha changes to meet the challenge of new conditions. Professional scholarship was considered crucial for understanding the furthering of this process. More recently, however, as a result of revived nationalism, new emphasis has been put upon lay education. Thus, a network of day schools and higher institutions of learning in which rabbinic tradition occupies a major role in the curriculum has been established. Scores of young Conservative Jews now search in the Talmud for answers to crucial problems, such as abortion and civil violence.

Classical (19th-century) Reform Judaism not only dissassociated itself from the Talmud but negated it. More recently, however, Reform leaders have been inclined to reestablish some measure of ritual practice and rabbinic climate. Thus, it is now not unusual to find them stating their decisions in the form of *responsa* and using the rabbinic style of argument and even the casuistic type of Talmudic dialectic (*pilpul*) to justify their religious practices.

Talmudic scholarship. Although Talmudic scholarship continues to be advanced by individuals in a number of countries, its two main centres are in Israel and the United States. The Israeli centre has tended to focus upon research of a critical nature. Like Bible criticism, this work is divided between source criticism (*i.e.,* discovering the different sources, their dates, and the methods by which Talmudic literature was formed) and textual criticism (*i.e.,* establishing the correct text and reading). Research is also being done on Haggadic concepts and thinking, Talmudic law, and Halakhic development.

Talmudic scholarship in the United States has tended to be more philosophically and historically oriented. There has been great interest in the development of Halakha and in folklore and custom. Essential work has been done and continues to be done in the areas of source criticism. A work unique in scope and method is S. Lieberman's commentary on the Tosefta. (H.Z.D./L.H.S./Ed.)

Basic beliefs and doctrines

The Judaic and the human

Judaism is not and cannot be viewed as an abstract intellectual system, although some of its affirmations may be couched in such terms. It affirms divine sovereignty disclosed in creation (nature) and in history, without necessarily insisting upon—but at the same time not rejecting—metaphysical speculation about the divine (see below *Jewish philosophy*). It insists that the community has been confronted by the divine not as abstraction but as person, with whom the community and its members enter into relationship. It is—as the concept Torah indicates—a program of human action, rooted in this personal confrontation. Further, the response of this particular people to its encounter with God is viewed as significant for all mankind. The community is called upon to express its loyalty to God and the Covenant by exhibiting solidarity within its corporate life on every level—including every aspect of human behaviour, from the most public to the most private. Thus, even Jewish worship is communal celebration of the meetings with God in history and in nature. Yet the particular existence of the Covenant people is not thought of as contradicting but rather as enhancing human solidarity. This people, together with all men, is called upon to create political, economic, and social forms that will affirm divine sovereignty—embody it in communal existence. This task is carried out in the belief not that man will succeed solely by his own efforts in these endeavours but that these sought-after human relationships have both their source and their goal in God—who assures their actualization. Within the sphere of his existence in the community, each Jew is called upon to realize the Covenant in his personal intention and behaviour.

In considering the basic affirmations of Judaism from this point of view it is best to allow indigenous formulations rather than systematic statements borrowed from other traditions to govern the presentation.

GOD

An early statement of basic beliefs and doctrines emerged in the liturgy of the synagogue some time during the last pre-Christian and first Christian centuries, although there is evidence that such formulations were not absent from the Temple cult that came to an end in the year 70 CE. A section of the *Siddur* (order of worship, or prayerbook) that has as its focus the recitation of a series of biblical passages (Deut. 6:4–9; Deut. 11:13–21; Num. 15:37–41) takes its name from the first of these, Shema ("Hear"): "Hear, O Israel! the Lord is our God, the Lord alone" (or ". . . the Lord our God, the Lord is one"). In the Shema—often regarded as the Jewish confession of faith, or creed—the biblical material and accompanying benedictions are arranged to provide a unified statement about God and his relationship to the world and Israel, as well as Israel's obligations toward and response to God. In this statement, God, who is the Creator of the universe and who has chosen Israel in love ("Blessed art thou, O Lord, who has chosen thy people Israel in love"), expressed by the giving of Torah, is declared to be "one"; his love is to be reciprocated by men who lovingly obey Torah and whose obedience is rewarded and rebellion punished. The goal of this obedience is God's "redemption" of Israel, a role foreshadowed by his action in bringing Israel out of Egypt.

Divine love and justice

Unity and uniqueness. At the centre of this liturgical formulation of belief is the concept of the divine unity. In its original setting, it may have served as the theological statement of the reform under Josiah, king of Judah, in the 7th century BCE when worship was centred exclusively in Jerusalem, and all other cultic centres were rejected, so that the existence of one shrine only was understood as affirming one deity. The idea, however, acquired further meaning. It was understood toward the end of the pre-Christian era to proclaim—over against dualistic religious formulations in the Greco-Roman world—the unity of divine love and divine justice, as expressed in the divine names YHWH and Elohim, respectively. A further expansion of this affirmation is found in the first two benedictions of this liturgical section, which together proclaim that the God who is the Creator of the universe and the God who is Israel's ruler and lawgiver are one and the same—as over against religious positions that insisted that the Creator God and the lawgiver God were separate and even inimical. Subsequently, this affirmation was developed in philosophical and mystical terms by both medieval and modern thinkers.

Creativity. As has been noted, this "creed," or "confession of faith," underscores in the first benediction the relation of God to the world as that of Creator to creation. "Blessed art thou, O Lord our God, King of the Universe, who forms light and creates darkness, who makes peace and creates all things." It adds the assertion that his activity is not in the past but is ongoing and continuous, for "he makes new continually, each day, the work of creation"; thus, unlike the deity of the Stoic world view, he remains actively present in nature. This "creed" is concerned as well to come to terms with the ever-present problem of evil. Paraphrasing Isa. 45:7, "I form the light and create darkness; I make peace, and create evil," it changes the last word to "all" (or "all things") rather than "evil." The change was clearly made to avoid the implication that God is the source of moral evil. Judaism, however, did not flinch from confronting the problem of pain and suffering in the world and affirming the paradox of suffering and divine sovereignty, of pain and divine providence, refusing to accept the concept of a partial God—a God that is Lord over only the harmonious and pleasant aspects of reality.

Divine Providence

Activity in the world. The second and the third benedictions deal with divine activity within the realm of history and human life. God is teacher of men through the giving of instruction (Torah; see above); he acts in the life of mankind in historical events; he has chosen a particular people—Israel—in love to witness to his presence and his desire for a perfected society; he will, as redeemer, enable man to experience that perfection. These activities, together with creation itself, are understood to express divine compassion and kindness as well as justice (judgment), recognizing the sometimes paradoxical relation between them. Taken together, they disclose Divine Providence—God's continual activity in the world. The constant renewal of creation (nature) is itself an act of compassion overriding strict justice and affording mankind further opportunity to fulfill the divinely appointed obligation. Yet the basically moral nature of God is asserted in the second of the biblical passages that form the core of this liturgical statement (Deut. 11:13–21). Here, in the language of its agricultural setting, the community is promised reward for obedience and punishment for disobedience. The intention of the passage is clear: obedience is rewarded by the preservation of order, so that the community and its members find wholeness in life; while disobedience—rebellion against divine sovereignty—shatters order, so that the community is overwhelmed by adversity. The passage of time has made the original language unsatisfactory (promising rain, crops, and fat cattle), but the basic principle remains, affirming that, however difficult it is to recognize the fact, there is a divine law and judge. Support for this affirmation is drawn from the third biblical passage (Num. 15:37–41), which explains that the fringes the

Israelites are commanded to wear on the corners of their garments are reminders to observe the commandments of God, who brought forth Israel from Egyptian bondage. The theme of divine redemption is elaborated in the concluding benediction to point toward a future in which the as yet fragmentary rule of God will be brought to completion: "Blessed is his name whose glorious kingdom is for ever and ever."

Otherness and nearness. Within this complex of ideas, other themes are interwoven. In the concept of the divine Creator there is a somewhat impersonal or remote quality—of a power above and apart from the world—which is underscored by such expressions as the trifold declaration of God's holiness, or divine otherness, in Isaiah 6:3: "Holy, holy, holy is the Lord of hosts . . ." The development of surrogate divine names for biblical usage, as well as the substitution of Adonai ("my Lord") for the tetragrammaton (YHWH) in the reading of the Bible itself, suggests an acute awareness of the otherness of God. Yet this has as its countertheme the affirmation of divine nearness. In the biblical narrative it is God himself who is the directly active participant in events, an idea that is emphasized in the liturgical narrative (Haggada) recited during the Passover meal (seder): "and the Lord brought us forth out of Egypt—not by an angel, and not by a seraph, and not by a messenger . . . ." The surrogate divine name *Shekhina,* the Present One, is derived from a Hebrew root meaning "to dwell," again calling attention to divine nearness ("presence"). The relationship between these two affirmations, otherness and nearness, is expressed in a Midrashic statement, "in every place that divine awesome majesty is mentioned in Scripture, divine abasement is spoken of, too."

The divine "thou"

Closely connected with these ideas is that of divine personhood, most particularly disclosed in the use of the pronoun "thou" in direct address to God. The community and the individual, confronted by the Creator, teacher, redeemer, addresses the divine as living person, not as theological abstraction. The basic liturgical form, the *berakha* ("blessing"), is usually couched in the second person singular: "Blessed art thou . . . ." This relationship, through which remoteness is overcome and presentness is established, illuminates creation, Torah, and redemption, for it reveals the meaning of love. From it flow the various possibilities of expressing the divine–human relationship in personal, intimate language. Sometimes, especially in mystical thought, such language becomes extravagant, foreshadowed by such vivid biblical metaphors as the husband–wife relation in Hosea; the "adoption" motif in Ezek. 16; and the firstborn-son relation (Ex. 4:22). Nonetheless, although terms of personal intimacy are used widely to express Israel's and man's relationship with God, such usage is restrained by the accompanying sense of divine otherness. This is to be seen in the liturgical "blessings," where, following the direct address to God, in which the second person singular pronoun is used, the verbs, with great regularity, are in the third person singular, thus providing the requisite tension between nearness and otherness, between the impersonal and the personal.

Modern views of God. The Judaic affirmations about God have not always been given the same emphasis nor have they been understood in the same way. This was true in the Middle Ages, among both philosophers and mystics, as well as in modern times. In the 19th century, western European Jewish thinkers attempted to express and transform these affirmations in terms of German Idealist philosophy: more recently, philosophical Naturalism was offered as the suitable content of Judaism, while still retaining the traditional God language. The meaningfulness of the whole body of such affirmations, moreover, has been called into question by the philosophical schools of Logical Positivism and Linguistic Analysis. Most recently, the destruction of 6,000,000 Jews during the Nazi period has raised the issue of the validity of such concepts as God's presence in history, divine redemption, the covenant, and the chosen people. In every case, however, it is with the structure of ideas here noted that these challenges must deal.

ISRAEL (THE JEWISH PEOPLE)

Choice and covenant. The concluding phrase of the second benediction of the liturgical section referred to above reads: "who has chosen thy people Israel in love." Here the basis of the relationship between God and Israel set forth in the biblical narrative is clearly and succinctly stated: the choice of this people was determined by no other factor than divine love. The patriarchal narratives, beginning with the 12th chapter of Genesis, presuppose the choice, which is set forth explicitly in Deut. 7:6–8 in the New Jewish Version:

> For you are a people consecrated to the Lord your God: of all the peoples on earth the Lord your God chose you to be His treasured people. It is not because you are the most numerous of peoples that the Lord set His heart on you and chose you—indeed you are the smallest of peoples; but it was because the Lord loved you and kept the oath He made with your fathers that the Lord freed you with a mighty hand and rescued you from the house of bondage, from the power of Pharaoh king of Egypt.

Later rabbinic traditions on occasion sought to base the choice upon some special merit of Israel, and the medieval poet and theologian Judah ha-Levi suggested that the openness to divine influence originally present in Adam continued only within the people of Israel.

The binding agreement between God and Israel

However understood, the background of this choice is the recurring disobedience of mankind narrated in Genesis 2–11. Abraham and his descendants are singled out not merely as the object of the divine blessing but also as its channel to all mankind. The choice, however, demands a reciprocal response from Abraham and his lineage. That response is obedience, as exemplified in the first instance by Abraham's readiness to leave his "native land" and "father's house" (Gen. 12:1). This twofold relationship was formalized in a mutually binding agreement, a covenant between the two parties. The covenant, thought by some modern biblical scholars to reflect the form of ancient suzerainty treaties, indicates (as in the Ten Utterances) the source of Israel's obligation—the acts of God in history—and the specific requirements those acts impose. The formalization of this relationship was accomplished by certain cultic acts that may, according to some contemporary scholars, have been reenacted on a regular basis at various sacred sites in the land, eventually being centralized in Jerusalem. The content of the covenantal obligations thus formalized was Torah. Israel was bound in obedience, and Israel's failure to obey provided the occasions for the prophetic messages. The prophets, as spokesmen for God, called the community to renewed obedience, threatened and promised disaster if such was not forthcoming, and—recalling the source of the choice in divine love—sought to explain its persistence even when, strictly understood, the covenant should have been repudiated by God.

The choice of Israel has its concrete expression in the requirements of the precepts (*mitzwot,* singular *mitzwa*) that are part of Torah. The blessing recited before the public reading of the pentateuchal portions on Sabbath, festivals, holy days, fasts, and certain weekdays refers to God as "He who chose us from among all the peoples and gave us His Torah," thus emphasizing the intimate relationship between the elective and revelatory aspects of God.

Priest and witness to the nations

Israel's role was not defined solely in terms of its own obedience to the commandments. As noted, Abraham and his descendants were seen as the means by which the estrangement of disobedient mankind from God was to be overcome. Torah was the formative principle underlying the community's fulfillment of this obligation. Israel was to be "a kingdom of priests and a holy nation" (Ex. 19:6) functioning within mankind and for its sake. This task is enunciated with particular earnestness in the writings of the prophets. In Isa. 43–44, Israel is declared to be God's witness and servant who is to bring the knowledge of God to the nations. In chapter 42 of the same book Israel is declared to be a "covenant of the people, a light to the nations, to open the blind eyes, to bring out the prisoners from the prisons, and them that sit in darkness out of the prison house" (42:6–7). This double motif of a chosen people and witness to the nations, joined to that of the righteous king, developed in the biblical and postbiblical periods into messianism in its several varieties (see below *Eschatology*).

The intimate relation between choice, covenant, and Torah determined the modality of Israel's existence. Religious faith, far from being restricted to or encapsulated in the cult, found its expression in the totality of communal and individual life. The obligation of the people was to be the true community, in which the relationship between its members was open, in which social distance was repudiated, and in which response to the divine will expressed in Torah was called for equally from all. One of the important recurring themes of the prophetic movement was the adamant rejection of any tendency to limit divine sovereignty to the partial area of "religion," understood as the realm of the priesthood and cult. Subsequent developments continued this theme, although it appeared in a number of other forms. Pharisaic Judaism and its continuation, rabbinic Judaism, down to modern times has resolutely held to the idea of the all-pervasive functioning of Torah, so that however the various Jewish communities over the centuries may have failed to fulfill the ideal, the self-image of the people was that of "holy community."

Israel and the nations. The double motif of "treasured people" and "witness" was not without its tensions as it functioned in ongoing history. Tensions are especially visible in the period following the return from the Babylonian Exile at the end of the 6th and beginning of the 5th centuries BCE. It is, however, doubtful whether the use of such terms as nationalism, particularism, or exclusivism (as opposed to universalism) are of any great help in understanding the situation. Emphasis has, for example, been laid upon Ezra 9:2 and 10:2, in which the reestablished community is commanded to give up wives taken from "the peoples of the land." This is taken as indication of the narrow, exclusivistic, nationalistic nature of Judaism, without reference to the situation in which a harassed contingent of returned exiles sought to maintain itself in a territory surrounded by politically unfriendly if not hostile neighbours. Nor does this represent racialism, since foreigners were admitted to the Jewish community, and in the following centuries some groups engaged in extensive missionary activities, appealing to the individuals of the nations surrounding them to join themselves to the God of Israel, who was the one true God, the Creator of the heavens and earth.

The welding of exclusivism and universalism

A more balanced view recognizes that within the Jewish community religious universalism was affirmed at the same time and by the same people who understood the nature of Jewish existence in politically particularistic (*i.e.*, nationalistic) terms. To neglect either side is to distort the picture. In no case was the universalism disengaged from the reality of the existing community, even when it was expressed in terms of the ultimate fulfillment of the divine purpose, the restoration of the true covenantal relationship between God and all mankind. Nor was political particularism, even under circumstances of great provocation and resentment, misanthropic. The most satisfactory figure in describing the situation of the restored community, and one that continues to be useful in dealing with later episodes, is that of the human heartbeat, made up of two functions, the systole, or contraction, and the diastole, or expansion. There have been several periods of contraction and of expansion throughout the history of Judaism. The emphasis within the abiding tension has been determined by the historical situation in which the community has found itself. To generalize in one direction or the other is fatal to an understanding of the history and faith of the "holy community."

The people and the land. Closely related to the concept of Israel as the chosen, or Covenant, people is the role of the land of Israel. In the patriarchal stories, settlement in Canaan is an integral part of the fulfillment, from the divine side, of the Covenant. The goal of the Israelites escaping from Egypt is the same land, and entry into it is understood in the same fashion. The return from the Babylonian Exile, too, is seen in the same light. As there was the choice of a people, so was there the choice of a land—and for much the same reason. It was to provide the setting in which the community could come into being as it carried out the divine commandments. This choice of the land contrasts significantly with the predominant ideas of other peoples in the ancient world, in which the deity or divinities were usually bound to a particular parcel of ground outside of which they lost their effectiveness or reality. Though some such concepts may very well have crept into Israelite thought during the period of the kings (from Saul to Jehoiachin), the crisis of the Babylonian Exile was met by a renewal of the affirmation that the God of Israel was, as Lord of all the earth, free from territorial restraint, although He had chosen a particular territory for this chosen people. Here again the twofold nature of Jewish thought becomes apparent, and both sides are to be affirmed or the view is distorted. Following the two revolts against Rome (66–73 CE and 132–135 CE), the Jews of the ever-widening dispersion continued, as they had before these disasters, to cherish the land. Once again it became the symbol of fulfillment, so that return to it was looked upon as an integral part of messianic restoration.

The land as ideal and reality

The liturgical patterns of the community, insofar as they were concerned with natural phenomena (*e.g.*, planting, rainfall, harvest, and the annual cycle) rather than historical events, were based on geography, topography, and agricultural practices of the land, viewed as paradigmatic. Although Jews continued to live in the land, yet for most in the distant dispersion it was idealized and viewed primarily in eschatological terms—at the end of days, in the world to come. The 11th-century poet Judah ha-Levi not only longed for it in verse but also gave it a significant role in his theological interpretation of Judaism and eventually sought to return to it from his native Spain. It was not, however, until the 19th century that the land began to play a role other than the goal of pilgrimage or of occasional settlement by pietists and mystics. At the end of that century the power of the utopian concept was released in eastern Europe in a cultural renaissance that focussed, in part, on a return to the land and, in western and central Europe, in a political movement coloured by nationalist motifs in European thought. The coming together of these two gave rise to Zionism. The political movement reflected a dissatisfaction with the view of the Jews as merely a body or organization of religious believers—like the Christian churches—an interpretation that had become dominant following the political emancipation of the Jews in the period after Napoleon. The political emphasis of Zionism aroused considerable opposition from those Jews who were convinced of the necessity of a churchly definition of Judaism parallel to the Roman Catholic and Protestant communions. While this conflict erupted in bitter debate during the first half of the 20th century, the events of the Nazi period in Europe brought it to a close, except for some sporadic renewals on the part of numerically insignificant groups. For the most part, although there are few satisfactory formulations—theological or secular—there is a working consensus that acknowledges a significant role to the land and recognizes that a churchly definition of the Jewish community, while strategically acceptable in some situations, does not do justice to history and is not theologically sound if it suggests that Judaism merely consists of abstract doctrines and dogmas. Some Jews, however, argue that whatever the past has been, the future of the Jewish community is with those movements in the modern world that cut across or transcend the particularity Zionism represents.

Modern views of the people Israel. The nature of the people Israel and of the land of Israel has been variously interpreted in the history of Jewish thought. In modern times, some interpretations have been deeply influenced by contemporary political and social discussions in the general community. Thus, for example, Zionist theoreticians were influenced by concepts of political nationalism on the one hand and by socialist ideas on the other. Further, the challenge to traditional theological concepts in the 19th century raised issues about the meaning of the choice of Israel, and Jewish thinkers borrowed from romantic nationalism such ideas as the "genius" of the people.

Modern reinterpretations or rejections of chosenness

Most recently, attempts have been made to approach the question sociologically, dismissing the theological mode as unhelpful. The concept of the chosen people is then understood to indicate a specific role deliberately undertaken by the Jewish people and similar to that espoused by other groups (*e.g.,* "Manifest Destiny" by the American people). The establishment of the State of Israel has motivated some thinkers to call for a repudiation of the idea, in keeping with the position that normal existence for the Jews requires the dismissal of such concepts. Although only a small minority of Jewish thinkers espouse this position, the doctrine of the choice is not without its theological difficulties even for those who continue to affirm it.

MAN

The image of God. In Gen. 1:26, 27; 5:1; and 9:6 two terms occur, "image" and "likeness," that seem to indicate clearly the biblical understanding of man's essential nature: he is created in the image and likeness of God. Yet the texts in which they are used are not entirely unambiguous; the idea they point to does not appear elsewhere in Scriptures; and the concept is skirted cautiously in the rabbinic interpretations. What the image and likeness of God or the divine image refer to in the biblical text is not made explicit, and, in the light of the psychosomatic unity of man that dominates the biblical concepts, it is not possible to escape entirely from the implication of "bodily" similarity. What the terms meant in their context at the time and whether they reflect mythological usages taken over from other Middle Eastern thought is a question that is by no means answered. Evidence of the problematic nature of the concept is found in rabbinic Judaism. Akiba (2nd century CE) ignored the usages in Gen. 1 and 5 and emphasized 9:6, understanding it to mean, contrary to the usual interpretation, "after an image, God made man," that is, in the Platonic sense of a heavenly archetype. He did not wish to allow any resemblance between God and any created being. Other interpretations sought to avoid the difficulty by rendering *elohim* (a plural form) not as "God" but as "divine beings" (*i.e.,* angels: "God created man after the image of divine beings [*elohim*]").

The earthly–spiritual creature. In those parts of the Jewish community of antiquity that were deeply influenced by Greek philosophical ideas, a dualistic interpretation of man was offered. Here the divine likeness suggested is that of the immortal, intellectual soul as contrasted to the body. Still other thinkers, both ancient and modern, have understood that likeness to be ethical, with particular emphasis placed on freedom of the will. What is evident is that no doctrine of man can be erected on the basis of these several verses alone, but that a broader view must be taken, in which they are assimilated. A careful examination of the biblical material, particularly the words *nefesh, neshama,* and *ruaḥ,* which are often too broadly translated as "soul" and "spirit," indicates that these must not be understood as referring to the psychical side of a psychophysical pair. A man did not possess a *nefesh* but rather was a *nefesh,* as Gen. 2:7 says: "*wayehi ha-adam le-nefesh ḥayya*" (". . . and the man became a living being"). Man was, for most of the biblical writers, what has been called "a unit of vital power," not a dual creature separable into two distinct parts of unequal importance and value. While this understanding of the nature of man dominated biblical thought, in apocalyptic literature (2nd century BCE–2nd century CE) the term *nefesh* began to be viewed as a separable psychical entity with existence apart from body. Although this was not entirely divorced from the unitary biblical view, nonetheless a functional body–soul dualism was present in such literature. In the Alexandrian version of Hellenistic Judaism the orientation toward Greek philosophy, particularly the Platonic view of the soul imprisoned in the flesh, led to a clear-cut dualism with a negative attitude toward the body. Rabbinic thought remained closer to the biblical position, at least in its understanding of man as a psychosomatic unit, although the temporary separation of the components after death was an accepted position.

"Soul," "spirit," or "life"

The biblical view of man as an inseparable psychosomatic unit meant that death was understood to be his dissolution. Yet, although man ceased to be, this dissolution was not utter extinction. Some of the power that functioned in the unit may have continued to exist, but it was not to be understood any longer as life. The existence of the dead in *sheol,* the netherworld, was not living but the shadow or echo of living. For most of the biblical writers this existence was without experience, either of God or of anything else; it was unrelated to events. To call it immortality is to empty that term of any vital significance. However, this concept of *sheol,* along with belief in the possibility of occasional miraculous restorations of dead individuals to life, and perhaps even the idea of the revival of the people of Israel from the "death" of exile, provided a foothold for the development of belief in the resurrection of the dead body at some time in the future. The stimulus for this may have come from ancient Iranian religion, in which the dualistic cosmic struggle is eventually won by life through the resurrection of the dead. This idea began to appear in sketchy form in postexilic writings (Isa. 26:19; Dan. 12:2). In this view there is life only in the psychosomatic unit now restored. This restoration was bound up with the eschatological hope of Israel (see *Eschatology,* below) and was limited to the righteous. In subsequent apocalyptic literature a sharper distinction between body and soul was entertained, and the latter was conceived of as existing separately in a disembodied state after death. Although at this point the doctrine of the resurrection of the body was not put aside, nonetheless, the direction of thinking changed. The shades of *sheol* were now thought of as souls, and real personal survival—with continuity between life on earth and in *sheol*—was posited. Now Greek ideas, with their individualistic bent, began to have influence, so that the idea of resurrection that was in some way related to a final historical consummation, began to recede. True life after death was now seen as release from the bondage of the body, so that in place of, or alongside of, the afterlife of physical resurrection was set the afterlife of the immortal soul.

Sheol, resurrection, and immortality

It was not the status of the soul, however, that concerned either the biblical or the rabbinic thinkers. What emerged from the latter's discussions of the biblical themes was an emphasis on the ethical import of man's composite makeup. Man is in a state of tension or equilibrium between the two foci of creation, the "heavenly" and the "earthly." He necessarily participates in both, and, as such, is the one responsible creature who can truly serve his Creator, for he alone, having both sides of creation in him, may choose between them. It is the ability to make an ethical choice that is the distinguishing mark of man. This ability is not derived from the "heavenly" side but resides in the double basis of man's existence. It is important to recognize this as something other than a body–soul dualism in which the soul is the source of good and the body the basis of evil. Such an attitude, however, did appear in some rabbinic material and was often affirmed in medieval philosophical and mystical speculations and by some of the later moralists. These are genuine variations and developments of the biblical-rabbinic ideas and may not be dismissed as aberrations. They represent authentic attempts to come to terms with other currents of thought and with the problems and uncertainties inherent within the earlier materials themselves.

The ethically bound creature. Mankind is then viewed, however this position is arrived at, as ethically involved. The first 11 chapters of Genesis are posited upon this responsibility, for the implicit assumption of the prepatriarchal stories is man's ability to choose between obedience and disobedience. Rabbinic Judaism, taking up the covenant-making episode between God and Noah (Gen. 9:8–17), developed it as the basis of mankind's ethical obligation. All men, not merely Israel, were engaged in a covenant relationship with God, which was spelled out in explicit precepts—variously enumerated as six, seven, or even 10 and occasionally as many as 30—that reflect general humanitarian behaviour and are intended to assure the maintenance of the natural order by the establishment of a proper human society. The Covenant with Israel was meant to bring into being a community that would advance the development of this society through its own obedience and witness.

The good and evil impulses

Man's nature, viewed ethically, was explained in rabbinic Judaism not only as a tension between the "heavenly" and "earthly" components but also as a tension between two "impulses." Here again, fragmentary and allusive biblical materials were developed into more comprehensive statements. The biblical word *yetzer* means "plan," that which is formed in man's mind. In the two occurrences of the word in Genesis (6:5; 8:21), the plan or formation of man's mind is described as *raʿ*, perhaps "evil" in the moral sense or maybe no more than "disorderly," "confused," "undisciplined." The other biblical occurrences do not have this modifier. Nonetheless, the Aramaic translations (Targumim) invariably denominated it as *bisha* ("wicked") wherever it occurred. Rabbinic literature created a technical term *yetzer ha-raʿ* ("the evil impulse") to denote the source within man of his disobedience, and, subsequently, the counterterm *yetzer ha-ṭov* ("the good impulse") to indicate man's obedience. These more clearly suggest the ethical quality of man's duality, while their opposition and conflict point to man's freedom and the ethical choices he makes. Indeed, it is primarily within the realm of the ethical that Judaism posits freedom, recognizing the bound, or determined, quality of much of his existence (*e.g.,* his natural environment or physiological makeup).

It is this ethically free creature who stands within the covenant relationship and who may choose to be obedient or disobedient. Sin, then, is ultimately deliberate disobedience or rebellion against the divine sovereign. This is more easily observed in relation to Israel, for it is here that the central concern of Judaism is most evident and the subject discussed in greatest detail. It should be noted, however, that since, according to Judaic tradition, all mankind stands within a covenant relation to God and is commanded to be moral and just, essentially the same choice is made universally. In technical language, the acceptance of divine sovereignty by the people of Israel and by the individual within that community is called "receiving the yoke of the kingship." This involves intellectual commitment to a basic belief, as expressed by the Deuteronomic proclamation: "Hear, O Israel, the Lord, our God, the Lord is one!" At the same time it imposes obligations in terms of communal and individual behaviour. These two responses are understood to be inextricably bound together, so that rejection of the divine sovereign is manifest as denial of God both intellectually and practically. It amounts to "breaking the yoke of the kingship." In more specific terms, sin is sometimes summed up under three major, interrelated headings: idolatry, murder, and illicit sexual behaviour, each and all of which explicitly and implicitly involve rebellion, for they involve activities that deny—if not God's existence—his commanding relationship and the requirement of man's response. Such behaviour destroys the community and sets individual against individual, thus thwarting the ultimate purpose of God, the perfected human society.

Teshuva ("turning")

If, however, man is free to choose rebellion and to suffer its consequences, he is also able to turn back to God and to become reconciled with him. The Bible—most particularly the prophetic writings—is filled with this idea, although the term *teshuva* ("turning") came into use only in rabbinic sources. Basically it grows out of the covenant and God's unwillingness—despite man's failures—to break off his relationship. In rabbinic thought it is apparently assumed that even the direst warnings of utter disaster and rejection imply the possibility of turning back to God, motivated by remorse and the desire for restoration. Divine readiness and human openness are the two sides of the process of reconciliation. What was expressed in prophetic literature in the immediate historicopolitical situation was stated in the synagogal liturgy in connection with pentateuchal and prophetic lessons and the homilies developed from them. Thus, the divine invitation was constantly being offered. Man was called upon to atone for his rebellion by positive action that repudiated his failure. He was summoned to reconstitute wholeness in his individual life and community in society.

Historically viewed, Jewish existence, as it developed under rabbinic leadership, following after the two disastrous rebellions against Rome, was an attempt to reconstitute a community of faith expressed in worship and in an ordered society that would enable the individual to live a hallowed life of response to the divine will. Although this plan was not spelled out in detail, it was probably understood to be the paradigm for the eventual reconstruction of humanity.

Medieval and modern views of man. The Jewish view of man is certainly less clearly articulated than its affirmations concerning God. Nonetheless, it is evident that its central concern was ethical. The question of how man as individual and community was to behave was the focus of interest. Yet it is also clear that metaphysical concerns, however rudimentary in the beginning, were included in the developing discussion. Medieval philosophers sought an accommodation between the doctrine of the resurrection of the body and the concept of the immortality of the soul. The greatest of them, Moses Maimonides (1135–1204), propounded an extremely subtle position that equated immortality with the cleaving of the human intellect to the active intellect of the universe, thus limiting it to philosophic adepts. In the modern period, the impact of various philosophical and psychological schools has further fragmented the situation so that little or no consensus is evident, although resurrection or immortality language is still used even when its content is uncertain. But alongside this lack of agreement, the view that man is to be understood, however else, as a creature who makes free ethical choices for which he is responsible remains—although variously articulated—the basic affirmation of Judaism about man.

ETHICS AND SOCIETY

The imitation of God

The ethical emphasis of Judaism. Jewish affirmations about God and man intersect in the concept of Torah as the ordering of human existence in the direction of the divine. Man, however else understood, is an ethically responsible creature responsive to the presence of God in nature and in history. Although that responsiveness is expressed on many levels, it is within the horizontal relationship of man to man that it is most explicitly called for. The pentateuchal legislation sets down, albeit within the limitations of the structures of the ancient Middle East, the patterns of interpersonal relations. The prophetic messages are deeply concerned with these demands and see the disregard of them as the source of social and individual disorder. No segment of society, even the most exalted, is free of ethical obligation. Indeed, the transformation of prophetism from its earlier form as ecstaticism and soothsaying is seen in the ethical confrontation of David by Nathan ("Thou art the man") for seducing Bathsheba and arranging to have her husband killed (II Sam. 12). What is particularly striking is the affirmation that God is not only the source of ethical obligation but is himself the paradigm of it. In the so-called Code of Holiness (Lev. 19), it is imitation of divine holiness that is offered as the basis of human behaviour in the ethical sphere as well as the cultic-ceremonial. Concern for the economically vulnerable members of the community; obligations toward neighbours, hired labourers, and the physically handicapped; interfamilial relationships; and attitudes toward strangers (*i.e.,* non-Israelites) were all motivated by the basic injunction, "You shall be holy, for I, the Lord your God, am Holy." Acceptable human behaviour is, therefore, "walking in all His ways" (Deut. 11:22). The dialectic relation between God and man in the literary prophets also exhibits divine righteousness and divine compassion as patterns to be emulated in the life of the community.

This theme, *imitatio Dei* ("imitation of God"), as developed in rabbinic Judaism, is expressed succinctly in a comment on the verse from Deuteronomy quoted above. In response to the question of how it is possible to walk "in all His ways," the reply is made (*Sifre* Deut. 85a): "As He is merciful and gracious, so be you merciful and gracious. As He is righteous so be you righteous. As He is holy, strive to be holy." Indeed even more daringly, God is described as clothing the naked, nursing the sick, comforting the mourners, burying the dead, so that man may recognize his own obligations.

Interpenetration of communal and individual ethics. What stands out in the entire development of Jewish ethical formulations is the constant interpenetration of communal and individual obligations and concerns. Although in the Book of Ezekiel (see especially chapter 18) emphasis is laid on individual responsibility, "the person who sins shall die," in contrast to the more widespread statement of communal involvement, "visiting the sins of the fathers upon the children" (*e.g.,* Ex. 20:5), these two aspects of ethical conduct are never entirely distinguished in Judaism. The just society requires the just man, and the just man functions within the just society. The concrete expression of ethical requirements in legal precepts took place with both ends in view, so that the process of beginning the holy community and the formation of the *ḥasid* ("pious"), the man of steadfast devotion to God, were concomitant processes. The relationship between the two is, of course, often mediated by the historical situation, so that in some periods one or the other moves to the centre of practical interest. In particular, the end of the Judaean state (70–135 CE) truncated the communal aspect of ethical obligations, often limiting discussion to apolitical responsibilities rather than to the full range of social involvements. The reestablishment of the State of Israel in the 20th century has, therefore, reopened for discussion areas that have for millennia been either ignored or relegated to the realm of abstraction. What this implies is that the full ethical responsibility of the Jew cannot be carried out solely within the realm of individual relationships but must include involvement in the life of a fully articulated community.

The just man in a just society

This double involvement is most vividly apparent in the biblical period, when both were equally present as divine command and demand. In the rabbinic period, because of the new political context, the communal aspect receded, so that discussion was mainly oriented toward the relationships between the members of the Jewish community or between individuals as such, and away from political responsibilities in the larger society. Nonetheless, the virtues that were understood to govern these relationships were, in their biblical setting, communal as well. Righteousness and compassion had been obligations of the state, governing the relationship between political units, as the first two chapters of Amos make evident. At the same time, as Micah 6:8 shows, doing justly, loving mercy, and walking humbly with God made up the pattern of the individual's obligations as well. Given the situation of the dispersion of the Jews following the revolts against Rome in the 1st and 2nd centuries CE, the individual pattern became the object of primary considerations. It is important to recognize that while theoretical ethical systems were not developed until the Middle Ages under the influence of philosophical concerns, nonetheless, even in the early period it was understood that behind the practical system of Halakha, the enumeration of legal precepts, there stood the dynamic of ethical theory. An attempt was made to reduce the hundreds of precepts to a small number expressing the ethical essence of Torah.

The key moral virtues. In keeping with the rabbinic understanding of Torah, study also was viewed as an ethical virtue. A passage in the traditional Prayer Book enumerates a series of virtuous acts—honouring parents, deeds of steadfast love, attendance twice daily at worship, hospitality to wayfarers, visiting the sick, dowering brides, accompanying the dead to the grave, devotion in prayer, peacemaking in the community and in family-life—and concludes by setting study of Torah as the premier virtue. Here is exhibited the complex variety of ethical behaviour called for within the Jewish tradition. To parental respect and family tranquillity are added, in other contexts, the responsibility of parents for children, the duties of husband and wife in the establishment and maintenance of a family, and ethical obligations that extend from the conjugal rights of each to the protection of the wife if the marriage is dissolved. The biblical description of God as upholding the cause of the fatherless and the widow and befriending the stranger, providing him with food and clothing (Deut. 10:18), remained a motivating factor in the structure of the community. Ethical requirements in economic life are expressed concretely in such a passage as Lev. 19:35–36: "You shall do no wrong in judgment, in measures of length or weight or quantity. You shall have just balances, just weights, a just *ephah,* and a just *hin*"; and in Amos' bitter condemnation of those who "sell the righteous for silver, and the needy for a pair of shoes" (Amos 2:6). Such injunctions, together with many other specific precepts and expressions of moral requirements, established the basis for a wide-ranging program that sought to govern, both in detail and in general, the economic life of the individual and the community. Not only are relations within the human sphere the object of ethical concern but nature also is so regarded. The animal world, in the biblical view, requires merciful consideration, so that not only man is commanded to rest on the Sabbath but his domestic animals are to share the rest with him (Ex. 20:10; 23:12). Mistreatment of beasts of burden is prohibited (Deut. 22:4); and wanton destruction of animal life falls under the ban (*ibidem:* 6–7). In the rabbinic attitude toward brute creation, even inanimate nature is the object of human solicitude. Thus, for example, the food-yielding trees of a city under siege may not be destroyed, according to Deuteronomic legislation (Deut. 20:14–20). The enlargement of this and other biblical precepts resulted in the generalized rabbinic prohibition "You shall not destroy" that governs man's use of his environment.

Protection of the weak

The relation to non-Jewish communities and cultures. As noted above, the end of the Jewish state reduced the scope of ethical judgments in the political sphere; nonetheless, relations between the Jewish community and other societies—particularly political units: the Roman and Christian empires, the Islāmic states, and other regimes—provided opportunities for the exploration of the ethical implications of such encounters. Since most of these were victor-victim, superior-inferior, power-powerless situations, with the Jews the weaker party, prudential considerations were dominant. Despite this, Jewish authorities sought to bring to bear upon these external arrangements the ethical standards that governed the internal structures.

The whole problem of the relationship between the Jewish community, in whatever form it has existed and does exist, and other social units has been vastly complicated. Ideally, the relation is that of witness to the divine intent in the world. Practically, it has swung between the extremes of isolation and assimilation, in which the ideal has, on occasion, been lost sight of. Culturally, from its earliest beginnings, the people Israel has met and engaged the ideas, forms, behaviour and attitudes of its neighbours in constructive development. It borrowed as it contributed and reformulated what it received in terms of its own commitments and affirmations. On more than a few occasions, as in the period of settlement in Canaan, it rejected the religiocultural ideas and forms of the native population. On others, it actively sought out—as in the Islāmic period in Spain (8th to 15th centuries)—ideas and cultural patterns of its neighbours, viewing them from its own perspective and embracing them when they were found to be of value. Indeed, the whole history of Israel's relationship with the world may be comprehended in the metaphor, used previously, of the heartbeat with its systole and diastole. No period of its existence discloses either total rejection of or abject surrender to other cultural and political structures but rather a tension, with the focal point always in motion at varying rates. Being more than a "confession" in the Christian sense, Judaism's adjustment to and relation with other sociopolitical units involved larger aspects of communal and individual life than merely the religious. Whether or not, under such circumstances, it is helpful to describe Judaism as a civilization, it is important to recognize that, viewed functionally, much more must be included than is usually subsumed under the common usage of "religion."

Isolation and assimilation

The formulation of Jewish ethical doctrines. The ethical concerns of Judaism have found frequent literary expression. Not only were rabbinic writings constantly directed toward the establishment of legal patterns that embody such concerns but in the medieval period the issues were dealt with in treatises on morals; in ethical wills, in which a father instructed his children about their obligations and behaviour; in sermons; and in other forms. In the 19th century the traditionalist Musar ("moral instructor") movement in eastern Europe and the philosophical discussions of the nascent Reform movement in the West focussed upon ethics. Indeed, since the political and social emancipation of the Jews, ethical and social rather than theological questions have tended to be given priority. Often the positions espoused have turned out to be, nonetheless, "judaized" versions of philosophic ethics or of political programs. In some instances, as in the case of the distinguished German-Jewish philosopher Hermann Cohen, the result has been a Jewishly compelling restatement of a secular philosophic ethics. In others, it has resulted in no more than a pastiche. More crucial, however, is the question of a unique Jewish ethics and of its authority. The reestablishment of a Jewish state renews the possibility that the full range of ethical decisions, including communal as well as individual responsibility, may be confronted. In such a situation, the ideal task of the people moves out of the realm of speculation to become actual again.

The Musar movement

THE UNIVERSE

Creation and Providence: God's world. Although the first chapter of Genesis affirms divine creation, it does not offer an entirely unambiguous view of the origin of the universe, as the debate over the correct understanding of Gen. 1:1 in former as in modern times discloses. (Was there or was there not a preexisting matter, void, or chaos?) Yet, basically, the interest of the author was not in the mode of creation, a later concern perhaps reflected in the various translations of the verse: "In the beginning God created," which could signify what medieval philosophers designated *creatio ex nihilo* ("creation out of nothing"); and "when God began to create," which could indicate some concept of prime matter. He was concerned rather to affirm that the totality of existence, inanimate (Gen. 1:3–19), living (20–25), and human (26–31), derived immediately from the same divine source; and, thus, that it is a universe. As divine creation, it is transparent to the presence of God, so that the Psalmist said: "The heavens declare the glory of God, and the expanse proclaims [that it is] the work of his hands" (19:1). Indeed, the repeated phrase: "And God saw how good it was" (Gen. 1:4, 10, 12, 18, 25, 31) may be understood as the ground of this affirmation, for the workmanship discloses the workman. The observed order of the universe is further understood by the biblical author to be the direct result of a covenantal relationship established between the world and God: "So long as the earth endures, seedtime and harvest. Cold and heat, summer and winter, day and night, shall not cease." (Gen. 8:22). This doctrine of the providential ordering of the universe, reaffirmed in rabbinic Judaism, is not without its difficulties, as in the liturgical change made in Isa. 45:7 to avoid ascribing evil to God. Nonetheless, despite the problem of theodicy (the problem of evil in a world made and ordered by God), Judaism has not acquiesced to the mood reported in the Palestinian Targum to Gen. 4:8: "He did not create the world in mercy nor does he rule in mercy." Rather, it has affirmed a benevolent and compassionate God.

Creation and history

It is the physical world—divine creation—that provides the stage for history, which is the place of the divine human encounter. An early Midrash, in response to the question as to why Scripture begins with the story of creation, points out that it was necessary in order to establish the identity of the Creator with the Giver of Torah, an argument basic to the liturgical structure of the Shema. This relationship is further emphasized in the Qiddush, the prayer of sanctification recited at the beginning of the Sabbath. That day is designated "a remembrance of creation" and "a recollection of the going-forth from Egypt." Thus, creation (nature) and history are understood to be inextricably bound up, for both derive from the same divine source. This being so, redemption—the reconciliation of God and man through and in history—does not ignore or exclude the natural world. Using the imagery of an extravagantly fecund world of nature, rabbinic thought expressed its view of the all-inclusive effects of the restored relationship.

Man's place in the universe. Man as creature is, of course, subject to the natural order. It is, indeed, in the world and through the world that man carries out his relationship to God. The commandments of Torah are obeyed not solely as observances between man and God but as actions between man and man, between man and the world. Although the creation story designates man as ruler over the earth and its inhabitants (Gen. 1:26–28; see also Ps. 8:5–9), nonetheless, far from being an arbitrary master, man's dominion is limited by Torah, for its regulations are concerned not only with transactions between man and man but also lay out his responsibilities to the land he cultivates, the produce of the soil, the animals he domesticates. Bound in the network of existence he, as the moral creature, is responsible for it in all of its parts.

Man's responsibility to nature

Even the destruction of the Jewish commonwealth in the 1st and 2nd centuries CE did not alienate the Jew from these responsibilities, as the elaborate system of Mishna and Gemara gives evidence. The gradual but consistent exclusion of the Jewish community from immediate connection with large segments of the natural world, through legislation in Christendom and Islām, tended to dull the Jew's awareness of it; the recurring references to it in the religious calendar, however, and the observation of harvest festivals even by citydwellers continued to remind the community of its ties. Thus, at the end of the 19th century, the nascent Zionist movement recognized that the regeneration of the Jewish people involved, among other requirements, a responsible relation to the natural order expressed in its attitude toward and treatment of the land.

As indicated in other contexts, the particular emphasis placed on one or the other side of the frequent twofoldness of the Jewish view has depended upon the situation in which the community has found itself. If nature as the place of divine disclosure has, during long periods of Jewish existence, assumed a somewhat subordinate role, it has never been rejected or been seen to be irrelevant to the divine purpose. Indeed, in Jewish eschatology, its restoration is part of the goal of history.

Intermediary beings: angels and demons. The exact nature of the nonhuman beings mentioned in Scripture—angels or messengers—is not altogether clear and their roles seem ephemeral. In the postexilic period, perhaps under Iranian influence, and in the late biblical and postbiblical literature, these beings emerge as more complete and often as clearly identifiable individuals with their own personal names. The unfocussed biblical view gave way to an elaborate hierarchy of functionaries who acted, in some apocalyptic visions, as a veritable heavenly bureaucracy. Nevertheless, despite a consensus concerning their existence, there was little agreement as to their role or importance. In some Midrashim God takes counsel with them; in other sources the rabbis urge men not to involve them but to approach God directly. Actually, they belong to that marginal area between religion and folklore. Like their counterfigures, the demons, they have a residual existence rooted in various layers of the Jewish experience and interpretation of the universe. At some times they are highly individualized and sharply realized; at others, they flit in and out of the imagination like bats in the evening. The medieval philosophers Aristotelized or Platonized them; the early mystics Neoplatonized them; the Kabbalists continually invented new ones and fitted them into their complicated network of cosmic existence. Nonetheless, their role, even in periods of considerable emphasis, was peripheral. They were outside the great movements and meanings of Jewish thought.

Contemporary philosophical speculation about the nature of the universe has, of course, required a response from Jewish thinkers. But, given the particular temper of a period in which metaphysics has not been central to much of theological discussion, no major statement has yet developed that has taken hold of the dominant positions and attempted to view them from the Jewish creationist perspective. The attempt within Reconstructionism to provide a naturalistic framework for Judaism, while courageous, lacks the breadth and depth of the great philosophical approaches.

ESCHATOLOGY

The future age of mankind and the world. The choice of Israel, according to the biblical writings, had occurred because of mankind's continual failure, by rebellion against its Creator, to fulfill its divine potential. The subsequent failure of Israel to become the holy community and thereby a witness to the nations gave rise to the prophetic movement that summoned the people to obedience. An integral part of prophetic summoning, side by side with threats of punishment and warnings of disaster, was the envisioning of a truly holy community, a society fully responding to the divine imperative. This kingdom of the future was conceived of as entirely natural, functioning as any normal sociopolitical unit and under the leadership of a human ruler, who would, however, carry out his tasks within the sphere of divine sovereignty, serving primarily to exhibit his own obedience and thus to stimulate the obedience of the entire people. This human monarch of the future was often, although not always, portrayed in terms of an idealized David, using such features of his life and reign as would underscore submission to God and emphasized social stability, economic satisfaction, and peace. During the period of the monarchy, the prophetic demand was directed toward each succeeding king, with the hope or even the expectation that he would be or become the new David, or the ideal ruler.

The Davidic model

The Babylonian Exile added a new measure of urgency to this expectation, although it was not expressed in any uniform fashion. The later chapters of the Book of Ezekiel provide in largely impersonal fashion the constitution for the new commonwealth but do not describe the peculiar characteristics of the ruler, while the later chapters of the Book of Isaiah focus on several figures—including Cyrus the Mede—who are seen as the divine instruments ushering in a new era. It is important to recognize that while such figures have extraordinary virtues ascribed to them, these virtues are neither superhuman nor suprahuman but such as are ultimately required of all Israel and of all men. The frustrations of the postexilic period, when the several attempts to bring into being the holy community had no more than partial success and were thwarted by the imperial designs of the great powers—as they had been in the preexilic period as well—led to an emphasis upon the futuristic quality of the messianic hope. This was abetted undoubtedly by external influences, such as Iranian thought, in which the cosmic rather than the historic aspect of a future era dominated. Since ancient cosmic myths—in good measure demythologized—had been part of the Israelite intellectual inheritance, evidenced at least in literary usages throughout Scriptures, the impact of such neighbouring ideas was to reinvigorate the mythic elements. Thus, hopes for the future at the end of the Persian period and on through the Hellenistic developments after *c.* 330 BCE comprised both historical expectations focussed upon a sociopolitical community and cosmic-mythic visions that moved on a broader stage. The latter were, of course, never entirely absent from the historical expectations and situations, for a renewal of nature was viewed as integral to the functioning of true society. The obedient community required, and was to be granted, a natural world in which true human relations could exist. In its most vivid forms, apocalypses (*i.e.,* visionary disclosures of the future), the literature of the period affords a remarkable insight into the agonies and urgencies of the people. After the failures, saving events, and disappointments of the past are recounted, the present, in transparent disguise, is portrayed and the immediately hoped-for intervention of God is described in awesome detail as a means of affirming and confirming the faith of those who saw themselves as the remnant, or perhaps the promise, of the holy community.

The socio-political and cosmic-mythic aspects

The king-messiah and his reign. Put schematically, Israel's hope was for the restoration of divine sovereignty over all of creation. Concretely, that hope found a considerable variety of expressions. Of all such expressions, that which centred around the idealized king began to assume an ever more important (but never exclusive) role. Many of the writings that report the ideas and attitudes of the Jewish community in the period immediately preceding and following the rise of Christianity are either ignorant of or more probably indifferent to the personal element. God is envisioned as the protagonist of the end, actively intervening or sending his messengers (*i.e.,* angels), to perform specific acts in ending the old and inaugurating the new era. On the other hand, in some writings of the period the anointed king-messiah (Hebrew, *mashiaḥ,* "anointed")—the title reflects the episode in I Sam. 16 in which David is thus singled out as the divinely chosen ruler—becomes more sharply defined as the central figure in the culminating events and, given the cosmic-mythic components, assumes suprahuman and in some instances, even quasi-divine, aspects. It is clear, then, that the doctrine of last things in Judaism is not necessarily messianic, if that term is properly limited to an inauguration of a future era through the action of a human, suprahuman, or quasi-divine person. Nonetheless, it must be recognized that the messianic version of eschatology played a more compelling role in rabbinic Judaism than other modes. The same is true with regard to the locus of the "world (or age) to come." Given the ingredients noted above, it was possible to construct various eschatological landscapes ranging from the mundane to the celestial, from Jerusalem in the hills of Judah to a heavenly city. Indeed, confronted with an embarrassment of riches, the medieval theologians sought to combine them into an inclusive system that intricately involved as large a variety of the possibilities as could be brought together. In such patterns the messianic this-worldly emphasis was understood as a preliminary movement toward an ultimate resolution. The ideal ruler, the new David, would reestablish the kingdom in its own land (in "Zion," or Palestine) and would reign in righteousness, equity, justice, and truth, thus bringing into being the holy nation and summoning all mankind to dwell under divine sovereignty. As a component of this reestablished kingdom, the righteous dead of Israel would be resurrected to enjoy life in the true community that did not exist in their days. This kingdom, however long it was destined to endure, was not permanent. It would come to an end either at a predetermined time or as victim of the unrepentant nations and cosmic foes, at which point the ultimate intervention by God would take place. All the wicked throughout history would be recalled to life, judged, and doomed; all the righteous would be transformed and transported into a new world; *i.e.,* creation would be totally restored. Particular emphases that one or the other of these ideas received, the ways in which they were interpreted—philosophically, mystically, or ethically—were determined most frequently by the situations and conditions in which the Jewish community found itself. With such a considerable body of ideas at its disposal and with the details of none of them ever receiving the kind of affirmation that statements about God, Torah, and Israel had, freedom of speculation in the realm of eschatology was little restricted. Thus, Joseph Albo (15th century) in his work on Jewish "dogmas," the *Sefer ha-ʿiqqarim,* was not inhibited from denying that belief in the messiah was fundamental. The mystical movements of the Middle Ages found in eschatological hopes a crucial centre. The early Kabbala was little interested in messianism, for it interiorized the expectations in the direction of personal redemption. Following the disasters of the late 15th to 17th centuries (*e.g.,* the expulsion of the Jews from Spain and the Cossack massacre of the Jews in Poland) however, messianic speculation in all of its varieties underwent a luxuriant growth, finally running wild in the movements surrounding Shabbetai Tzevi of Smyrna and later Jacob Frank of Offenbach. These tragedies for the Jewish communities once again resulted in a futurizing of the hopes or at least a limiting of their application (see also below, *Jewish mysticism*).

Resurrection of the dead and Last Judgment

Secularization of messianism. In the 19th century, with the political emancipation of the Jews in western Europe and the development of an optimistic evolutionism, messianism was transformed by many liberal thinkers into a version of the idea of progress whose goal was often thought of as immediately attainable through enlightened social and political action. When disillusionment with the emancipation set in, messianism was even more completely secularized in some segments of the community who saw its meaning and fulfillment in some form of socialism—again, rather close at hand. In others, it was absorbed into the emerging political nationalism—Zionism. Similar developments took place in eastern Europe, with parallel transformations. In more recent times, particularly since the events symbolized by the name Auschwitz (a Nazi death camp in Poland, where millions of Jews were exterminated), the earlier modern interpretations, particularly of messianism, but also of eschatology as a whole, have been considered inadequate. Although no compelling statement has been forthcoming, Jewish thinkers in the second half of the 20th century have been attempting once again to come to grips with eschatological concepts in all of their varieties and forms.

Basic practices and institutions

THE HALLOWING OF EVERYDAY EXISTENCE

The centrality of Halakha

Systematic presentations of the affirmations of the Jewish community never served as the sole mode of expressing beliefs of the people. Side by side with speculation—Haggadic, philosophic, mystical, or ethical—there stood, not in a secondary role but as the other of the double focuses, Halakha ("practice," "rules of conduct"), the paradigmatic statement of the behaviour, individual and communal, that embodied concretely the beliefs conceptualized in speculation. Life in the holy community was understood to embrace every level of human existence. The prophets vigorously resisted attempts to limit the sovereignty of the God of Israel to organized worship and ritual. The Pharisees, even while the Jerusalem Temple cult was still in existence, sought to reduce priestly exclusiveness by enlarging the scope of sacral rules to include, as far as possible, all of the people. Rabbinic Judaism, Pharisaism's surviving descendant, continued the process of democratization and sought, through its system of interpretation, to find in every occasion of life a means of affirming divine concern and presence. Viewed negatively by some Protestant theologians, this development has been judged to stifle spontaneity. Yet spontaneity is not necessarily lacking in a world governed by Halakha, although the danger of the stylized routine in religious and ethical life is apparent. Nonetheless, the intention of the Halakhic attitude is to remind the Jew constantly that each and every occasion of life is a locus of divine disclosure. This is most clearly seen in the *berakhot,* the "blessings," that are prescribed to accompany the performance of a broad spectrum of human actions, from the commonplace routines of daily life to the restricted gestures of the cultic-liturgical year In these, God is addressed directly in the second person singular, his sovereignty is affirmed, and his activity as Creator, Giver of Torah, or redeemer, expressed in a wide variety of eulogies, is proclaimed. There are no areas of human behaviour in which man cannot be met by God, and in terms of its intention, the Halakhic pattern is designed to make such possibilities experienced realities. Yet, again, it must be noted that the situation of the Jewish community determines in a very large way how the intention is actualized. On more than one occasion the Halakhic pattern has served as a defense against a hostile environment and has thus tended to become scrupulosity (an obsessive concern with minute details), but the dynamic of the intention itself has as often broken through to re-establish its integrity and hallow life in its wholeness.

THE TRADITIONAL PATTERN OF INDIVIDUAL AND FAMILIAL PRACTICES

Perspective on the traditional pattern of an individual's life is obtained by examining a passage from the Babylonian Talmud (tractate *Berakhot* 60b) that was subsequently reworked into a liturgical structure but which in its original form exhibits the intention discussed above. In this passage, the blessings accompanying a man's waking and returning to the routines of life are prescribed. There is a brief thanksgiving on awakening for being restored to conscious life; then the impingement of the external world is responded to in a benediction over the cock's crowing; following this, each ordinary act, opening one's eyes, stretching and sitting up, dressing, standing up, walking, tying one's shoes, fastening one's belt, covering the head, washing the hands and face, has its accompanying blessing, reminding a man that the world and the life to which he has returned exist in the presence of God. These are followed by a supplication in which the petitioner asks that his life during the day may be worthy in all of its relationships. Then, as the first order of daily business, Torah, both written (Bible) and oral (Mishna), is briefly studied, introduced by eulogies of God as Giver of Torah. Finally, there is a prayer for the establishment of the Kingdom of God, for each day contains within itself the possibility of ultimate fulfillment. As indicated, this was originally not a part of public worship (even today it is, strictly speaking, not part of the synagogue service, although it is most frequently recited there) but was personal preparation for a life to be lived in the presence of God.

The liturgy of daily life

From H. Hapgood, *The Spirit of the Ghetto;* © 1965 by Funk & Wagnalls Co., Inc., reproduced by permission of Funk & Wagnalls Publishing Co., Inc.

Boy putting on phylacteries in preparation for morning prayer, drawing by Jacob Epstein, *c.* 1902.

Such individual responsibility marks much of Jewish observance, so that the synagogue—far from being the focus of observance—shares with the home and the workaday world the opportunities for the divine-human encounter. The table blessings, Qiddush (the "sanctification" of the Sabbath and festivals), the erection of the booth (*sukka*) for Sukkot (the Feast of Tabernacles), the seder (the festive Passover meal) with its symbols and narration of the Exodus, the lighting of the lamps during the eight days of Ḥanukka (the Feast of Dedication), are all the obligation of the individual and the family and have their place in the home. It is here, too, that woman's role is defined and here, as contrasted with the synagogue, that she functions centrally. Given the traditional dietary regimen of the Jewish community—the exclusion of swine, carrion eaters, shellfish, and other creatures, the separation of meat and dairy products, the ritual slaughtering of animals, the required separation and burning of a small portion of dough (*ḥalla*) when baking, the supervision of the Passover food requirements, and many other stipulations—there exists a large and meticulously governed

The central role of women in family religion area within the home that is indeed the sphere of woman's religion. There seems not to have been a hierarchy of values in which the home-centred, as contrasted with the synagogue-oriented, practices were given an inferior status. In modern times, however, particularly in Western civilization where the pervasiveness of religious obligation has been replaced by ecclesiastical institutionalism, on the prevailing Christian model, this whole crucial area has lost much of its meaning as a place of divine-human meeting. Thus, for many it is only the synagogue that provides such an opportunity, and the individual act has been reduced on the scale of values. With this downgrading, woman's religion has lost its significance so that her status—when parallels are drawn to her role in the larger society—has been reduced to one of inferiority. However attenuated personal religious responsibility may have become in some environments or transformed into stylized cultural forms, the intention that informs the Halakhic structure, the hallowing of the individual's total existence, remains a potent force within the Jewish community.

THE TRADITIONAL PATTERN OF SYNAGOGUE PRACTICES

The other focus of observance is the synagogue. The origins of this institution are obscure and a number of hypotheses have been proposed to account for the appearance of this essentially lay-oriented form of worship. What seems certain is that during the period of the Second Temple—following the return from Babylon and continuing until the Temple destruction in 70 CE—there were, side by side with the official cult, other modes of worship more or less independent of the priesthood and nonsacrificial in form. The reports by the philosopher Philo and the historian Josephus in the 1st century, buttressed by the Qumrān document (Dead Sea Scrolls), provide some knowledge of the practices of the contemporary Essenes; rabbinic sources, including the earliest layers of the traditional order of worship, enable us to understand another, apparently Pharisaic, mode; the brief allusions to the practices of James and his Jewish Christian companions in the book of Acts suggest yet other varieties. In any case, the grouping that formed the cadre of what eventually became rabbinic Judaism observed some form of worship that, with the destruction of the Temple cult, was able to provide a new centre and even to absorb enough from the defunct priestly institution to suggest continuity and legitimacy. This was probably the basic pattern for synagogal liturgy in the millennia that followed.

Readings from the Torah and prophets

At the heart of synagogal worship is the public reading of Scriptures. This takes place at the morning service on Sabbaths, holy days, and festivals, on Monday and Thursday mornings, and on Sabbath afternoons. The readings from the Pentateuch are presently arranged in an annual cycle so that, beginning on the Sabbath following the autumnal festivals with Gen. 1:1, the entire five books are read through the rest of the year. The texts for festivals, holy days, and fasts reflect the particular significance of those occasions. In addition, a second portion from the prophetic writings (in the Jewish tradition these include Joshua, Judges, Samuel, and Kings, as well as the three major and 12 minor prophets, but not Daniel) is read on many of these occasions. All of this takes place within the structure of public worship and is provided with ceremonies during which the Sefer Torah ("Book of the Torah"), the pentateuchal scroll, is removed from the ark (cabinet) at the front of the synagogue, and carried in procession to the reading desk; from it, the pertinent text is chanted by the reader. The text for the service is divided into subsections varying from seven on the Sabbath to three at the weekday morning service, and individuals are called forward to recite the blessings eulogizing God as Giver of Torah before and after each of these. The order of worship is composed of the preparatory blessings and prayers noted above, to which are added passages recalling the Temple sacrificial cult (thus relating the present form of worship to the past); the recitation of a number of Psalms and biblical prayers; the Shema and its accompanying benedictions, introduced by a call to worship that marks the beginning of formal public worship; the prayer (*tefilla*) in the strict sense of petition; confession and supplication (*taḥanun*) on weekdays; the reading of Scripture; and concluding acts of worship. This general structure of the morning service varies somewhat, with additions and subtractions for the afternoon and evening services and for Sabbath, holy days, and festivals.

The "Eighteen Benedictions"

The prayer (*tefilla*), just mentioned, is often called the *shemone ʿesre,* the "Eighteen Benedictions"—although it actually has 19—or the *ʿamida,* "standing," because it is recited in that position. It is made up of three introductory benedictions: praise of the God of the Fathers, of God the Redeemer who resurrects the dead, and of God the holy one who fills the earth with his glory, and of three concluding acts; a prayer for the acceptance of the service, a thanksgiving, and a prayer for peace—with a series of intermediate petitions for knowledge, well-being, acceptance of repentance, forgiveness of sin, and others. On the Sabbath and festivals these are replaced by benedictions that mention the specific occasion but are not petitionary, it being considered inappropriate to attend to workday concerns at these times.

While the general outline of this order of service is found throughout the entire Jewish world, the details have varied, both in different periods and in geographic and cultural areas. The public service, requiring the presence of at least 10 males, the *minyan* ("quorum"), is generally led by a synagogal official, the *ḥazzan,* or cantor, but any Jewish male with the requisite knowledge may act in this capacity since there is, quite strictly, no clerical class in the community to whom such leadership is limited (see *The rabbi,* below).

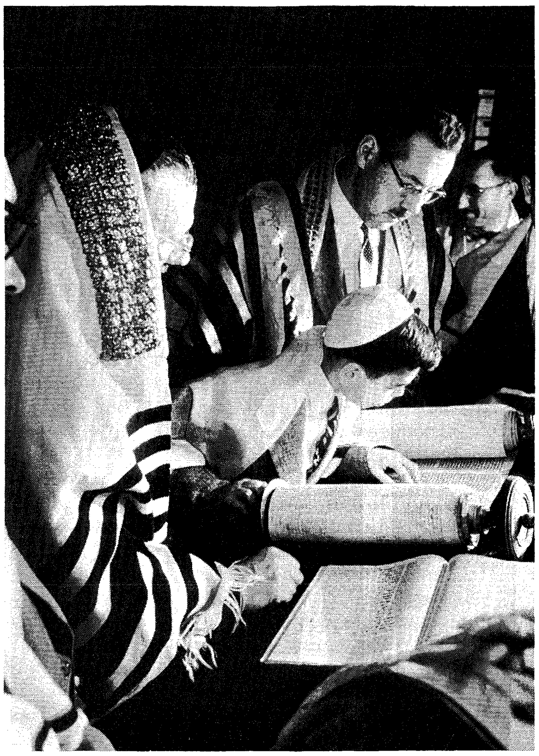
The synagogue room itself has a very simple basic form although, of course, it may be embellished considerably. The only requirements are a container for the Torah scroll(s), the *aron ha-qodesh* ("the holy ark")—a chest against the east wall, or a recessed closet with doors and a curtain; a prayer desk (*ʿamud*) facing the ark at which the reader stands when reciting the service; and the pulpit (*bima*)—according to some requirements in or close to the centre of the room—from which the Torah is read. In the Spanish-Portuguese tradition, only one desk (called *teva*) is used. The ark contains one or more scrolls, on which are written the five books of Moses. These are variously ornamented, depending upon the cultural region: European communities decking them in coverings of cloth; Oriental (North African and Near Eastern) placing them in wooden or metal containers. In addition, silver ornaments, in the form of towers or crowns, are often set on the tops of two rods on which the scroll is wound, and a breastplate and a pointer are suspended from them.

Accommodations for the worshippers vary according to the cultural milieu, from rugs and cushions in Oriental synagogues to pews and standing desks in European ones. Given this essential simplicity, the synagogue room itself may be used for other purposes than worship, *e.g.,* study and community assembly. Again, this varies with the cultural pattern.

CEREMONIES MARKING THE INDIVIDUAL LIFE CYCLES

Circumcision and naming

There are within Jewish life two cycles corresponding to the individual and the synagogal focuses, although they necessarily intermingle. The life of the individual is marked by observances that single out the notable events of personal existence. A male child is circumcised on the eighth day following birth, as a covenantal sign (Gen. 17); the rite of circumcision (*berit mila*) is accompanied by appropriate benedictions and ceremonies, including naming. Females are named in the synagogue, generally on the Sabbath following birth, when the father is called to recite the benedictions over the reading of Torah. A firstborn son, if he does not belong to a priestly or a levitical family, is rcdeemed at one month (in accordance with Ex. 13:12–13 and Num. 18:14–16) by the payment of a stipulated sum to a cohen (a putative member of the priestly family). On arrival at the age of 13, a boy is called publicly to recite the Torah benedictions, thus signifying his religious coming-of-age; he is thenceforth obligated to observe the commandments as his own responsibility—he is now a Bar Mitzwa ("Son of the Commandment"). Marriage (*ḥatuna,* also *Qiddushin,* "sanctifications") involves



Boy reading the Torah at synagogue services, an important part of the Bar Mitzwa ceremony.

Cornell Capa—Magnum

a double ceremony, performed together in modern times but separated in ancient times by a year. First is the betrothal (*erusin*), which includes the reading of the marriage contract (*ketubba*) and the giving of the ring with a declaration, "Behold you are consecrated to me by this ring according to the law of Moses and Israel," accompanied by certain benedictions. This is followed by the marriage proper (*nissuʾin*), consisting of the reciting of the seven marriage benedictions. The ceremony is performed under a *ḥuppa*, a canopy, that symbolizes the bridal bower.

The burial service is marked by simplicity. The body is prepared for the grave by the *ḥevraʾ qaddishaʾ* (the holy society), clad only in a simple shroud, and the interment takes place as soon after death as possible. In Israel no coffin is used. There are observances connected with death, many of which belong to the realm of folklore rather than Halakhic tradition. A mourning period of 30 days is observed, of which the first seven (*Shivʿa*) are the most rigorous. During the 11 months following a death, the bereaved recite a particular form of a synagogal doxology (Qaddish) during the public service as an act of memorial. The doxology itself, entirely devoid of any mention of death, is a praise of God and a prayer for the establishment of the coming Kingdom. It is also recited annually on the anniversary of the death (*yahrzeit*). (L.H.S.)

THE CYCLE OF THE RELIGIOUS YEAR

The term Jewish religious year as used in this section encompasses the cycle of Sabbaths and holidays that are commonly observed by the Jewish religious community—and officially in Israel by the Jewish secular community as well. The Sabbath and festivals are bound to the Jewish calendar, reoccur at fixed intervals, and are celebrated at home and in the synagogue according to ritual set forth in Jewish law and hallowed by Jewish custom. According to Jewish teaching, the Sabbath and festivals are, in the first instance, commemorative. The Sabbath, for example, commemorates the Creation, and Passover commemorates the Exodus from Egypt over 3,000 years ago. The past is not merely recalled; it is also relived through the Sabbath and festival observances. Creative physical activity ceases on the Sabbath as it did, according to Genesis, when the Creation was completed; Jews leave their homes and reside in booths during the Sukkot festival as did their biblical ancestors. Moreover, Sabbath and festival themes are considered to be perpetually significant, recurring and renewed in every generation. Thus the revelation of the Torah (the divine teaching or law) at Sinai, commemorated on Shavuot, is considered an ongoing process which recurs whenever a commitment is made to Torah study.

An important aspect of Sabbath and festival observance is sanctification. The Sabbath and festivals sanctified the Jews more than the Jews sanctified the Sabbath and festivals. Mundane meals became sacred meals; joy and relaxation became sacred obligations (*mitzwot*). No less significant is the contribution of the Sabbath and festivals toward communal awareness. Thus, neither Sabbath nor festival can be properly observed in the synagogue according to the ancient tradition if fewer than 10 male Jews are present. Again, a Jew prays on Rosh Hashana and mourns on Tisha be-Av not only for his own fate but for the fate of all Jews. The sense of social cohesiveness fostered by the Sabbath and festival observances has stood the Jews well throughout their long, often tortuous history.

The seven-day week, the notion of a weekly day of rest, and many Christian and Islāmic holiday observances owe their origins to the Jewish calendar, Sabbath, and festivals.

The Jewish calendar. *Lunisolar structure.* The Jewish calendar is lunisolar—*i.e.,* regulated by the positions of both the moon and the sun. It consists usually of 12 alternating lunar months of 29 and 30 days each (except for Ḥeshvan and Kislev, which sometimes have either 29 or 30 days), and totals 353, 354, or 355 days per year. The average lunar year (354 days) is adjusted to the solar year (365¼ days) by the periodic introduction of leap years in order to assure that the major festivals fall in their proper season. The leap year consists of an additional 30-day month called First Adar, which always precedes the month of (Second) Adar. A leap year consists of either 383, 384, or 385 days and occurs seven times during every 19-year period (the so-called Metonic cycle). Among the consequences of the lunisolar structure are these: (1) The number of days in a year may vary considerably, from 353 to 385 days. (2) The first day of a month can fall on any day of the week, that day varying from year to year. Consequently, the days of the week upon which an annual Jewish festival falls vary from year to year despite the festival's fixed position in the Jewish month.

Leap year: Second Adar

Months and notable days. The months of the Jewish religious year, their approximate equivalent in the Western Gregorian calendar, and their notable days, are as follows:

Tishri (September–October)
1, 2 Rosh Hashana (New Year)
3 Tzom Gedaliahu (Fast of Gedaliah)
10 Yom Kippur (Day of Atonement)
15–21 Sukkot (Tabernacles)
22 Shemini Atzeret (Eighth Day of the Solemn Assembly)
23 Simḥat Torah (Rejoicing of the Law)

Ḥeshvan, or Marḥeshvan (October–November)

Kislev (November–December)
25 Ḥanukka (Feast of Dedication) begins

Ṭevet (December–January)
2–3 Ḥanukka ends
10 ʿAsara be-Tevet (Fast of Tevet 10)

Shevaṭ (January–February)
15 Tu bi-Shevaṭ (15th of Shevaṭ: New Year for Trees)

Adar (February–March)
13 Taʿanit Esther (Fast of Esther)
14, 15 Purim (Feast of Lots)

Nisan (March–April)
15–22 Pesaḥ (Passover)

Iyyar (April–May)
18 Lag ba-Omer (33rd Day of the Omer Counting)

Sivan (May–June)
6, 7 Shavuot (Feast of Weeks, or Pentecost)

Tammuz (June–July)
17 Shivaʿ ʿAsar be-Tammuz (Fast of Tammuz 17)

Av (July–August)
9 Tisha be-Av (Fast of Av 9)

Elul (August–September)

During leap year, the Adar holidays are postponed to Second Adar.

Since 1948 many Jewish calendars list Iyyar 5—Israel Independence Day—among the Jewish holidays.

Regulation by the Sanhedrin and patriarchate

Origin and development. The origin of the Jewish calendar can no longer be accurately traced. Some scholars suggest that a solar year prevailed in ancient Israel, but no convincing proofs have been offered, and it is more likely that a lunisolar calendar similar to that of ancient Babylonia prevailed in ancient Israel. In late Second Temple times (*i.e.,* 1st century BCE to 70 CE), calendrical matters were regulated by the Sanhedrin, or council of elders, at Jerusalem. The testimony of two witnesses who had observed the New Moon was ordinarily required to proclaim a new month. Leap years were proclaimed by a council of three or more rabbis with the approval of the *nasi,* or president, of the Sanhedrin. With the decline of the Sanhedrin, calendrical matters were decided by the Palestinian patriarchate (the official heads of the Jewish community under Roman rule). Jewish persecution under Constantius II (reigned 337–361) and advances in astronomical science led to the gradual replacement of observation by calculation. According to Hai ben Sherira (died 1038)—the head of a leading Talmudic academy in Babylonia—Hillel II, a Palestinian patriarch, introduced a fixed and continuous calendar in 359 CE. A summary of the regulations governing the present calendar is provided by Maimonides, the great medieval philosopher and legist, in his *Code: Sanctification of the New Moon,* chapters 6–10.

Fragments of writings discovered in a geniza (depository for sacred writings withdrawn from circulation) have brought to light a calendrical dispute between Aaron ben Meir, a 10th-century Palestinian descendant of the patriarchal (Hillel) family, and the Babylonian Jewish authorities, including Saʿadia ben Joseph—an eminent 10th-century philosopher and *gaon* (head of a talmudic academy). Ben Meir's calculations provided that Passover in 922 be celebrated two days earlier than the date fixed by the normative calendar. After a bitter exchange of letters, the controversy subsided in favour of the Babylonian authorities, whose hegemony in calendrical matters was never again challenged.

Calendars of various sectarian Jewish communities deviated considerably from the normative calendar described above. The Dead Sea (or Qumrān) community (made famous by the Dead Sea Scrolls discoveries) adopted the calendrical system of the noncanonical books of *Jubilees* and *Enoch,* which was essentially a solar calendar. Elements of this same calendar reappear among the Mishawites, a sect founded in the 9th century.

The Karaites, a sect founded in the 8th century, refused, with some exceptions, to recognize the normative fixed calendar and reintroduced observation of the New Moon. Leap years were determined by observing the maturation of the barley crop in Palestine. Consequently, Karaites often celebrated the festivals on dates different from those fixed by the rabbis. Later, in medieval times, the Karaites adopted some of the normative calendrical practices, while rejecting others.

The Sabbath. The Jewish Sabbath (from Hebrew *shavat,* "to rest") is observed throughout the year on the seventh day of the week—Saturday. According to biblical tradition, it commemorates the original seventh day on which God rested after completing the creation.

Scholars have not succeeded in tracing the origin of the seven-day week, nor can they account for the origin of the Sabbath. A seven-day week does not accord well with either a solar or lunar calendar. Some scholars, pointing to the Akkadian term *shapattu,* suggest a Babylonian origin for the seven-day week and the Sabbath. But *shapattu,* which refers to the day of the Full Moon and is nowhere described as a day of rest, has little in common with the Jewish Sabbath. It appears that the notion of the Sabbath as a holy day of rest, linking God to his people and recurring every seventh day, was unique to ancient Israel.

Importance. The central significance of the Sabbath for Judaism is reflected in the traditional commentative and interpretative literature called Talmud and Midrash (*e.g.,* "if you wish to destroy the Jewish people, abolish their Sabbath first") and in numerous legends and adages from more recent literature (*e.g.,* "more than Israel kept the Sabbath, the Sabbath kept Israel"). Some of the basic teachings of Judaism affirmed by the Sabbath are God's acts of creation, God's role in history, and God's covenant with Israel. Moreover, the Sabbath is the only Jewish holiday the observance of which is enjoined by the Ten Commandments. Jews are obligated to sanctify the Sabbath at home and in the synagogue by observing the Sabbath laws and engaging in worship and study. The leisure hours afforded by the ban against work on the Sabbath were put to good use by the rabbis, who used them to promote intellectual activity and spiritual regeneration among Jews. Other days of rest, such as the Christian Sunday and the Islāmic Friday, owe their origins to the Jewish Sabbath.

Worship, study, and abstinence from work

Observances. The biblical ban against work on the Sabbath, while never clearly defined, includes such activities as baking and cooking, travelling, kindling fire, gathering wood, buying and selling, and bearing burdens from one domain into another. The Talmudic rabbis listed 39 major categories of prohibited work, including agricultural activity (*e.g.,* plowing and reaping), work entailed in the manufacture of cloth (*e.g.,* spinning and weaving), work entailed in preparing documents (*e.g.,* writing), and other forms of constructive work.

At home, the Sabbath begins Friday evening some 20 minutes before sunset, with the kindling of the Sabbath candles by the wife, or in her absence by the husband. In the synagogue, the Sabbath is ushered in at sunset with the recital of selected psalms and the *Lekha Dodi,* a 16th-century Kabbalistic (mystical) poem. The refrain of the latter goes: "Come, my beloved, to meet the bride," the "bride" being the Sabbath. After the evening service, each Jewish household begins the first of three festive Sabbath meals by reciting the *Qiddush* ("sanctification" of the Sabbath) over a cup of wine. This is followed by a ritual washing of the hands and the breaking of bread; two loaves of bread (commemorating the double portions of manna described in Exodus) being placed before the breaker of bread at each Sabbath meal. After the festive meal, the remainder of the evening is devoted to study or relaxation. The distinctive features of the Sabbath morning synagogue service include the public reading of the Torah, or Five Books of Moses (the portion read varies from week to week) and, generally, the sermon, both of which serve to educate the listeners. Following the service, the second Sabbath meal begins, again preceded by *Qiddush* (of lesser significance), and conforming for the most part to the first Sabbath meal. The afternoon synagogue service is followed by the third festive meal (without *Qiddush*). After the evening service, the Sabbath comes to a close with the Havdala ("Distinction") ceremony, which consists of a benediction noting the distinction between Sabbath and weekday, usually recited over a cup of wine accompanied by a spice box and candle.

The Havdala ceremony

The Jewish holidays. The major Jewish holidays are the Pilgrim Festivals: Pesaḥ (Passover), Shavuot (Feast of Weeks, or Pentecost), and Sukkot (Tabernacles); and the High Holidays: Rosh Hashana (New Year) and Yom Kippur (Day of Atonement). In common, their observance is required by the Torah and work is prohibited for the duration of the holiday (except on the intermediary days of the Pesaḥ and Sukkot festivals, when work the neglect of which entails monetary loss is permitted). Purim (Feast of Lots) and Ḥanukka (Feast of Dedication), while not mentioned in the Torah (and therefore of lesser solemnity), were instituted by Jewish authorities in the Persian and Greco-Roman periods. Lacking the work restrictions characteristic of the major festivals, they are sometimes regarded as minor festivals. In addition, there are the five fasts: ʿAsara be-Ṭevet (Fast of 10 Ṭevet), Shivaʿ ʿAsar be-Tammuz (Fast of Tammuz 17), Tisha be-Av (Fast of Av 9), Tzom Gedaliahu (Fast of Gedaliah), and Taʿanit Esther (Fast of Esther); and the lesser holidays—*i.e.,* holidays the observances of which are few and not always clearly defined—such as Rosh Ḥodesh (First Day of the Month), Ṭu bi-Shevaṭ (New Year for Trees), and Lag ba-ʿOmer (33rd Day of Omer Counting). The fasts and the lesser holidays also lack the work restrictions characteristic of the major festivals. Some of the fasts and Rosh Ḥodesh are mentioned in Scripture, but most of the details concerning their proper observance, as well as those concerning the other lesser holidays, were provided by the Talmudic and medieval rabbis.

Pilgrim festivals. In Temple times, all males were required to appear at the Temple three times annually and actively participate in the festal offerings and celebrations. These were the joyous pilgrim festivals of Pesaḥ, Shavuot, and Sukkot. Originally, they marked the major agricultural seasons in ancient Israel and commemorated Israel's early history; but after the destruction of the Second Temple in 70 CE, emphasis was almost exclusively placed on the commemorative aspect.

In modern Israel, Pesaḥ, Shavuot, and Sukkot are celebrated for the number of days prescribed by Scripture, namely, seven days, one day, and eight days, respectively (with Shemini Atzeret added to Sukkot). Due to calendrical uncertainties which arose in Second Temple times (6th century BCE to 1st century CE), each festival is celebrated for an additional day in the Diaspora.

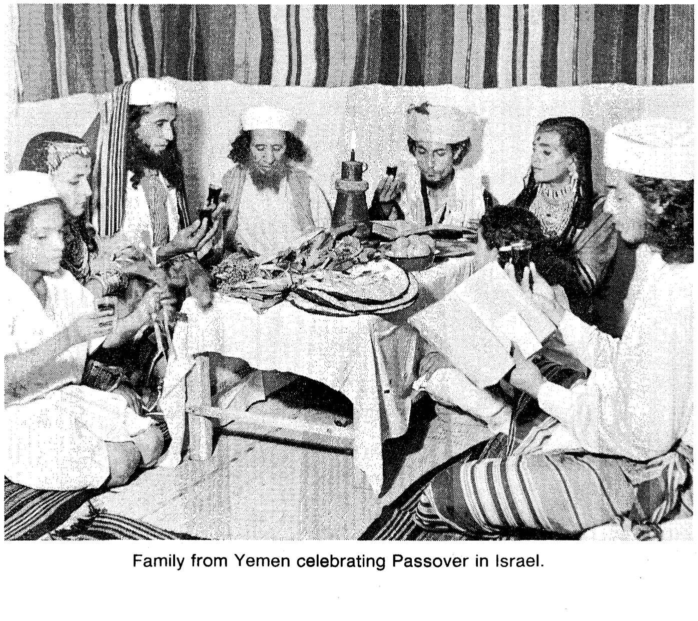
Pesaḥ
(Passover)

Pesaḥ commemorates the Exodus from Egypt and the servitude that preceded it. As such, it is the most significant of the commemorative holidays, for it celebrates the very inception of the Jewish people—*i.e.,* the event which provided the basis for the covenant between God and Israel. The term *pesaḥ* refers originally to the paschal (Passover) lamb sacrificed on the eve of the Exodus, the blood of which marked the Jewish homes to be spared from God's plague; its etymological significance, however, remains uncertain. The Hebrew root is usually rendered "passed over"—*i.e.,* God passed over the homes of the Israelites when inflicting the last plague on the Egyptians—hence the term Passover. The festival is also called Ḥag, Matzot ("Festival of Unleav- ened Bread"), for unleavened bread is the only kind of bread consumed during Passover.

Leaven (*seʾor*) and foods containing leaven (*ḥametz*) are neither to be owned nor consumed during Pesaḥ. Aside from meats, fresh fruits, and vegetables, it is customary to consume only those foods prepared under rabbinic supervision and labelled "kosher for Passover," warranting that they are completely free of contact with leaven. In many homes, special sets of crockery, cutlery, and cooking utensils are acquired for Passover use. On the evening preceding the 14th day of Nisan, the home is thoroughly searched for any trace of leaven (*bediqat ḥametz*). The following morning the remaining particles of leaven are destroyed by fire (*biʿur ḥametz*). From then until after Pesaḥ, no leaven is consumed. Many Jews sell their more valuable leaven products to non-Jews before Passover (*mekhirat ḥametz*), repurchasing the foodstuffs immediately after the holiday.

The unleavened bread (*matza*) consists entirely of flour and water, great care being taken to prevent any fermentation before baking. Hand-baked *matza* is flat, rounded, and perforated. Since the 19th century, many Jews have preferred the square-shaped, machine-made *matza.*

Popperfoto



Family from Yemen celebrating Passover in Israel.

Passover eve is ushered in at the synagogue service on the evening before Passover, after which each family partakes of the seder ("order of service); *i.e.,* an elaborate festival meal in which every ritual is regulated by the rabbis. (In the Diaspora, the seder is also celebrated on the second evening of Passover.) The table is bedecked with an assortment of foods symbolizing the passage from slavery (*e.g.,* bitter herbs) into freedom (*e.g.,* wine). The Haggada (literally "narration"), a printed manual comprised of appropriate passages culled from Scripture, Talmud, and Midrash, accompanied by medieval hymns, serves as a guide for the ensuing ceremonies and is recited as the evening proceeds. The seder opens with the cup of sanctification (Qiddush), the first of four cups of wine drunk by the celebrants. An invitation is extended to the needy to join the seder ceremonies, after which the youngest son asks four prescribed questions expressing his surprise at the many departures from usual mealtime procedure. ("How different this night is from all other nights!") The father then explains that the Jews were once slaves in Egypt, were then liberated by God, and now commemorate the servitude and freedom by means of the seder ceremonies. Special blessings are recited over the unleavened bread and the bitter herbs (*maror*), after which the main courses are served. The meal closes with a serving of *matza* recalling the paschal lamb, consumption of which concluded the meal in Temple times. The seder concludes with the joyous recital of hymns praising God's glorious acts in history and anticipating a messianic redemption to come.

The Passover liturgy is considerably expanded and includes the daily recitation of Psalms 113–118 (Hallel, "praise"), public readings from the Torah, and an additional service (*musaf*). On the first day of Pesaḥ, a prayer for dew in the Holy Land is recited; on the last day, the memorial service for the departed (*yizkor*) is added.

Shavuot
(Feast of
Weeks, or
Pentecost)

Originally an agricultural festival marking the wheat harvest, Shavuot commemorates the revelation of the Torah at Sinai. Shavuot ("weeks") takes its name from the seven weeks of grain harvest separating Passover and Shavuot. The festival is also called Ḥag ha-Qazir (Harvest Festival) and Yom ha-Bikkurim (Day of First Fruits). Greek-speaking Jews called it *pentēkostē,* meaning "the fiftieth" day after the sheaf offering. In rabbinic literature, Shavuot is called *atzeret* ("cessation, conclusion"), perhaps because the cessation of work is one of its distinctive features, or possibly because it was viewed as concluding the Passover season. In liturgical texts it is described as the "season of the giving of our Torah." The association of Shavuot with the revelation at Sinai, while not attested to in Scripture, is alluded to in the Pseudepigrapha (a collection of noncanonical writings). In rabbinic literature the association first appears in 2nd-century materials. The association, probably an ancient one, was derived in part from the book of Exodus, which dates the revelation at Sinai to the third month (counting from Nisan), *i.e.,* Sivan.

Scripture does not provide an absolute date for Shavuot. Instead, 50 days (or seven weeks) are reckoned from the day the sheaf offering (ʿOmer) of the harvest was brought to the Temple, the 50th day being Shavuot. According to the Talmudic rabbis, the sheaf offering was brought on the 16th of Nisan; hence Shavuot always fell on or about the 6th of Sivan. Jewish sectarians, such as the Sadducees, rejected the rabbinic tradition concerning the date of the sheaf ceremony, preferring a later date, and celebrated Shavuot accordingly.

In Temple times, aside from the daily offerings, festival offerings, and first-fruit gifts, a special cereal offering consisting of two breads prepared from the new wheat crop was offered at the Temple. Since the destruction of the Second Temple, Shavuot observances have been dominated by its commemorative aspect. Many Jews spend the entire Shavuot night studying Torah, a custom first mentioned in the *Zohar* ("Book of Splendour"), a Kabbalistic work edited and published in the 13th–14th centuries. Some prefer to recite the *tiqqun lel Shavuʿot* ("Shavuot night service"), an anthology of passages from Scripture and the Oral Law (Mishna) compiled in the late medieval period. An expanded liturgy includes Hallel, public readings from the Torah, *yizkor* (in many congregations), and *musaf.* The Book of Ruth is read at the synagogue service, possibly because of its harvest-season setting.

Sukkot (Tabernacles)

Sukkot ("booths"), an ancient harvest festival that commemorates the booths the Israelites resided in after the Exodus, was the most prominent of the three pilgrim festivals in ancient Israel. Also called Ḥag ha-Asif (Festival of Ingathering), it has retained its joyous, festive character through the ages. It begins on Tishri 15 and is celebrated for seven days. The concluding eighth day (plus a ninth day in the Diaspora), Shemini Atzeret, is a separate holiday. In Temple times, each day of Sukkot had its own prescribed number of sacrificial offerings. Other observances, recorded in the Mishna tractate *Sukka,* include the daily recitation of Hallel, daily circumambulation of the Temple altar, a daily water libation ceremony, and the nightly *bet ha-shoʿeva* or *bet ha-sheʾuvah* ("place of water drawing") festivities starting on the evening preceding the second day. The last mentioned featured torch dancing, flute playing, and other forms of musical and choral entertainment.

Ideally, Jews are to reside in booths—walled structures covered with thatched roofs—for the duration of the festival; in practice, most observant Jews take their meals in the *sukka* ("booth") but reside at home. A palm-tree branch (*lulav*), bound up together with myrtle (*hadas*) and willow (*ʿarava*) branches, is held together with a citron (*etrog*) and waved. Medieval exegetes provided ample (if not always persuasive) justification for the Bible's choice of these particular branches and fruit as symbols of rejoicing. The numerous regulations governing the *sukka, lulav,* and *etrog* comprise the major portion of the treatment of Sukkot in the codes of Jewish law. The daily Sukkot liturgy includes the recitation of Hallel, public readings from the Torah, the *musaf* service, and the circumambulation of the synagogue dais. On the last day of Sukkot, called Hoshana Rabba (Great Hoshana) after the first words of a prayer (hoshana, "save us") recited then, seven such circumambulations take place. Kabbalistic (mystical) teaching has virtually transformed Hoshana Rabba into a solemn day of judgment.

Hoshana Rabba is followed by Shemini Atzeret (Eighth Day of Solemn Assembly), which is celebrated on Tishri 22 (in the Diaspora also Tishri 23). None of the more distinctive Sukkot observances apply to Shemini Atzeret; but Hallel, public reading from the Torah, *yizkor* (in many congregations), *musaf,* and a prayer for rain in the Holy Land are included in its liturgy. Simḥat Torah (Rejoicing of the Law) marks the annual completion of the cycle of public readings from the Torah. The festival originated shortly before the gaonic period (*c.* 600–1050 CE) in Babylon, where it was customary to conclude the public readings annually. In Palestine, where the public readings were concluded approximately every three years, Simḥat Torah was not celebrated annually until after the gaonic period. Israeli Jews celebrate Simḥat Torah and Shemini Atzeret on the same day; in the Diaspora, Simḥat Torah is celebrated on the second day of Shemini Atzeret. Its joyous celebrations bring the Sukkot season to an appropriate close.

Ten Days of Penitence. The Ten Days of Penitence begin on Rosh Hashana and close with Yom Kippur. Already in Talmudic times they were viewed as forming an especially appropriate period of introspection and repentance. Penitential prayers (*seliḥot*) are recited prior to the daily morning service and, in general, during the period scrupulous observance of the Law is expected.

Rosh Hashana (New Year)

According to Mishnaic teaching, the New Year festival ushers in the Days of Judgment for all of mankind. Despite its solemnity, the festive character of Rosh Hashana is in no way diminished. In Scripture it is called "a day when the horn is sounded"; in the liturgy "a day of remembrance." In the land of Israel and in the Diaspora, Rosh Hashana is celebrated on the first two days of Tishri. Originally celebrated by all Jews on Tishri 1, calendrical uncertainty led to its being celebrated an additional day in the Diaspora and, depending upon the circumstances, one or two days in Palestine. After the calendar was fixed in 359, it was regularly celebrated in Palestine on Tishri 1 until the 12th century, when Provençal scholars introduced the two-day observance. Considerable speculation in recent literature concerning the origin of the Jewish New Year festival proves mostly that its early history can only be conjectured, not reconstructed.

Scroll of the Law being shown to the congregation on Simḥat Torah in a synagogue on the Tunisian island of Jazīrat Jarbah.

BBC Hulton Picture Library

The most distinctive Rosh Hashana observance is the sounding of the ram's horn (*shofar*) at the synagogue service. Medieval commentators suggest that the blasts acclaim God as Ruler of the universe, recall the divine revelation at Sinai, and are a call for spiritual reawakening and repentance. An expanded New Year liturgy stresses God's sovereignty, his concern for man, and his readiness to forgive those who repent. On the first day of Rosh Hashana (except when it falls on the Sabbath) it is customary for many to recite penitential prayers at a river, symbolically casting their sins into the river; this ceremony is called *tashlikh* ("thou wilt cast"). Other symbolic ceremonies, such as eating bread and apples dipped in honey, accompanied with prayers for a "sweet" and propitious year, are performed at the festive meals.

Yom Kippur (Day of Atonement)

The most solemn of the Jewish festivals, Yom Kippur is a day when sins are confessed and expiated and man and God are reconciled. It is also the last of the Days of Judgment and the holiest day of the Jewish year. Celebrated on Tishri 10, it is marked by fasting, penitence, and prayer. Work, eating, drinking, washing, anointing one's body, sexual intercourse, and donning leather shoes are all forbidden.

In Temple times, Yom Kippur provided the only occasion for the entry of the high priest into the Holy of Holies; details of the expiatory rites performed by the high priest and others are recorded in the Mishna and recounted in the liturgy. Present-day observances begin with a festive meal shortly before Yom Kippur eve. The Kol Nidre prayer (recited before the evening service) is a legal formula which absolves Jews from fulfilling solemn vows, thus safeguarding them from accidentally violating a vow's stipulations. The formula first appears in gaonic sources (derived from the Babylonian Talmudic academies, 6th–11th centuries) but may be older; the haunting melody that accompanies it is of medieval origin. Virtually the entire day is spent in prayer at the synagogue, the closing service (*neʿila*) concluding with the sounding of the ram's horn.

Minor festivals: Ḥanukka and Purim. Ḥanukka and Purim are joyous festivals lacking the work restrictions characteristic of the major festivals.

Ḥanukka (Feast of Dedication)

Ḥanukka commemorates the Maccabean (or Hasmonean) victories over the forces of the Seleucid king Antiochus IV Epiphanes (reigned 175–164 BCE), and the rededication of the Temple Kislev 25, 164 BCE. Led by Mattathias and his son Judah Maccabee, the Maccabees were the first Jews who fought to defend their religious beliefs rather than their lives. Ḥanukka is celebrated for eight days beginning on Kislev 25. The Ḥanukka lamp or candelabrum (*menora*), which recalls the Temple lampstand, is kindled each evening. One candle is lit the first evening; an additional candle is lit each subsequent evening until eight candles are lit on the last evening. According to the Talmud (Shabbat 21b), the ritually pure oil available at the rededication of the Temple was sufficient for only one day's light but miraculously lasted for eight days, hence the eight-day celebration of Ḥanukka. Evidence from the Apocrypha (writings excluded from the Jewish canon but included in the Roman Catholic and Eastern Orthodox canons) and rabbinic literature shows an association between Sukkot and Ḥanukka, possibly accounting for the latter's eight-day duration. Ḥanukka joy is expressed in festive meals, song, games, and gifts to children. The liturgy includes Hallel, public readings from the Torah, and the *ʿal ha-nissim* ("for the miracles") prayer. The Scroll of Antiochus, an early medieval account of Ḥanukka, is read in some synagogues and homes.

Purim (Feast of Lots)

As recorded in the biblical Book of Esther, Purim commemorates the delivery of the Persian Jewish community from the plottings of Haman, Ahasuerus' (perhaps Xerxes, king of Persia, 486–465 BCE) prime minister. Mordecai and his cousin Esther, the King's Jewish wife, interceded on behalf of the Jewish community, rescinded the royal edict authorizing a pogrom against the Jews, and instituted the Purim festival. The historicity of the biblical account is questioned by many modern scholars. It is now generally conceded that the Book of Esther was written in the Persian period (it contains Persian but not Greek words) and reflects Persian custom. Except for the Book of Esther, the earliest mention of the Purim festival is from the 2nd–1st centuries BCE. The name of the festival was derived from the Akkadian *pûru,* meaning "lot."

In most Jewish communities, Purim is celebrated on Adar 14 (some also celebrate it on the 15th, others only on the 15th). On the evening preceding Purim, men, women, and children gather in the synagogue to hear the Book of Esther read from a scroll (*megilla*). The reading is repeated Purim morning. A festive meal during the day is accompanied by much song, wine, and merriment. Masquerades, Purim plays, and other forms of parody are common. Friends exchange gifts of foodstuffs and also present gifts to the poor. Aside from the Esther readings, the liturgy includes public reading from the Torah and recital of the Purim version of the *ʿal hanissim* prayer.

The five fasts. The commemorative aspects of the fasts are bound up with their penitential aspects, all of which find expression in the liturgy. Thus the Jew not only relives the tragic history of his people with each fast, but is also afforded an opportunity to search within himself and focus on his own (and his people's) present and future. Penitential prayers (*seliḥot*) are recited on all fasts, and the Torah is read at the morning and afternoon services.

ʿAsara be-Ṭevet (Fast of Ṭevet 10) commemorates the beginning of the siege of Jerusalem by Nebuchadnezzar, king of Babylonia, in 588 BCE.

Shivaʿ ʿAsar be-Tammuz (Fast of Tammuz 17) commemorates the first breach in the wall of Jerusalem by the Romans in 70 CE. It initiates three weeks of semi-mourning that culminate with Tisha be-Av.

Tisha be-Av (Fast of Av 9) commemorates the destruction of the First and Second Temples in 586 BCE and 70 CE. The most solemn of the five fasts, its self-denials are more rigorous than those prescribed for the others, and, like Yom Kippur, the fast begins at sunset. The book of Lamentations is read at the evening service, followed by poetic laments that are also recited Tisha be-Av morning.

Tzom Gedaliahu (Fast of Gedaliah) commemorates the slaying of Gedaliah, governor of Judah after the destruction of the First Temple.

Taʿanit Esther (Fast of Esther), which commemorates Esther's fast (*cf.* Esther 4:16), is first mentioned in gaonic literature.

The lesser holidays. A major festival in the biblical period, Rosh Ḥodesh (First Day of the Month) gradually lost most of its festive character. Since Talmudic times, it has been customary to recite Hallel on Rosh Ḥodesh. In the medieval period, aside from the liturgical practices carried over from the Talmudic period, it was celebrated with a festive meal. Always more diligently observed in Palestine than in the Diaspora, attempts to revive its full festive character are being made in modern Israel.

First mentioned in the Mishna, where it marks the New Year for tithing purposes, Tu bi-Shevat (New Year for Trees) assumed a festive character in the gaonic period, and later in the medieval period it became customary to eat assorted fruits on the holiday. In modern times it is associated with the planting of trees in Israel.

Lag ba-ʿOmer (33rd Day of the ʿOmer Counting) is a joyous interlude in the otherwise somber period of ʿOmer counting (*i.e.,* of the 49 days to Shavuot), which is traditionally observed as a time of semi-mourning. Usually celebrated as a school holiday with outings, it is first mentioned in medieval sources, which attribute its origin to the cessation of a plague that was decimating the students of Akiba, an influential rabbinic sage in the 2nd century, and to the anniversary of the death of another great rabbi, Simeon ben Yoḥai (died *c.* 170 CE).

The situation today. Modern attitudes toward the Sabbath and festivals vary considerably. Acculturated Jews under the sway of Western secularism often are ignorant of, or choose to neglect, traditional observances. Attitudes of committed Jews in the Western world are patterned mostly along the lines of accepted Orthodox, Conservative, and Reform practice. Thus for example, driving to synagogue services on the Sabbath is unthinkable in Orthodox circles, a matter of dispute among Conservative rabbis, and normative practice for Reform Jews. Among Orthodox Jews, who best preserve the traditional observances, contemporary discussion centres mostly on technological advances and their effect on Halakhic practice (the behaviour laid down in the written and oral Torah). Whether or not hearing aids may be worn on the Sabbath, and how crossing the international dateline affects observance of Sabbaths and festivals typify the sort of problem raised in Orthodox *responsa* ("replies" to questions on law and observance). Recent (and often heated) discussion in Conservative literature raises the possibility of abolishing the obligatory character of the additional festival days in the Diaspora (except for the second day of Rosh Hashana), thus unifying Jewish practice throughout the world. Reform Jews, the most innovative of the three groups, observe neither the additional festival days (including the second day of Rosh Hashana) nor the fasts and have introduced numerous modifications in the liturgy as well as in the observances. In recent years more radical Reform congregations have experimented freely with "psychedelic" sound and light effects and other novel forms of synagogue service.

In Israel Sabbath is the national day of rest, and Jewish holidays are vacation periods. Municipal ordinances govern public observance of the Sabbath and festivals; their enactment and enforcement vary with the political influence of the local Orthodox Jewish community. Attempts to interpret festivals along nationalistic lines are common; some kibbutzim (communal farms) stress the agricultural significance of the festivals. Independence Day is a national holiday; the preceding day, Remembrance Day, commemorates Israel's war dead. Yom Hashoa (Holocaust Day)—marking the systematic destruction of European Jewry between 1933 and 1945 and recalling the short-lived Ghetto uprisings—is commemorated officially on Nisan

27; many religious Israelis prefer to commemorate it on Ṭebet 10 (a fast day) now called *yom ha-qaddish* (day upon which the mourner's prayer is recited). Since the June 1967 war, Iyyar 28—Liberation of Jerusalem Day—is celebrated unofficially by many Israelis. Appropriate services are conducted on all the aforementioned holidays by most segments of Israel's religious community.

In Israel and the Diaspora, Jewish theologians often stress the timelessness and contemporaneity of holiday observances. Nevertheless, "revised" Passover Haggadot (plural of Haggada) in which contemporary issues are accorded a central position, appear regularly.

Scholarly research into the origin of the festivals, if unabated, has not advanced significantly in recent years, nor is it likely to unless new evidence is forthcoming. Attempts to trace the development and spread of festival observances have fared better, and studies such as A. Yaari's *History of the Simḥat Torah Festival* (in Hebrew) bode well for the future. (S.Z.L.)

HOLY PLACES: THE LAND OF ISRAEL AND JERUSALEM

The Jerusalem Temple

The land of Israel, as is evident from the biblical narratives, played a significant role in the life and thought of the Israelites. It was the promised home, for the sake of which Abraham left his birthplace; the haven toward which moved the tribes who escaped from Egyptian servitude; the hope of the exiles in Babylon. In the long centuries following the destruction of the Judean state by the Romans, it remained inextricably bound up with messianic and eschatological expectations. During the early period of settlement, there seem to have been many sacred localities, with one or another functioning for a time as a central shrine for all of the tribes, without displacing the others. Even the establishment of Jerusalem as the political capital by David and the building of a royal chapel there by Solomon did not bring to an end local cult centres. It was not until the reign of Josiah of Judah (640–609) that a reform centralized the cult in Jerusalem and attempted—although not entirely successfully—to end worship at local shrines. However irregular was the effectiveness of this reform, the Babylonian Exile and the subsequent return saw Jerusalem and its Temple win out over its rivals and become—in law, in fact, and in sentiment—the centre of Jewish cultic life. As noted above, this did not inhibit the rise and development of other forms of worship and even—on a few occasions—other cult centres. Nonetheless, no matter how unpopular the priesthood of the Jerusalem Temple became with some segments of the population—the Qumrān community seems to have denied its legality, and the Pharisees complained bitterly about its arrogance and exactions, attempting when politically feasible to impose and enforce Pharisaic regulations upon it—reverence for the Temple itself seems to have remained a widespread sentiment. With the destruction of the Temple by the Romans in 70 CE, such reverence was transformed both by messianic expectations and eschatological hopes into fervent devotion, which, over the following centuries, became idealized and even supernaturalized. The most ardently articulated statement of the crucial role of the land of Israel and the Jerusalem Temple is found in the *Sefer ha-Kuzari* of Judah ha-Levi in which the two are seen as absolutely indispensable for the proper relation between the people Israel and God. Symbolizing the significance of the land and of the city is the practice of facing in their direction during worship. The earliest architectural evidence derived from synagogue remains in Galilee indicates that the attempt was made to arrange the building in such a way that the worshippers faced directly toward Jerusalem. This practice may have continued even in the Diaspora, but at a later date the present practice of setting the holy ark in or before the east wall was established, so that "facing Jerusalem" is now more symbolic than actual.

THE SACRED LANGUAGE: HEBREW AND THE VERNACULAR TONGUES

Use of the vernacular in worship

The transformation of Hebrew into a sacred language is, of course, bound up with the political fate of the people. In the period following the return from the Babylonian Exile, Aramaic, a cognate of Hebrew, functioned as the international or imperial language in official life and certainly gained a foothold as a vernacular. It did not, despite claims made by some scholars, displace the everyday Hebrew of the people. The language of the Mishna, far from being a scholar's dialect, seems to reflect—in the same way as the Koine (common) Greek of the New Testament—popular speech. Displacement of Hebrew—both in its literary form in Scriptures and in its popular usage—did take place in the Diaspora, however, as evidenced by the need to translate Scriptures into Greek in some communities and into Aramaic in others. As far as the emerging order of worship is concerned, there seems also to have been an inclination on the part of some authorities to permit even the recitation of the Shema complex in the vernacular. Struggles over these issues within the communities continued for a number of centuries in various places, but the development of formal literary Hebrew—a sacred tongue, to be used side by side with the Hebrew Scriptures in worship—brought them to an end. Although the communities of the Diaspora used the vernaculars of their environment in day-to-day living and even—as in the case of the communities of the Islāmic world—for philosophical, theological, and other scholarly writings, in worship, Hebrew remained the standard until modern times when some of the reform movements in western Europe sought partially—and a very small fraction even totally—to displace it.

THE RABBINATE

Legal, judicial, and congregational roles. The rabbinate, with its peculiar nature and functions, is the result of a series of developments going back to the period that followed the disastrous second revolt against Rome (132–135 CE). The term rabbi ("my teacher") was originally an honorific title for the graduates of the academy directed by the *nasi* or patriarch, the head of the Jewish community in Palestine in that era, who was also a Roman imperial official. The curriculum of the school was Torah, written and oral, according to the Pharisaic tradition and formulation. The *nasi* appointed rabbis to the law court (the Bet Din) and as legal officers of local communities: acting with the local elders, they supervised and controlled the life of the community and its members in all of its aspects. A similar situation obtained in Babylon under the Parthian and Sāsānian empires, where the *resh galuta* or exilarch ("head of the exile") appointed rabbinical officials to legal and administrative posts. In time the patriarchate and exilarchate disappeared, but the rabbinate, nourished by independent rabbinical academies, survived. An authorized scholar, when called to become the judicial officer of a community, would at the same time become the head of the local academy and would, after adequate preparation and examination, grant authorization to his pupils, who were then eligible to be called to rabbinical posts. There was, thus, a diffusion of authority, the communities calling, rather than a superior official appointing, their rabbis. What must be kept in mind is that these rabbis were not ecclesiastical personages but communal officials, responsible for the governance of the entire range of life of what was understood to be the *qehilla qedosha,* the "holy community."

The modern rabbi

In modern times and particularly in the Western world, the total change in Jewish communal existence required a transformation of this ancient structure. The rabbinate became, for the most part, an ecclesiastical rather than a communal agency, reflecting the requirements of civic life in modern national states. The education of rabbis who now function within this new situation is carried on in seminaries whose structure and curriculum have been influenced by European and American academic institutions. The majority of their graduates serve as congregational rabbis, in roles similar to those of ministers and priests in the Christian denominations, but with some other functions deriving from the particular situation and nature of the Jewish community.

Even in the State of Israel, where certain larger areas, such as that of family law, are still reserved to the rabbinate, it nonetheless functions more as a counterpart to other ecclesiastical organizations, Christian and Muslim, than as an overarching and all-inclusive communal agency that embodies, as in the past, involvement in every aspect of community and personal life.

Chief rabbinates. The existence of the offices of chief rabbi in the State of Israel derives from the situation in the Turkish Empire when the various religious communities functioned as quasi-political entities in that multiethnic conglomerate. Israel has two chief rabbis, one for the Ashkenazic (European) and one for the Sefardic (Oriental) communities—they no longer function, however, as the heads of whole communities but only of ecclesiastical organizations. The same is true in those countries outside Israel that have the office of chief rabbi; *e.g.,* Great Britain and France. Here they function vis à vis the governments like their ecclesiastic counterparts in the Christian churches. While they have certain kinds of limited authority because of their official position, their jurisdiction extends only over those members of the total Jewish community who are ready to accept it; others form their own ecclesiastical units and act without reference to the chief rabbinate. In some situations, particularly in the United States where there is no similar structure, the title chief or grand rabbi has been assumed occasionally by individuals as the means of asserting superior dignity or even (fruitlessly) authority.

GENERAL COUNCILS OR CONFERENCES

The precise nature of the Sanhedrin (Council Court) in the last years of the Jewish commonwealth is a much disputed matter. The several councils mentioned in Talmudic literature are equally difficult to define with exactitude. There are references scattered throughout medieval literature that suggest the existence of councils and synods but their composition and authority are also uncertain. Around 1000 a synod was held in the Rhineland in which French and German communities participated under the guidance of Rabbenu Gershom, the leading rabbinic authority of the region. The late Middle Ages saw the rise in eastern Europe of the Waʿad Arbaʿ Aratzot (Council of the Four Lands) composed of communal representatives from Great Poland, Little Poland, Russian Poland (Volhynia), and Lithuania. At the beginning of the modern era Napoleon (1806) summoned an Assembly of Notables—representatives of communities under French dominion—to deal with questions arising from the dissolution of the older status of the Jews and their naturalization as individuals into the new national states. Those decisions of the Assembly that involved questions of Jewish law were subsequently submitted to a Grand Sanhedrin called into being by Napoleon to provide some sort of Halakhic justification for the acts the French imperial government had required of the Jewish communities. During the 19th century the demand for the reform of Jewish life—principally the liturgy of the synagogue, but many other aspects as well—evoked a series of rabbinical conferences and synods that debated the questions and sought to guide the changes thought to be necessary. A similar procedure was followed on the American scene. In both instances, after an initial period in which radicals, moderates, and conservatives argued their respective cases in the same forum, polarization set in and intellectual differences were transformed into competing organizations. In the 1970s the several tendencies within the Jewish communities in North America were institutionalized in rabbinical conferences and congregational unions—Orthodox, Conservative, and Reform—whose influence was in large measure limited to their adherents. In the United States the Synagogue Council of America claims to be the "united voice" of American Jewry in common concerns. There is also a worldwide body in Reform or Liberal Judaism—the World Union for Progressive Judaism.

MODERN VARIATIONS

Reinterpretations of the Halakhic system

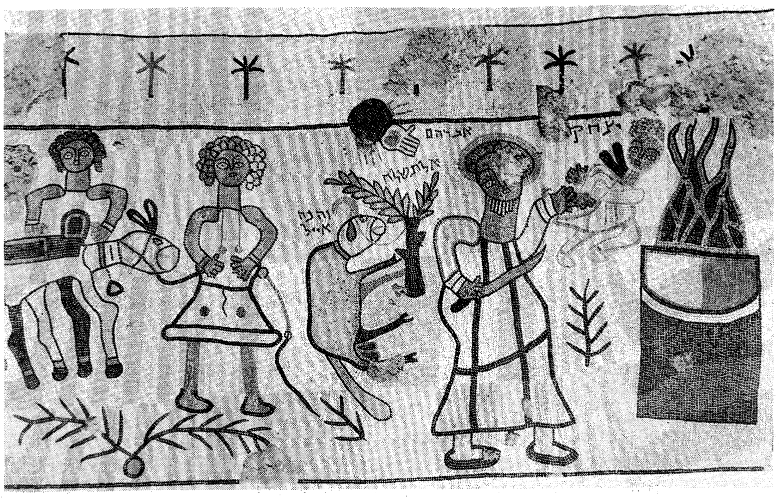
The above sketch of basic practices and institutions has attempted to describe the so-called traditional situation, although it has been indicated that even here there are variations—actually more than have been noted. In addition, reference has been made to some shifts and changes that represent a giving up of traditional practices on the basis of intellectual decisions about the nature of Judaism, its beliefs, practices, and institutions. Such changes are far too numerous to describe in detail. What is more important is to indicate their motivation. Basically, it is the view that the Halakhic system is not, as a whole and in all of its parts, divinely revealed but is rather a human process that seeks to expose in mutable forms the meaning of the divine-human encounter. Thus viewed, the practices and institutions are understood to be historically determined, reflecting the multifaceted experience of the people Israel as it has sought to live in the presence of God. Historical scholarship has, from this point of view, disclosed the origins, rise, development, and decline of these structures in the past and thus authorizes such changes in the present and future as appear to fulfill the needs of the community and its members. An examination of the specific deviations from the traditional forms makes clear that the application of this position, or attitude, has been subject to wide variation during the 19th and 20th centuries, in which it has operated. Some have seen it as a call for the disengagement from much if not all of the traditional pattern, and a recognition that only the spiritual essence is of importance or consequence for Judaism. Others have argued that an indiscriminate use of historicism (the explanation of values and forms in terms of their historical conditions) is unjustified and that the burden of proof is always upon those who would introduce changes. In the post-World War II period, the question has been whether a reconstituted Halakhic system might not be a requirement of the day.

Art and iconography

The synagogue at Doura-Europus

The anti-iconic principle and its modifications. Although the Second Commandment (Ex. 20:4; Deut. 5:8), "You shall not make yourself a graven image, or any likeness of anything that is in heaven above, or that is in the earth beneath, or that is in the water under the earth," has indeed been understood as absolutely prohibiting any and all artistic representation, this is not the only way in which these words may be interpreted. What is intended is a prohibition against the construction of such likenesses as were the object of worship in the cultural area in which the Israelites dwelt. Even in the Bible there are reports of artistic productivity in the construction of the tent sanctuary and its ritual vessels (Ex. 25–31) and of the Temple in Jerusalem (I Kings 6–7). The literalness and rigour with which the commandment was interpreted depended upon the larger situation of the community, so that during periods of external pressures toward religious conformity, such as the reign of Antiochus IV Epiphanes in Antioch (175–164 BCE), the anti-iconic attitude sharpened. Similarly, during the Roman occupation, the presence of the battle standards of the legions with their animal representations was looked upon as an affront, while extreme Pietists would not even handle Roman coinage because of the images stamped on it. On the other hand, the walls of a 3rd-century-CE synagogue in Doura-Europus in Syria are covered from floor to ceiling with biblical scenes with human representations, and a number of synagogues in Palestine had elaborate mosaic floors with the signs of the zodiac, representations of the seasons, and the like. Further, illuminated manuscripts from the medieval period in Europe were frequently decorated with biblical figures, some quite clearly copied from Christian prototypes. A fascinating mediating position is to be seen in a Haggada, in which the human figures have bird heads. Synagogues from a later, although preemancipation, period (before the 18th century) were often decorated with animal figures. In the modern period the avoidance of human figures has not been entirely accomplished, although nothing like the decorations of Doura-Europus has appeared.

Ceremonial objects and symbols. Nonetheless, given this general anti-iconic attitude, much of Jewish artistic endeavour has been directed toward the creation of ceremonial objects: Qiddush goblets, candlesticks and candelabra, spice boxes for the Havdala ceremony at the end of the Sabbath, ornamented containers for the *mezuza* (a



Sacrifice of Isaac, detail of a mosaic from the synagogue of Bet Alfa, Jezreel, Israel, 6th century AD.

Picture from the photographic archive of the Jewish Theological Seminary of America, New York. Frank J. Darmstaedter.

parchment on which is written Deut 6:4–9 and 11:13–21, fastened to the doorpost on the right side as one enters), the silver crowns placed on the Torah scrolls, together with the mantles and breastplates for the same, and many other objects designed to embellish the performance of the large number of ritual acts of the individual and the community. All of these vary in artistic quality, from the work of simple artisans to exquisitely produced works of master craftsmen.

Architecture. The building of synagogues, too, is an expression of artistic interest and concern, as well as of religious and social function. Nothing is known of these edifices, if indeed there were any, until the Greco-Roman period. Then the Roman basilica often provided the appropriate model. What was required was a spacious hall for assembly, and galleries for the women, and this form served that purpose very well. However synagogues were furnished before the destruction of the Second Temple, after that event some attempt seems to have been made to transfer some of the latter's appurtenances to the former, a move that was successfully resisted. When possible, the synagogue stood on a hill. Before it stood a walled entrance court with a fountain for ablutions. Before the Temple destruction, the building may have been oriented with its doors facing eastward, but afterward they faced Jerusalem; still later, when the holy ark containing the Torah scrolls was placed in a fixed position, the orientation was reversed so that the central gate would not be blocked; ultimately, the ark was placed in or against the east wall, without reference to the actual direction of Jerusalem. As the Diaspora grew larger, the new communities adapted the architectural forms of the enveloping culture. The surviving buildings of the Muslim period in Spain are often built with the horseshoe arches and decorated with the exquisite stucco arabesques that mark the era. The medieval period in Christian Europe saw a revival of a very strict anti-iconic attitude and a gradual rejection of the church edifice in favour of secular buildings as a model for the synagogue. The increasingly limited role of the Jew in that society and the enlargement of restrictions by church and state made it necessary to modify the synagogal structure. The doors no longer were in the wall facing the ark; the courtyard grew smaller; galleries were discontinued (side rooms now serving as the women's section); and a double- rather than a triple-aisled construction was largely favoured. Similar developments took place in eastern Europe with the building of fortress-synagogues and the remarkable wooden synagogues of Poland. In the early postemancipation period, Baroque style had its day, followed by Greek temples, Romanesque, Gothic, and pseudo-Byzantine churches, and pseudo-Moorish mosques. In the most recent period, the various schools of functionalism and their commercial descendants have come to the fore. The best of these have brought together fine architectural design and beautifully conceived and executed decoration. The interior arrangement, even in some traditional synagogues, has been influenced by the Protestant sermon-centred form of worship, so that some of the unique forms that marked older structures are absent. The holy ark is, however, still a centre of attention and has often been treated in interesting and striking ways.

Synagogal sites and structures

Paintings and illustrations. As noted above, the use of paintings in the decoration of synagogues goes back to at least the 3rd century CE and is found in the late pre-emancipation and modern synagogues as well. Manuscripts, too, were illuminated with miniatures and the Renaissance period saw the appearance of beautifully decorated Scrolls of Esther and *ketubbot* (marriage contracts). Nonetheless, the appearance of Jewish artists in painting and sculpture is a modern phenomenon. Beginning in the 19th century, interest grew apace and more and more Jews are to be found, often in the avant-garde of these fields. Some, such as Marc Chagall and Jacques Lipchitz, have done specifically religious art.

Liturgical cantillation

Music. The description of the synagogue service above noted the role of the *ḥazzan,* or cantor. It is he who reads the service and declaims the scriptural lessons to certain set musical modes that vary with the season and occasion. Many of these call for melodic responses on the part of the congregation. The origins and varying developments

Picture from the photographic archive of the Jewish Theological Seminary of America, New York. Frank J. Darmstaedter

Pewter Passover plate, German, 17th century. Adam and Eve are represented in the centre, surrounded by the signs of the zodiac and a depiction of the sacrifice of Isaac. The Hebrew lettering on the rim indicates the order of the Passover eve service. In the Jewish Museum, New York.

of these chants are ancient, often obscure, and equally complicated. Whatever the basic materials, these were enlarged, varied, corrupted, and reworked over the centuries in the various environments in which the Jewish communities have lived. In modern times musicologists have begun to examine with great care the history of synagogal music, analyzing its basic structures and its relationship to the music of Christian liturgical traditions. In the 19th century in Western Europe much of the traditional music was either discarded or re-worked under the influence of western forms and styles. In addition the pipe-organ was introduced and was the centre of stormy controversy.

Literature. Literature has been throughout the ages the home of Jewish artistic activity. The Hebrew Bible is a work of monumental artistry, exhibiting grandeur of form and language in historical narrative, poetry, rhetoric, and aphorism. The extrascriptural writings of the period, although their originals have often vanished, still disclose literary genius of a high order in translation. The documents of the rabbinic tradition are not often looked at with an eye to their literary worth but much of the material, particularly the Haggadic portions of the Midrashim, reveals a noteworthy sensitivity to the uses of language. In the medieval period much attention was given to the production of *piyyuṭim,* liturgical poetry with which to embellish the *Siddur* (prayer book), itself a collection containing much imaginative, as well as pedestrian, writing. In the Islāmic world, under the influence of Arabic poetry, Hebrew poetry rose to a high peak in both liturgical and secular forms. The Middle Ages in the Rhineland also saw the beginnings of the Jewish form of Middle High German that was, over the centuries, to develop into an autonomous Jewish language, Yiddish, which, in the 19th century, became a literary vehicle of very high order. The same period saw the beginnings of the recreation of Hebrew into a literary language that has become the basis of the spoken vernacular of the State of Israel and of a flourishing literature encompassing every branch of the field. Since the emancipation at the end of the 18th century, Jews in western Europe and later in the United States have turned to literature in the vernaculars of their countries, and have produced writers of note dealing with both Jewish and general themes. (L.H.S.)

Jewish philosophy

The term Jewish philosophy refers to various kinds of reflective thought engaged in by persons identified as being Jews, in one sense or another. At times, as in the Middle Ages, this meant any methodical and disciplined thought, whether on general philosophical subjects or on specifically Judaic themes, pursued by Jews. In other eras, as in modern times, concentration on the latter has been considered a decisive criterion, so that philosophers who are Jewish but unconcerned with Judaism or the Jewish heritage and destiny in their thought are not ordinarily classified as Jewish philosophers.

PRE-HELLENISTIC AND HELLENISTIC THOUGHT

Bible and Apocrypha. Philosophy arose in Judaism under Greek influence; however, a kind of philosophical approach may be discerned in early Jewish religious works apparently subject to little or no Greek influence. The books of Job and Qohelet (Ecclesiastes) were favourite works of medieval philosophers, who took them as philosophical discussions untinged by theological preconceptions. The book of Proverbs introduces, in an apparently theological context, the concept of Wisdom (Ḥokhma), which was to have a primordial significance for Jewish philosophical and theological thought, and presents it as the first and favourite of God's creations. It is also praised in the book of the Wisdom of Jesus the Son of Sirach (Ecclesiasticus) as instilled by God into all his works and granted in abundance to those he loves. It is sometimes equated with fearing God and keeping the Law; however, in other passages piety seems to be regarded as superior to Wisdom. The Wisdom of Solomon, probably originally written in Greek, praises Wisdom, which is held to be an image of God's goodness and a reflection of the eternal light. God is said to have given the author knowledge of the composition of the world, of the powers, the elements, the nature of animals, the divisions of time, and the positions of the stars. In its vocabulary and perhaps in some of its doctrines, the work shows the influence of Greek philosophical conceptions. It has had considerable influence on Christian theology.

The concept of Wisdom

Philo Judaeus. The first systematic attempt to apply Greek philosophical concepts to Jewish doctrines was made by Philo Judaeus (Philo of Alexandria) in the 1st century CE. Philo, a scholar who combined Greek and Jewish learning, was influenced by Platonic and Stoic writings, and probably also by certain postbiblical Jewish beliefs and speculations. He apparently had some knowledge of the Oral Law, which was being evolved in his time, and he also knew of the Essenes, a contemporary rigorous sect, whom he praised highly.

Philo's main contribution was interpretative. He provided Jewish conceptions with the hallmark of intellectual and cultural respectability by stating them in Greek philosophical terms; he also showed that many Greek notions were consonant with Jewish doctrine, as he conceived it, and with the allegorical sense of biblical texts, as he read them. He had two schemes of reference—Jewish religious tradition and Greek philosophy—and the fact that he took care to stress the primacy of the former may have been more than mere lip service. It may be argued with some plausibility that in central points of his thought, such as his conception of the Logos (the divine Reason or Word), Philo used philosophical notions as trappings for an originally religious belief. A main function of the Logos in his thought is to serve as an intermediary between the transcendent, unknowable God and the world, a view that probably has a close connection with the view of his Jewish contemporaries concerning the world of God, by means of which he accomplishes his designs. On basic philosophical or theological problems, such as the creation of the world or the freedom of will, Philo's writings provide either vague or contradictory answers. He placed mystic ecstasy, of which he may have had personal experience, above philosophical and theological speculations.

The concept of Logos

Philo's approach, his method of interpretation, and his way of thinking, as well as some of his conceptions—primarily that of the Logos—exerted a considerable influence on early Christian thought but not, to any comparable extent, upon Jewish thought in that period. Later, in the Middle Ages, knowledge of Philo among Jews was either very slight or nonexistent. Not until modern times was his role in Jewish religious thought recognized.

Other ancient sources. Some traces of a knowledge of popular, mainly Stoic philosophy may be found in the Mishna, a codification of the Oral Law composed in Palestine in the 2nd century CE, and in the subsequent Talmudic literature set down in writing in Palestine and Babylonia. On the whole these traces are rather slight. Some scholars believe that the influence of Greek philosophy on Palestinian Jewry was far-reaching, but the case, to say the least, is not proven. Jewish theological and cosmological speculations occur in the Midrashim (plural of Midrash,), which, under the guise of interpreting biblical verses, propound allegorical interpretations, legends, and myths, and in the *Sefer Yetzira* ("Book of Creation"), a work that is a combination of a cosmogony and a grammar and that was fictitiously attributed to Abraham. There is no clear evidence of the period in which it was written; both the 3rd century and the 6th or 7th century have been suggested. The book became a key work in later Jewish mysticism.

MEDIEVAL PHILOSOPHY

Islāmic background

In the 9th and 10th centuries, after a long hiatus, systematic philosophy and ideology reappeared among Jews, a phenomenon indicative of their accession to Islāmic civilization. The evolution of Islām in the 9th and 10th centuries showed that Greek scientific and philosophic lore could be separated, at least to some extent, from its pagan associations and could be transformed into another language and another culture; it also tended to show that a culture in which the sciences and philosophy or the sciences and theology or both of these combinations were an indispensable part could be based upon a monotheistic prophetic religion that in all relevant essentials, including adherence to a basic religious law, was closely akin to Judaism. The question of whether philosophy is compatible with religious law (the answer being sometimes negative) constituted the main theme of the foremost medieval Jewish thinkers. From approximately the 9th to the 13th centuries Jewish philosophical and theological thought participated in the evolution of Islāmic philosophy and theology and manifested only in a limited sense a specifically Jewish character. Jewish philosophers showed no particular preference for philosophic texts written by Jewish authors over those composed by Muslims, and in many cases the significant works of Jewish thinkers constitute a reply or a reaction to the ideas of Islāmic philosophic and scientific writings.

Jewish kalām. Although several Jewish intellectuals in 9th–10th-century Babylonia were steeped in Greek philosophy, the most productive and influential Jewish thinkers of this period represented a very different tendency, that of the Muʿtazilite *kalām. Kalām* (literally, "speech") is an Arabic term used both in Islāmic and in Jewish vocabulary to designate several theological schools that were ostensibly opposed to Greek, particularly Aristotelian, philosophy. The Aristotelians, both Islāmic and Jewish, regarded *kalām* theologians (called the *mutakallimūn*) with a certain contempt, holding them to be mere apologists, watchdogs of religion, and indifferent to truth. Herein they did not do justice to their adversaries, for many representatives of the Muʿtazilite school of *kalām,* formed in the 8th century, displayed a genuine speculative impulse. Its theology, forged in disputes with Zoroastrians, Manichaeans, and Christians, claimed to be based on reason.

Saʿadia ben Joseph. This belief in reason, as well as some of the tenets of Muʿtazilite theology, were taken over by Saʿadia ben Joseph, who was also influenced, either directly or through the intermediary of an Arabic philosopher, by the arguments of a Christian 6th-century philosopher, John Philoponus, against certain Aristotelian and Neoplatonic positions. Saʿadia's main theological work, *Kitāb al-amānāt wa al-iʿtiqādāt (Beliefs and Opinions),* is modelled on similar Muʿtazilite treatises and on a Muʿtazilite classification of theological subject matter known as the Five Principles. Like many Muʿtazilite authors, Saʿadia starts out by setting forth in his introduction a list and theory of the various sources of knowledge.

Saʿadia's four sources of knowledge

Saʿadia distinguished four sources: (1) the five senses, (2) the intellect, or reason, (3) necessary inferences, and (4) reliable information given by trustworthy persons. In Saʿadia's sense of the word, intellect, or reason (*al-ʿaql*), means first and foremost an immediate, a priori cognition, independent of sense experience. In *Beliefs and Opinions* the intellect is characterized as having immediate ethical cognitions—that is, as discerning what is good and what is evil—in opposition to the medieval Aristotelians, who did not regard even the most general ethical rules as a priori cognitions. The third source of knowledge comprises inferences of the type "if there is smoke, there is fire," which are based on data furnished by the first two sources of knowledge. The fourth source of knowledge is meant to validate the teachings of Scripture and of the religious tradition. Teachings of Scripture must be held true because of the trustworthiness of the men who propounded them. One of the main purposes of the work is to show that the knowledge deriving from the fourth source concords with that discovered by means of the other three, or, in other words, that religion and human reason agree.

Saʿadia opposed Aristotle's view that the natural order was eternal. He held, with other partisans of the Muʿtazilite *kalām,* that the demonstration of the temporal creation of the world must precede and pave the way for the proof of the existence of God the Creator. Given the demonstrated truth that the world has a beginning in time, it can be proved that it could have been produced only through the action of a creator. It can further be proved that there can have been only one Creator.

Divine justice and human ethical judgment

The theology of Saʿadia, like that of the Muʿtazilites, hinges on two principles: the unity of God and the principle of justice. The latter principle takes issue with the view (widespread in Islām and present also in Judaism) that the definition of what is just and what is good depends solely on God's will, to which none of the moral criteria found among men is applicable; according to this view, a revelation from God can convert an action now generally recognized as evil into a good action. Against this way of thinking, Saʿadia and the Muʿtazilites believed that being good and just or evil and unjust are intrinsic characteristics of human actions and cannot be changed by divine decree. The notions of justice and of good, as conceived by man, are binding on God himself. Since, according to Saʿadia, man has a priori knowledge of good and evil, just and unjust, the fact that human ethical judgments are valid for God means that man's ethical cognitions are also those of the Deity.

The function of religious law—of central importance in traditional Judaism and Islām—is to impose on man the accomplishment of good actions and to prohibit bad ones. Because Saʿadia believed that man has a priori knowledge of good and evil and that this knowledge coincides with the principles underlying the most important portions of the revealed law, he was forced to ask the question whether this law is not superfluous. He could, however, point out that, whereas the human intellect recognizes that certain actions—for instance, murder or theft—are evil, it cannot by itself discover the best possible definition of what constitutes a particular transgression, nor can it, on its own, determine the punishment appropriate for a transgression. On both points, Saʿadia asserted, the commandments of religious law give the best possible answers.

The commandments that accord with the behests of the human intellect were designated by Saʿadia as the intellectual, or rational, commandments. According to him, they include the duty of manifesting gratitude to the Creator for the benefits he has bestowed upon man. Saʿadia recognized that a considerable number of commandments—for instance, those dealing with the prohibition of work on the Sabbath—do not belong to this category. He held, however, that the obligation to obey them may be derived from the rational commandment that makes it incumbent upon man to be grateful to God, for such gratitude entails obedience to his orders.

Karaite rationalism and atomism

The Karaites. Saʿadia's adoption of the rational Muʿtazilite theology was a part of his overall activity, directed toward the consolidation of rabbinical Judaism (based on the Mishna and Talmud), which was being attacked by the Karaites. This Jewish sect, founded by Anan ben David in the 8th century, rejected the authority of the Oral Law—that is, of the Mishna and the Talmud. In the 10th century and afterward, the Karaites accepted as their guides the Bible (Old Testament) and human reason, in the Muʿtazilite sense of the word. Their professed freedom from any involvement with postbiblical Jewish religious tradition facilitated a rational approach to theological doctrine. This approach led the Karaite authors to criticize their opponents, the adherents of rabbinical Judaism, for holding anthropomorphic beliefs based, in part, on texts of the Talmudic period. Karaite authors propounded, in conceptual terms, a theology of Jewish history in exile (*galut*). Life in exile is a diminished existence; nevertheless, the good or bad actions of the Jewish people (rather than their material strength or weakness) affect the course of history. Redemption may come when all Jews are converted to Karaism.

The Karaites adopted, wholesale, Muʿtazilite *kalām,* including its atomism. The Muʿtazilite atomists held that everything that exists consists of minute, discrete parts. This applies not only to bodies but also to space, to time, to motion, and to the "accidents"—that is, qualities, such as colour—which the Islāmic and Jewish atomists regarded as being joined to the corporeal atoms (but not determined by them, as had been believed by the Greek Atomists). An instant of time or a unit of motion does not continue the preceding instant or unit. All apparent processes are discontinuous, and there is causal connection between their successive units of change. The fact that cotton put into fire generally burns does not mean that fire is a cause of burning; rather, it may be explained as a "habit," signifying that this sequence of what is often wrongly held to be cause and effect has no character of necessity. God's free will is the only agent of everything that occurs, with the exception of one category—human actions. These are causes that produce effects; for instance, a man who throws a stone at another man, who is then killed, directly brings about the latter's death. This inconsistency on the part of the theologians was necessitated by the principle of justice, for it would be unjust to punish a man for a murder that was a result not of his action but of God's. This grudging admission that causality exists in certain strictly defined and circumscribed cases was occasioned by moral, not physical, considerations.

Jewish Neoplatonism. *Isaac Israeli.* Outside Babylonia, philosophical studies were pursued by Jews in the 9th and 10th centuries in Egypt and in the Maghrib (northwest Africa). The outstanding figure was Isaac ben Solomon Israeli, an Egyptian-born North African who has been called "the first Jewish Neoplationist." In his philosophical works, such as the "Book of Elements" and the "Book of Five Substances," he drew largely upon a 9th-century Muslim popularizer of Greek philosophy, Abū Yūsuf Yaʿqūb al-Kindī, and also, in all probability, upon a lost pseudo-Aristotelian text. The peculiar form of Neoplatonic doctrine that seems to have been set forth in this text had, directly and indirectly, a considerable influence on medieval Jewish philosophy.

According to Israeli, God creates through his will and power. The two things that were created first were form, identified with wisdom, and matter, which is designated as the genus of genera (the classes of things) and which is the substratum of everything, not only of bodies, but also of incorporeal substances. This conception of matter seems to derive from the Greek Neoplatonists Plotinus and Proclus, particularly from the latter. In Proclus' opinion generality was one of the main criteria for determining the ontological priority of an entity (relative place in the levels of reality). Matter, because of its indeterminacy, obviously has a high degree of generality; consequently, it figures among the entities having ontological priority. According to the Neoplatonic view, which Israeli seems to have adopted, the conjunction of matter and form gives rise to the intellect. A light sent forth from the intellect produces the rational soul, and in its turn it gives rise to the vegetative soul.

Israeli's doctrine of prophecy seems to be the earliest Jewish philosophical theory attributing prophecy to the influence of the intellect on the imaginative faculty. According to Israeli, this faculty receives from the intellect spiritual forms that are intermediate between corporeality and spirituality. This explanation implies that these forms, "with which the prophets armed themselves," are inferior to purely intellectual cognitions.

Solomon ibn Gabirol. In essentials the schema of creation and emanation propounded by Isaac Israeli and his Neoplatonic source or sources was taken over by Solomon ibn Gabirol, a celebrated 11th-century Hebrew liturgical poet, who seems to have been the earliest Jewish philosopher of Spain. His chief philosophical work, "Fountain of Life," written in Arabic, has been preserved in full only in a 12th-century Latin translation entitled *Fons vitae.* This work, which makes no reference to Judaism or to specifically Jewish doctrines, is a didactic dialogue between a disciple and a master who teaches him true philosophical knowledge. Despite its prolixity and many contradictions, it is an impressive work. Few medieval texts so effectively communicate the Neoplatonic conception of the existence of a number of planes of being that differ according to their ontological priority, the derivative and inferior ones constituting a reflection in a grosser mode of existence of those that are prior and superior.

Ibn Gabirol's doctrine of divine will

A central conception in Ibn Gabirol's philosophy is concerned with the divine will, which appears to be both part of and separate from the divine essence. Infinite according to its essence, the will is finite in its action. It is described as pervading everything that exists and as being the intermediary between the divine essence and matter and form. Will was one of a number of traditional appellations applied in various medieval theologies to the entity intermediate between the transcendent Deity and the world or to the aspect of the Deity involved in creation. According to a statement in *Fons vitae,* matter derives from the divine essence, whereas form derives from the divine will. This suggests that the difference between matter and form has some counterpart in the Godhead and also that universal matter is superior to universal form. Some of Ibn Gabirol's statements seem to bear out the impression of superiority of universal matter; other passages, however, appear to imply a superiority of universal form.

Form and matter, whether they be universal or particular, exist only in conjunction. All things, with the sole exception of God, are constituted through the union of the two, the intellect no less than the corporeal substance. In fact, the intellect is the first being in which universal matter and form are conjoined. The intellect contains and encompasses all things. It is through the grasp of the various planes of being, through ascending in knowledge to the world of the intellect and apprehending what is above it—the divine will and the world of the Deity—that man may "escape death" and reach "the source of life."

Judah ha-Levi. Judah ben Samuel ha-Levi (*c.* 1075–*c.* 1141), also from Spain and a celebrated Hebrew poet, was the first medieval Jewish thinker who consciously and consistently based his thought upon arguments drawn from Jewish history. His views are set forth in an Arabic dialogue, *al-Hazari* (Hebrew *Sefer ha-Kuzari*), the full title of which is translated as "The Book of Proof and Argument in Defense of the Despised Faith." This work is usually called *Kuzari*; *i.e.,* "the Khazar."

The *Kuzari*

Basing his narrative on the historical fact that the Khazars (a Turkic-speaking people in Central Eurasia) were converted to Judaism (*c.* 740), ha-Levi relates that their King, a pious man who did not belong to any of the great monotheistic religions, dreamed of an angel, who said to him, "Your intentions are pleasing to the Creator, but your works are not." To find the correct way of pleasing God, the King seeks the guidance of a philosopher, a Christian, a Muslim, then, finally, after hesitating to have recourse to a representative of a people degraded by its historical misfortune, of a Jewish scholar who converts him to Judaism. The words of the angel heard in a dream may, in accordance with both religious and philosophical doctrine, be regarded as a kind of revelation. The use of this element of the story enabled ha-Levi to suggest that it is not the spontaneous activity of human reason that impels man to undertake the quest for the true religion; for this, one needs the gift of prophecy, or, at least, a touch of the prophetic faculty (or a knowledge of the revelations of the past).

The argument of the philosopher whose advice is sought by the King brings this point home. This disquisition is a brilliant piece of writing, for it lays bare the essential differences between the Aristotelian God, who is totally ignorant of and consequently wholly indifferent to human individuals, and the God of religion. Within the framework of philosophical doctrine, the angel's words are quite meaningless. Not only is the God of the philosophers, who is a pure intellect, not concerned with man's works, but the (cultural) activities, involving both mind and body, to which the angel clearly referred, cannot, from the philosophical point of view, either help or hinder man in the pursuance of the philosophers' supreme goal—the attainment of union with the active intellect, a "light" of the divine nature. This union was supposed to confer knowledge of all intelligible things. Thus, man's supreme goal was here held to be of a purely intellectual nature.

In opposition to the philosopher's faith, the religion of ha-Levi's Jewish scholar is based upon the fact that God may have a close, direct relationship with man, who is not conceived primarily as a being endowed with intellect. The postulate that God can have intercourse with a creature made of the disgusting materials that go into the composition of the human body is scandalous to the King and prevents his acceptance of the doctrine concerning prophecy, expounded by the Muslim sage (just as the extraordinary nature of the Christological dogmas deters him from adopting Christianity).

The Jewish scholar's position is that it is contemplation

Knowledge of God through the contemplation of Jewish history

not of the cosmos but of Jewish history that procures knowledge of God. Ha-Levi was aware of the odium attaching to the doctrine of the superiority of one particular nation; he held, however, that only this doctrine explains God's dealing with mankind, which, like many other things, reason is unable to grasp. The controversies of the philosophers serve as proof of the failure of human intelligence to find valid solutions to the most important problems.

Ha-Levi's dialogue was also directed against the Karaites. He shows the necessity and celebrates the efficacy of a blind, unquestioning adhesion to tradition, which the Karaites rejected. Yet, he expounds a theology of Jewish exile that seems to have been influenced by Karaite doctrine. According to ha-Levi, even in exile the course of Jewish history is not determined like that of other nations by natural causes, such as material strength or weakness; the decisive factors are the religious observance or disobedience of the Jews. The advent of Christianity and of Islām prepares the other nations for conversion to Judaism, an event that will occur in the eschatological period (at the end of history).

Other Jewish thinkers c. 1050–c. 1150. During the period comprising the second half of the 11th century and the first half of the 12th century, many other Jewish thinkers appeared in Spain. Bahya ben Joseph ibn Pakuda wrote one of the most popular books of Jewish spiritual literature, *Kitāb al-hidāyah ilā farāʾiḍ alqulūb* ("Guidance to the Duties of the Heart"), which combines a theology influenced by Saʿadia with a moderate mysticism inspired by the teachings of the Muslim Ṣūfīs (mystics). The commandments of the heart—that is, those relating to men's thoughts and sentiments—are contrasted with the commandments of the limbs—that is, the Mosaic commandments enjoining or prohibiting certain actions. Bahya maintained that both sets of commandments should be observed (thus rejecting the antinomian position), but made clear that first and foremost he was interested in the commandments of the heart.

Abraham bar Hiyya Savasorda, a mathematician, astrologer, and philosopher, outlined in *Megillat ha-megalle* ("Scroll of the Revealer") a view of Jewish history that is reminiscent of ha-Levi but does not emphasize to the same degree the uniqueness of that history; it is also set forth in much less impressive fashion. Living in Barcelona under Christian rule, Bar Hiyya wrote his scientific and philosophical treatises not in Arabic but in Hebrew. Hebrew was also used by Abraham ibn Ezra (died *c.* 1167), a native of Spain who travelled extensively in Christian Europe. His commentaries on the Bible contributed to the diffusion among the Jews of Greek philosophical thought, to which Ibn Ezra made many, although as a rule disjointed, references. His astrological doctrine had a strong influence on some philosophers.

A borderline case: Abū al-Barakāt

The last outstanding Jewish philosopher of the Islāmic East, Abū al-Barakāt al-Baghdādī (who died as a very old man sometime after 1164), also belongs to this period. As a borderline case he illustrates a certain indeterminacy in the definition of a Jewish thinker. An inhabitant of Iraq, he was converted to Islām in his old age (for reasons of expediency, according to his biographers). His philosophy appears to have had a strong impact on Islāmic thought, whereas its influence upon Jewish philosophy and theology is very hard to pin down and may be practically nonexistent. His chief philosophical work, *Kitāb al-muʿtabar* ("The Book of That Which Has Been Established by Personal Reflection"), has very few references to Jewish texts or topics. Abū al-Barakāt rejects Aristotelian physics completely; according to him, time is the measure of being, and not, as Aristotle taught, the measure of motion, and he replaces Aristotle's bidimensional concept of place with the tridimensional notion of space, the existence of which is independent of the existence of bodies.

Jewish Aristotelianism. With regard to the adoption of Aristotelianism, including systems that in many essentials stem from but also profoundly modify the pure Aristotelian doctrine, there is a considerable time lag between the Islāmic East, on the one hand, and Muslim Spain and the Maghrib, on the other.

Abraham ibn Daud. Abraham ibn Daud (12th century), who is regarded as the first Jewish Aristotelian of Spain, was primarily a disciple of Avicenna, the great 11th-century Islāmic philosopher. According to a not unlikely hypothesis, he may have translated or helped to translate some of Avicenna's works into Latin, for Ibn Daud lived under Christian rule in Toledo, a town that in the 12th century was a centre for translators. His historical treatises, written in Hebrew, manifest his desire to familiarize his coreligionists with the historical tradition of the Latin world, which at that time was alien to most of them. But his philosophical work, *Sefer ha-emuna ha-rama* ("Book of Sublime Faith"), written in 1161 in Arabic, shows few, if any, signs of Christian influence.

The doctrine of emanation, set forth in the "Book of Sublime Religion," describes in the manner of Avicenna the procession of the ten incorporeal intellects, the first of which derives from God. This intellect produces the second intellect, and so on. Ibn Daud questioned in a fairly explicit manner Avicenna's views on the way the second intellect is produced; his discipleship did not by any means spell total adherence. Ibn Daud's psychology was also, and more distinctively, derived from Avicenna. The argumentation leading to a proof that the rational faculty is not corporeal attempts to derive the nature of the soul from the fact of immediate self-awareness. Like Avicenna, Ibn Daud tended to found psychology on a theory of consciousness.

Ibn Daud often referred to the accord that, in his view, existed between philosophy and religious tradition. As he remarked, the "Book of Sublime Faith" was not meant to be read either by readers who, in their simplicity, are satisfied with what they know of religious tradition or by those who have a thorough knowledge of philosophy. It was intended for readers of one type only, those who, being, on the one hand, acquainted with the religious tradition and having, on the other, some rudiments of philosophy, are "perplexed." It was for the same kind of people that Maimonides wrote his *Guide of the Perplexed.*

The greatest medieval Jewish philosopher

Maimonides. Maimonides (Moses ben Maimon, 1135–1204), a native of Spain, is incontestably the greatest name in Jewish medieval philosophy, but his reputation is not derived from any outstanding originality in philosophical thought. Rather, the distinction of Maimonides, who is also the most eminent codifier of Jewish religious law, is to be found in the vast scope of his attempt, in the *Dalālat al-hāʾirīn* (*Guide of the Perplexed*), to safeguard both religious law and philosophy (the public communication of which would be destructive of the law) without suppressing the issues between them and without trying to impose, on the theoretical plane, a final, universally binding solution of the conflict.

As Maimonides made clear in his introduction to the *Guide,* he regarded his self-imposed task as perilous, and he therefore had recourse to a whole system of precautions destined to conceal his true meaning from the people who, lacking the necessary qualifications, might misread the book and abandon observance of the law. According to Maimonides' explicit statement, these precautions include deliberately contradictory statements meant to mislead the undiscerning reader. The apparent or real contradictions that may be encountered in the *Guide* are perhaps most flagrant in Maimonides' doctrine concerning God. There seems to be no plausible hypothesis capable of explaining away the differences between the following three views:

1. God has an eternal will that is not bound by natural laws. Through an act of his will, he created the world in time and imposed on it the order of nature. This creation is the greatest of miracles; only if it is admitted can other miracles, which interfere with the causally determined concatenations of events, be regarded as possible. The philosophers' God, who is not free to cut the wings of a fly, is to be rejected. This conception is in keeping with the traditional religious view of God and is avowedly adopted by Maimonides because failure to do so would undermine religion.

2. Man is incapable of having any positive knowledge concerning God. No positive attributes—*e.g.,* wisdom or life—can be ascribed to God. Contrary to the attributes

predicated of created beings, the divine attributes are strictly negative; they state what God is not: for instance, he is not not-wise, and such a statement is not a positive assertion. Hence, only a negative theology is possible—saying what God is not. The way God acts can, however, be known. This knowledge is to be found in natural science.

3. God is an intellect. The formula current among medieval philosophers that maintains that in God the knowing subject, the object known, and the act of intellectual knowledge are identical derives from Aristotle's thesis that God knows only himself. Maimonides, however, in adopting the formula, interpreted it in the light of human psychology and epistemology (theory of knowledge), pointing out that, according to a theory of Aristotle, the act of human (not only of divine) cognition brings about an identity of the cognizing subject and cognized object. The parallel drawn by Maimonides between the human and the divine intellect quite evidently implies a certain similarity between the two; in other words, it is incompatible with the negative theology of other passages of the *Guide.* Nor can it be reconciled with his theological doctrine that the structure of the world—created in time—came into being through the action of God's will.

Maimonides' doctrine of prophecy

The enigma of the *Guide* would be nonexistent if Maimonides could be held to have believed that truth can be discovered in a suprarational way, through revelations vouchsafed to the prophets. This, however, is not the case. Maimonides held that the prophets (with the exception of Moses) combine great intellectual abilities, which qualify them to be philosophers, with a powerful imagination. The intellectual faculty of the philosophers and the prophets receives an overflow from the active intellect. In the case of the prophets, this overflow not only brings about intellectual activity but also passes over into the imaginative faculty, giving rise to visions and dreams. The fact that prophets have a strong imagination gives them no superiority in knowledge over philosophers, who do not have it. Moses, who belonged to a higher category than did the other prophets, did not have recourse to imagination.

The laws and religion as instituted by Moses are intended not only to ensure the bodily welfare and safety of the members of the community but also to facilitate the attainment of intellectual truths by individuals gifted enough to uncover the various hints embodied in religious laws and practices. This does not mean that all the beliefs inculcated by Judaism are true. Some indeed express philosophical truths, although in an inaccurate way, in a language suited to the intellectual capacity of the common people, who in general cannot grasp the import of the dogmas they are required to profess. Other beliefs, however, are false but necessary for the preservation of a public order upholding justice—*e.g.,* the belief that God is angry with wrongdoers.

As far as the Law—that is, the religious commandments—is concerned, two aspects of Maimonides' position may be distinguished. First, he maintained that it is unique in its excellence and valid for all time. This profession of faith, at least with regard to its assumptions about the future, lacked philosophical justification; however, it could be regarded as necessary for the survival of Judaism. Second, he asserted that certain precepts of the Mosaic Law were related to specific historical situations and the need to avoid too sharp a break with popular customs and practices, for instance, the commandments concerning sacrifice.

Influence of the *Guide of the Perplexed*

For at least four or five centuries the *Guide of the Perplexed* exercised a very strong influence in the European centres of Jewish thought; in the 13th century, when the *Guide* was twice translated into Hebrew, these centres were Spain, the south of France, and Italy. Rather paradoxically, in view of the unsystematic character of Maimonides' exposition, it was used as a standard textbook of philosophy and condemned as such when the teaching of philosophy came under attack. The performance of this function by the *Guide* was rendered possible, or at least facilitated by, the fact that from the 13th century onward the history of Jewish philosophy in European countries acquired a continuity it had never had before. This development seems to have resulted from the substitution of Hebrew for Arabic as the language of philosophical exposition. Because of the existence of a common and relatively homogeneous philosophical background—the Hebrew texts were much less numerous and less diverse than Arabic philosophical works—and the fact that Jewish philosophers reading and writing in Hebrew read the works of their contemporaries and immediate predecessors, something like a dialogue can be discerned. In striking contrast to the immediately preceding period, European Jewish philosophers in the 13th century and after frequently devoted a very considerable part of their treatises to discussions of the opinions of other Jewish philosophers. That many of the Jewish philosophers in question wrote commentaries on the *Guide* undoubtedly furthered this tendency.

Influence of Averroës and Christian Scholastics on medieval Jewish philosophy

Averroists. The influence of Maimonides' great Islāmic contemporary Averroës, many of whose commentaries and treatises were translated into Hebrew, was second only to that of Maimonides on Jewish intellectual development. Indeed, it may be argued that for philosophers, as distinct from the general reading public, it often came first. In certain cases commentators on the *Guide* tend, in spite of the frequent divergences between the two philosophers, to quote Averroës' opinions in order to clarify those of Maimonides.

The apparently significant influence of Christian Scholastic thought on Jewish philosophy was often not openly acknowledged by Jewish thinkers in the period beginning with the 13th century. Samuel ibn Tibbon, one of the translators of the *Guide* into Hebrew and a philosopher in his own right, remarked on the fact that the philosophical sciences were more widely known among Christians than among Muslims. Somewhat later, at the end of the 13th century and after, Jewish scholars in Italy translated into Hebrew varied texts of Thomas Aquinas and other Christian Scholastics; not infrequently, some of them acknowledged the debt they owed their Christian masters. In Spain and in the south of France, a different convention seems to have prevailed up to the second half of the 15th century. Whereas Jewish philosophers of these countries felt no reluctance about referring to Greek, Arabic, and other Jewish philosophers, as a rule they refrained from citing Christian thinkers whose views had, in all probability, influenced them. In the case of certain Jewish thinkers, this absence of reference to the Christian Scholastics served to disguise the fact that in many essentials they were representative of the philosophical trends, such as Latin Averroism, that were current among the Christian Scholastics of their time.

Quite evident is the resemblance between certain views professed by the Latin Averroists and the parallel opinions of Isaac Albalag, a Jewish philosopher who lived in the second half of the 13th century, probably in Catalonia, Spain, and who wrote a commentary in Hebrew on the *Tahāfut al-falāsifah* ("The Inconsistencies of the Philosophers"), an exposition of Avicenna's doctrine written by the Muslim philosopher al-Ghazālī. Albalag's assertion that both the teachings of the Bible and the truths demonstrated by reason must be believed even if they are contradictory clearly poses the question whether some historical connections exist between this view and the Latin Averroist doctrine that there are two sets of truths—the religious and the philosophical—and that these are not necessarily in accord. In most other points Albalag was a follower of the system of Averroës himself. This philosophical position may be exemplified by his rejection of the view that the world was created in time. He professed, it is true, to believe in what he called "absolute creation in time." This expression, however, merely signifies that at any given moment the continued existence of the world depends on God's existence, an opinion that is essentially in harmony with Averroës.

Joseph Caspi, a prolific 14th-century philosopher and exegetical commentator, maintained a somewhat unsystematic philosophical position that seems to have been influenced by Averroës. He expressed the opinion that knowledge of the future, including that possessed by God himself, is of a probabilistic nature. The prescience of the prophets is of the same nature. It is more than likely that Caspi's interest in this problem had some connection with the debate about future contingencies in which Christian Scholastics were engaged at that time.

Moses Narboni

Moses of Narbonne, or Moses Narboni, who lived in the south of France in the 14th century, was, like many other Jewish writers of this period, mainly a writer of commentaries. He wrote commentaries on biblical books, on treatises of Averroës, and on Maimonides' *Guide.* In his commentary on the *Guide,* Narboni often interprets the earlier Jewish philosopher's opinions by recourse to Averroës' views. Narboni also expounded and gave radical interpretations to certain conceptions that he understood as implied in the *Guide.* According to Narboni, God participates in all things, because he is the measure of all substances. God's existence appears to be bound up with that of the world, to which he has a relation analogous to that existing between a soul and its body (a comparison already made in the *Guide*).

Gersonides. Gersonides (Levi ben Gershom), another 14th-century Jewish philosopher born in the south of France, wrote the systematic philosophical work *Sefer milḥamot Adonai* ("The Book of the Wars of the Lord") as well as many philosophical commentaries. Gersonides apparently never explicitly mentioned Christian Scholastic philosophers; he cited Greek, Arabic, and Jewish thinkers only, and in many ways his system appears to have stemmed from the doctrines of Maimonides or Averroës, regardless of whether he agreed with them or not. For example, he explicitly rejected Maimonides' doctrine of negative theology. A comparison of his opinions and of the particular problems that engaged his attention, with the views and debates found in Scholastic writings of his period, however, suggests that he was also influenced by the Latins on certain points.

Gersonides disagreed both with the Aristotelian philosophers who maintained the eternity of the world and with the religious partisans who believed in the creation of the world in time out of nothing. He maintained that God created the world in time out of a preexistent body lacking all form. As conceived by Gersonides, this body seems to be similar to primal matter.

The problem of human freedom of action and a particular version of the problem of God's knowledge of future contingencies form an important part of Gersonides' doctrine. Gersonides, who, unlike the great Jewish and Muslim Aristotelians, believed in astrology, held that all happenings in the world except human actions are governed by a strict determinism. God's knowledge does not extend to individual human acts but embraces the general order of things; it grasps the laws of universal determinism but is incapable of apprehending events resulting from man's freedom. Thus, the object of God's knowledge is an ideal world order, which differs from the real world insofar as the latter is in some measure formed according to man's free will.

In political and social doctrine there is a fundamental difference between Maimonides and Gersonides. Gersonides does not appear to have assigned to the prophets any political function; according to him, their role consists in the prediction of future events. The providence exercised by the heavenly bodies ensures the existence in a given political society of men having an aptitude for and exercising the handicrafts and professions necessary for the survival of the community. He remarked that in this way the various human activities are distributed in a manner superior to that outlined in Plato's *Republic.* Thus, he rejected explicitly Plato's political philosophy, which, having been adapted to a society ruled through the laws promulgated by a prophet (Muḥammad), had been an important element of Jewish philosophy in the Arabic period.

Ḥasdai Crescas. Ḥasdai ben Abraham Crescas (1340–1410), a Spanish-Jewish thinker, like Gersonides had thorough knowledge of Jewish philosophy and partial knowledge of Islāmic philosophy, and both seem to have been influenced by Christian Scholastic thought; moreover, in certain important respects Crescas was influenced by Gersonides himself. In Crescas' main work, *Or Adonai* ("The Light of the Lord"), however, one of his objectives, quite contrary to Gersonides, was to expose the weakness and insufficiency of Aristotelian philosophy. This attitude may be placed in the wider context of the return to religion itself, as opposed to the Aristotelian rationalization of religion, and the vogue of Kabbala (Jewish theosophical mysticism), both of which were characteristic features of Spanish Jewry in Crescas' time. This change in attitude has been regarded as a reaction to the increasing precariousness of the position of the Jewish community in Spain.

Anti-Aristotelianism

The low estimation of the certainties and the rationalistic arrogance of the medieval Aristotelians coincided chronologically with a certain disintegration of and disaffection toward classical Aristotelian Scholasticism. Relevant to this decline were the so-called voluntarism of Duns Scotus, the Nominalism of William of Ockham and other 13th–14th-century Christian Scholastics, and the development, in the 14th century and after, of anti-Aristotelian physics at the University of Paris and elsewhere. Significantly there is a pronounced resemblance between Crescas' views and two of these trends, Scotism and the "new" physics.

Crescas accepted Gersonides' view that divine attributes cannot be negative, but unlike his predecessor he centred his explanation of the difference between the attributes of God and those of created existents on the antithesis between an infinite being and finite beings. It is through infinitude that God's essential attributes—wisdom, for instance—differ from the corresponding and otherwise similar attributes found in created beings. In Crescas', as in Spinoza's, doctrine (see below), God's attributes are also infinite in number. The central place assigned to the thesis of God's infinity in Crescas' system suggests the influence of Duns Scotus' theology, which is similarly founded upon the concept of divine infinity.

The problem of the infinite approached from an altogether different angle was one of the main themes of Crescas' critique of Maimonides' 25 propositions; these propositions, concerned mainly with Aristotelian physical doctrines, had been set forth in the *Guide* as the basis of Maimonides' proof for the existence of God. Crescas' declared purpose in criticizing and rejecting several of these propositions was to show that the traditional Aristotelian proofs (founded in the first place on physical doctrines) were not valid. In the course of his critique, Crescas attempted to disprove the Aristotelian thesis that the existence of an actual infinite is impossible. He held that space is not a limit but a tridimensional extension, that it is infinite, and that, contrary to Aristotle, the existence of a vacuum and of more worlds than one is possible. He also criticized as being impossible the thesis of the Aristotelian philosophers that there exists an infinite number of causes and effects, which have order and gradation. This thesis refers not to a temporal succession of causes and effects that have a similar ontological status but to a vertical series, descending from God to the lowest rung in creation. His attacks were likewise directed against the Aristotelians' conceptions of time and of matter.

Crescas' fundamental opposition to Aristotelianism is perhaps most evident in his rejection of the conception of intellectual activity as the supreme state of being for man and for God. Crescas' God is not first and foremost an intellect, and the supreme goal to which man can aspire is to love God with a love corresponding as far as possible to the infinite greatness of its object and to rejoice in the observance of his commandments. God, too, loves man, and his love, in spite of the lowliness of its object, is proportionate to his infinity.

Crescas attacked the separation of the intellect from the soul as conceived by the Aristotelians and attempted, perhaps in part under the influence of Judah ha-Levi, to refute the Aristotelian doctrine that the actualized intellect, in contradistinction to the soul, survives the death of the body. According to Crescas, the soul is a substance in its own right and can be separated from the body; it continues to subsist after the body's death.

Joseph Albo. Whereas Crescas unmistakably regarded the Aristotelian philosophers as adversaries to be criticized or combatted, the attitude of Joseph Albo (*c.* 1380–*c.* 1444), who regarded Crescas as his teacher, is much less clearly defined. Albo did not eschew self-contradiction, apparently considering it a legitimate precaution on the part of a philosophical or theological author; indeed, he

Albo's mixture of religious tradition and rational philosophy

indulged in it in a much more obvious way than did Maimonides. But, whereas the latter's fundamental philosophical position is fairly clear, the problem being how far he was prepared to deviate from Aristotelian doctrine in the interests of religion, there may be valid doubt whether Crescas and the Jewish religious tradition or Maimonides and Averroës were Albo's true masters. Mainly because of this perhaps deliberate failure to explain to the reader where he really stood, Albo has often been dismissed as an eclectic. He was strongly influenced not only by the authors just mentioned but also by Saʿadia. He seems to have had a considerable knowledge of Christian theology, even adopting for his own purposes certain Scholastic doctrines. He differs from Crescas and to some extent resembles Maimonides in having a marked interest in political theory.

The proclaimed theme of Albo's magnum opus, *Sefer ha-ʿiqqarim* ("Book of Principles"), is the investigation of the theory of Jewish religious dogmas, the number of which Maimonides, in a nonphilosophical work, had set at 13, whereas Albo, following a doctrine that in the last analysis seems to go back to Averroës, would limit the number to three: the existence of God, divine providence in reward and punishment, and the Torah as divine revelation. One section, usually including the philosophical and the traditional religious interpretations side by side, is devoted to each of these dogmas. Albo's principal and relatively novel contribution to Jewish doctrinal evolution is the classification, in his introduction, of natural, conventional, and divine law.

Natural law (the universal moral law inherent in human nature) is necessary, because man, being political by nature, must belong to a community, which may be restricted in size to one town or may extend over the whole earth. Natural law preserves society by promoting right and repressing injustice; thus, it restrains men from stealing, robbing, and murdering. The positive laws instituted by wise men take into account the particular nature of the people for whose benefit they are instituted, as well as other circumstances. This means that they differ from the natural law in not being universally applicable. Neither natural law nor the more elaborate conventional laws, however, lead men toward true spiritual happiness; this is the function of divine laws instituted by a prophet, which teach men true theoretical opinions. Whereas Maimonides maintained that Judaism was the only divine law promulgated by a true prophet, Albo considered that the commandments given to Noah for all mankind also constitute divine law, which ensures, although to a lesser degree than does Judaism, the happiness of its adherents. This position justifies a certain universalism; in accordance with a Talmudic saying, Albo believed that the pious among the non-Jews—that is, those who observe Noah's laws—have a share in the world to come. But he rejected the pretensions of Christianity and Islām to be divine laws.

MODERN PHILOSOPHY

The Iberian-Dutch philosophers. The expulsion of the Jews from Spain and Portugal produced a new centre of Jewish thought, Holland, where many of the exiled Jews found a new and safer domicile; the tolerance of the regime seemed to provide guarantees against external persecution. This did not prevent, and indeed may have furthered, the establishment of an oppressive internal orthodoxy that was prepared to chastise rebellious members of the community. This was evident in the cases of Uriel Acosta, or da Costa, and Benedict de Spinoza, two 17th-century philosophers who rebelled against Jewish orthodoxy and who were excommunicated for their views (Acosta twice).

The two excommunicados: Acosta and Spinoza

Uriel Acosta. Acosta came to Amsterdam from Portugal, where, belonging to a family of Marranos (Jews who had converted to escape religious persecution), he had been brought up in the Catholic faith; his philosophical position was to a great extent determined by his antagonism to the orthodox Judaism that he encountered in Amsterdam. His growing estrangement from generally accepted Jewish doctrine is attested by his Portuguese treatise *Sobre a Mortalidade da Alma* ("On the Mortality of the Soul"). He considered that the belief in the immortality of the soul has had many evil effects, for it impels men to choose an ascetic way of life and even to seek death. According to him, nothing has tormented men more than the belief in an inner, spiritual good and evil. At this stage Acosta affirmed the authority of the Bible from which, according to him, the mortality of the soul can be proved.

In his autobiography, written in Latin and entitled *Exemplar Humanae Vitae* ("Example of a Human Life"), he takes a more radical position. He proclaims the supreme excellency of the natural moral law (which, when arguing before Jews, he seems to identify with the divine commandments to Noah, thus suggesting a correspondence with the view of Albo). Accordingly, he denies the validity of the argument that natural law is inferior to Judaism and Christianity, because he believes that both these religions teach the love of one's enemies, a precept that is not a part of natural law and is a manifest impossibility.

Benedict de Spinoza. Although modern philosophers of Jewish origin are not considered as belonging to the history of Jewish philosophy unless they deal with Judaic themes, this restriction may not apply to Spinoza for the following reasons. (1) It was through the study of Jewish philosophical texts that Spinoza was first initiated into philosophy. (2) Spinoza's system is in part a radicalization of, or perhaps a logical corollary to, medieval Jewish doctrines; and the impact of Maimonides and of Crescas is evident. (3) A considerable portion of Spinoza's *Tractatus Theologico-Politicus* deals with problems related to Judaism.

Spinoza's view of prophecy

The first chapters of the *Tractatus* show that the doctrine of prophecy is of central importance to Spinoza's explanation of Judaism and that, in dealing with this subject, he used Maimonides' categories, although he applied them to different people or groups of people. Maimonides held that the prophets combined intellectual perfection, which made them philosophers, with perfection of the imaginative faculty. He also referred to a category of persons endowed with a strong imagination but possessing no extraordinary intellectual gifts; this category includes, for example, lawgivers and statesmen. Spinoza took over this last category but applied it to the prophets, whom he described as possessing vivid imaginations but as not necessarily having outstanding intellectual capacities. He denied that the biblical prophets were philosophers and used a philosophical and historical approach to the Scriptures to show that the contrary assertion is not borne out by the texts.

Spinoza also denied Maimonides' assertion that the prophecy of Moses was essentially different from that of the other prophets and that this was largely because Moses, in prophesying, had no recourse to the imaginative faculty. According to Spinoza, the distinctive fact about Moses' prophecy was that he heard the voice of God in a prophetic vision—that is, in a state in which his imagination was active. In this assertion Spinoza employed one of Maimonides' categories of prophecy, differentiated in the *Guide* according to certain characteristics of prophetic dreams and visions; however, Maimonides thought it improbable that the voice of God was ever heard in prophetic vision, and he held that this category is purely hypothetical. It seems evident that in his classification of Moses, Spinoza was concerned not with what really happened in history but with pigeonholing the biblical evidence into Maimonides' theoretical framework so that it fit in with his own theologico-political purpose: to show that there could be a religion superior to Judaism.

This purpose made it imperative to propound in the *Tractatus Theologico-Politicus* a theory concerning Jesus, whom Spinoza designates as Christus. The category and the status assigned to Jesus are by and large similar to those that Maimonides attributed to Moses. Thus, Jesus is referred to in the *Tractatus* as a religious teacher who makes recourse not to the imaginative faculty but solely to the intellect. His authority may be used to institute and strengthen the religion Spinoza called *religio catholica* ("universal religion"), which has little or nothing in common with any of the major manifestations of historic Christianity.

The difference between Judaism and Spinoza's *religio* *catholica* corresponds to the difference between Moses and Jesus. After leaving Egypt the Jews found themselves, in Spinoza's view, in the position of people who had no allegiance to any positive law; they had, as it were, reverted to a state of nature and were faced with the need to enter into a social pact. They were also an ignorant people and very prone to superstition. Moses, a man of outstanding ability, made use of the situation and the characteristics of the people in order to make them accept a social pact and a state founded upon it that, contrary to Spinoza's scheme for his ideal communities, were not based first and foremost upon utilitarian—that is, reasonable—consideration of the advantages of life in society over the state of nature.

The social pact concluded by the children of Israel in the desert was based upon a superstitious view of God as "King" and "Judge," to whom the children of Israel owed whatever political and military successes they obtained. It was to God rather than to the representatives of the popular will that the children of Israel transferred political sovereignty. In due course political sovereignty was vested in Moses, God's representative, and in his successors. It should be added that, in spite of Spinoza's insistence on the superstitious foundations of the ancient Israelite state, his account of its regime was not wholly unsympathetic. He believed, however, that it contained the seeds of its own destruction and that, with the extinction of this state, the social pact devised by Moses had lapsed and all the political and religious obligations incumbent upon the Jews had become null and void.

The rational state and its religion

It could be argued that, because the state conceived by Spinoza is based not on superstitious faith but on a social contract originating in rational, utilitarian considerations, it does not necessarily need to have its authority safeguarded and stabilized by means of religion. Nevertheless, Spinoza appears to have held the view—perhaps derived from a purely empirical knowledge of the behaviour of the common run of men—that there is a need for religion. In order to fulfill the need for some religion and to obviate the danger of harmful religions, he devised his *religio catholica,* the universal religion, which has the following distinctive traits: (1) Its main purpose, a practical one (which is furthered by recourse to the authority of Jesus), is to impel men to act in accordance with justice and charity. Such conduct is tantamount to obedience to the laws of the state and to the orders of the magistrates, in whom sovereignty is vested; for disobedience—even if it springs from compassionate motives—weakens the social pact, which safeguards the welfare of all the members of the community; in consequence, its evil effects outweigh whatever good it may produce. (2) Although religion, according to Spinoza, is not concerned with theoretical truth, in order to be effective the *religio catholica* requires dogmas, which he set forth in the *Tractatus.* These dogmas are formulated there in terms that can be interpreted in accordance both with the philosophical conception of God that Spinoza regarded as true and also with the superstitious ideas of ordinary people. It follows that if they are accepted as constituting by themselves the only creed that everybody is obliged to profess, people cannot be persecuted on account of their beliefs; Spinoza held that such a persecution may lead to civil war and may thus destroy the state. Philosophers are free to engage in the pursuit of truth and to attain, if they can, the supreme goal of man, freedom grounded in knowledge. There can be little doubt that the furtherance of the cause of tolerance for philosophical opinions was one of Spinoza's main objects in writing the *Tractatus.*

The relation between Spinoza's *Ethics,* his major philosophical work, and Jewish medieval philosophy is much more ambiguous than in the case of Spinoza's *Tractatus Theologico-Philosophicus.* In a way, Spinoza's metaphysical system, contained in the *Ethics,* can be regarded as being, in part, a spelling out of some extreme consequences, which could perhaps be legitimately drawn from medieval Aristotelianism; but the facts of the case are no doubt much more complicated than this.

German philosophers. *Moses Mendelssohn.* Moses Mendelssohn opened what may be called the German period of Jewish philosophy (*c.* 1750–*c.* 1830). This period, in which a considerable number of works on Jewish philosophy were written in German and often under the influence of German philosophy, is also marked by the emancipation of the Jews—that is, by the abrogation of discriminatory laws directed against them—and by their partial or complete assimilation. In this period in particular, the term Jewish philosophy applies especially to works the main purpose of which, or one of the main purposes of which, consists in proposing a definition of Judaism and a justification of its existence. The second task is often conceived as necessitating a confrontation of Judaism with Christianity rather than with philosophy, which served as a critical point of comparison for many medieval philosophers. This change seems to have been a result of the demarcation of the sphere of religion in such a way that, at least in the opinion of the philosophers, possible points of collision no longer existed between it and philosophy. This demarcation was largely furthered by the doctrine of Spinoza, from whom Mendelssohn and others took over and adopted for their own purposes certain fundamental ideas concerning Judaism. Like Spinoza, Mendelssohn held that it is not the task of Judaism to teach rational truths, although they may be referred to in the Bible. Contrary to what he called Athanasian Christianity—that is, the doctrine set forth in the Athanasian Creed—Judaism has no binding dogmas; it is centred on inculcating belief in certain historical events and on action—that is, observance of religious law (including the ceremonial commandments). Such observance is supposed to lead to happiness in this world and in the afterlife. Mendelssohn did not reject this view offhand, as Spinoza would have done; indeed, he seems to have been prepared to accept it, God's mysteries being inscrutable, and the radicalism and what may be called the consistency of Spinoza being the complete antithesis of Mendelssohn's apologetics. Non-Jews were supposed by Mendelssohn to owe allegiance to the natural moral law.

Solomon Formstecher. Whereas Mendelssohn continued the medieval tradition, at least to some extent, or adopted Spinoza's doctrine for his purposes, the Jewish philosophers of the first half of the 19th century may generally be regarded as disciples of the philosophers of their own time. In *Die Religion des Geistes* ("The Religion of the Spirit"), Solomon Formstecher (1808–89) may have been influenced by F.W.J. Schelling, an eminent German philosopher, in his conception of nature and spirit as manifestations of the divine. There are, in Formstecher's view, two types of religions that correspond to these manifestations: (1) the religion of nature, in which God is conceived as the principle of nature or as the world soul, and (2) the religion of the spirit, which conceives of God as an ethical being. According to the religion of the spirit, God has produced the world as his manifestation in full freedom and not, as the religion of nature tends to profess, because the world was necessary for his own existence.

The religions of nature and of spirit

The religion of the spirit, which corresponds to absolute religious truth, was first manifested in the Jewish people. The religious history of the world may be understood as a process of universalization of the Jewish religion. Thus, Christianity propagated Jewish conceptions among the nations; however, it combined them with pagan ideas. The pagan element is gradually being eliminated—Protestantism, for instance, in this respect, marks considerable progress. When at long last the Jewish element in Christianity is victorious, the Jews will be right to give up their isolation. The progress that will bring about this final religious union is already under way.

Samuel Hirsch. The main philosophical work of Samuel Hirsch, entitled *Die Religionsphilosophie der Juden* ("The Philosophy of Religion of the Jews"), was decisively influenced by G.W.F. Hegel. This influence is most evident in Hirsch's method and in the task that he assigned to the philosophy of religion—the transformation of religious consciousness into conceptual truth. Contrary to Hegel, however, he did not consider religious truth to be inadequate as compared to philosophical truth.

God revealed himself in the first stages of Jewish history by means of miracles and of prophecy. At present, he manifests himself in the miracle that is constituted by the existence of the Jewish people. At its beginning in the time of Jesus, Christianity was identical with Judaism. The decisive break between the two religions was caused by Paul. When the Pauline elements are eliminated from Christianity, it will be in all essentials in agreement with Judaism, which, however, will preserve its separate existence.

Nachman Krochmal. Nachman Krochmal, a native of Galicia (at that time, part of Austria), was the author of *More nevukhe ha-zman* ("Guide for the Perplexed for Our Time"), a treatise in Hebrew on the philosophy of history and on Jewish history that had a considerable influence. Krochmal's philosophical thought was centred on the notion of "spirit," Krochmal being mainly concerned with the "national spirit," the particular spirit that is proper to each people and that accounts for the peculiar characteristics differentiating one people from another in every domain of human activity. The national spirits of all peoples except the Jewish are, according to Krochmal, essentially particular. Hence, the national spirit either becomes extinct with the extinction of the nation or, if it is a powerful spirit, is assimilated by some other nation. The Jewish people have a special relation to the Universal Spirit, who is the God of Israel. This relation accounts for the perpetuity of the Jewish people.

Solomon Steinheim. Solomon Ludwig Steinheim, the author of *Die Offenbarung nach dem Lehrbegriff der Synagoge* ("The Revelation According to the Doctrine of the Synagogue"), was apparently influenced by the antirationalism of Friedrich Heinrich Jacobi, a German philosopher. His criticism of science is based on Jacobi's criticism, but he did not agree with Jacobi in opposing discursive reason to the intuitive knowledge of God—Steinheim contrasted human reason to divine revelation. The main point on which the revelation, vouchsafed to the prophets of Israel, is opposed to reason is to be found in the fact that the God posited by reason is subject to necessity, that he can act only in accordance with laws. Moreover, reason affirms that nothing can come from nothing. Accordingly, God is free to create not a good world, but only the best possible world. Revealed religion, on the other hand, affirms the freedom of God and the creation of the world out of nothing.

The Neo-Kantian philosopher of Judaism

Hermann Cohen. There seems to be little connection between the Jewish philosophers of the first half or two-thirds of the 19th century and Hermann Cohen (1842–1918), the head of the Neo-Kantian school centred at the University of Marburg. In a certain sense Cohen may be regarded as a rather unusual case among the philosophers of Judaism of his and the preceding generations because of the two aspects of his philosophical thought—the general and the Jewish—and the uneasy equilibrium between them. Judaism was by no means the only important theme of his philosophical system; it was one of several and not even his point of departure. There is no doubt that, for most of his life, Cohen was wholly committed to his brand of Kantianism, in the elaboration of which he displayed considerable originality—it has been maintained with some justification that his doctrine manifests a certain (unintentional) kinship with Hegel's. Cohen's idea of God, however, derives from an analysis and a development of certain conceptions of Immanuel Kant. In Cohen's view reason requires that nature be conceived of as conforming to one rational plan and that harmony exist between the domains of natural and of moral teleology (ultimate purposes or ends). These two requirements, in turn, necessitate the adoption of the idea of God—the word idea being used in the Kantian sense, which means that no assertion is made about the metaphysical reality of God.

Cohen seems to have changed his attitude in the last years of his life; although he did not explicitly renounce his previous positions, a considerable shift of emphasis can be discerned in his doctrines. The notion of the human individual—an individual who is weak and full of sin—comes to the fore, as well as the conception of a correlation, a relationship, between God and the individual. This relationship is one of love, the love of God for man and the love of man for God. It is difficult to reconcile the conception of God expounded in Cohen's work of his last period with his Kantian or Neo-Kantian attitude toward metaphysics.

Franz Rosenzweig. Franz Rosenzweig published his main philosophical work, *Der Stern der Erlösung* (*The Star of Redemption,* 1971), in 1921. This work begins with a rejection of the traditional philosophical attitude that denies the fear of death, maintaining, instead, that this fear is the beginning of the cognition of the All. Man should continue to fear death, despite the indifference of philosophy and its predilection for accepting death. Traditional philosophy is interested exclusively in the universal, and it is monistic—its aim is to discover one principle from which everything can be derived. This tendency of philosophy, however, denatures human experience, which knows not one but three separate domains (which Kant had referred to in a different context), namely, God, the world, and man.

According to Rosenzweig, God (like the world and like man) is known through experience (the experience of revelation). In Greek paganism, the most perfect manifestation of paganism, every one of these domains subsists by itself: the gods, the cosmos, and man as the tragic, solitary, silent hero. Biblical religion is concerned with the relation between the three: the relation between God and the world, which is creation; the relation between God and man, which is revelation; and the relation between man and the world, which leads to salvation. The philosophy that renounces the ambition to find one principle for everything that exists and that follows biblical religion in centring on the connections between the three domains and between the words and acts that bring about and develop these connections, Rosenzweig termed the narrative philosophy; the term and the concept were taken over from Schelling, whose influence Rosenzweig repeatedly emphasized.

The biblical faith brought forth two valid religions—Christianity and Judaism. The first is described by Rosenzweig as the eternal way; the Christian peoples seek in the vicissitudes of time and history the way to salvation. In contradistinction to them, the existence of the stateless Jewish people is not concerned with time and history; it is—notwithstanding the hope for final salvation—already an eternal life, renewed again and again according to the rhythm of the liturgical Jewish year.

Martin Buber. Since the early years of the 20th century, Martin Buber has exercised a powerful influence on both Jews and non-Jews. In his early period Buber was led, partly through empathy with Jewish and non-Jewish mysticism, to stress unitive experience and knowledge, in which the difference between one man and another and between man and God tend to disappear. But in his final period he taught, following, as he claimed, a suggestion of Ludwig Feuerbach, a 19th-century German philosopher, that man can only realize himself as a human being in a relation with another, who may be another man or God. This conception of the "I and Thou" relationship leads to the formulation of Buber's view of the dialogical life—the mutual, responsive relation between man and others—and accounts for the importance that he attaches to the category of "encounter." (S.Pi.)

"I and Thou"

Jewish mysticism

This section deals with the special nature and characteristics of Jewish mysticism, the main lines of its development, and its role in present-day religion and culture.

NATURE AND CHARACTERISTICS

The Judaic context. The term mysticism applies whenever a person is convinced that it is possible to establish *direct contact,* apart from sense perception and intellectual apprehension, with the divine—a reality undefinable by pure logic and believed to be the ultimate ground of being. Since mysticism springs from an aspiration to join and grasp that which falls outside ordinary experience, it is not easily restricted within precise limits. The boundary line that separates mysticism from metaphysics and cosmology (doctrines on the basic nature or structure of being and the world), from theosophies (systems of thought claiming special insights or revelation into the divine nature), and various forms of occultism (the study and control of supernatural powers), and from theurgy (the art of compelling or persuading divine powers) and even magic, often of the lowest kind, is not clear.

If mysticism is defined as the search for *direct* contact with the divine, however, it seems to be incompatible with Judaism. In its classical and normative form, Judaism appears as faith in a sole God who created the universe and who chose to reveal himself to a selected group by means of a rule of life he imposed on it—Torah ("Guidance" or "Teachings," incorrectly rendered as "Law"). The earthly destiny of the chosen nation, as well as the eternal salvation of the individual, in traditional Judaic beliefs, depends upon the observance of this rule of life, through which any relationship to God must take place. The fact is, however, that in the religious history of Judaism the quest for God goes beyond this relationship mediated by Torah, without ever dispensing with it (since that would take the seeker outside of Judaism) or pretending to reach the depths of the mystery of the divine, or still less to end in an ontological identification with it (where God and man are the same in nature and being).

Three types of Jewish mysticism. Three types of mysticism may be discerned in the history of Judaism: the ecstatic, the contemplative, and the esoteric. Though they are distinct types, in practice there are frequent overlappings and mixtures between them.

The first type is characterized by the quest for God—or, more precisely, for access to a supernatural realm, which is itself still infinitely remote from the inaccessible deity—by means of ecstatic experiences; this method is sometimes tainted by theurgy. The second follows the way of metaphysical meditation pushed to the limit, always bearing in its formulations the imprint of the cultural surroundings of the respective thinkers, who are exposed to influences from outside Judaism; this was the case with Philo of Alexandria (*c.* 15 BCE–after 40 CE) and a few of the Jewish thinkers of the Middle Ages, who drew their inspiration from Greco-Arabic Neoplatonism and sometimes also from Muslim mysticism.

Esoterism

The third type of mysticism claims an esoteric knowledge (hereafter called esoterism) that explores the divine life itself and its relationship to the extradivine level (the natural, finite realm) of being, a relationship that is subject to the "law of correspondences." From this perspective, the extradivine is a symbol of the divine; that is, a reality that reveals another, superior reality, whence reciprocal action of the one on the other (which corresponds to it) exists. This form of mysticism, akin to gnosis—the secret knowledge claimed by Gnosticism, a Hellenistic religio-philosophical movement—but purged, or almost purged, of the dualism that characterizes the latter, is what is commonly known as Kabbala (literally "tradition"). By extension, this term is also used to designate technical methods, used for highly diverse ends, ranging from the conditioning of the aspirant to ecstatic experiences to magical manipulations of a frankly superstitious character. If the concept of spiritual energy acting on matter and at a distance originally underlay these practices, it finally became unrecognizable and all that remained was a collection of "tricks of the trade."

The favour with which the doctrine of correspondences was regarded by ancient and medieval science, as well as the tendency in the three monotheistic religions (Judaism, Christianity, and Islām) to reconcile the results of rational reflection with the data of revelation, had the result of turning speculation on the origin and order of the universe toward mysticism.

Perennial human questions

It must also be noted that the quest for God implies the search for solutions to problems that go beyond those of religion in the narrow sense and that arise even when there is no interest in the relationship between man and supernatural powers. Man ponders the problems of his origins, his destiny, his happiness, his suffering—questions that arise outside of religion, as well as within nonmystical forms of religious life; the presence or absence of religious institutions or dogmas is of little importance when it comes to these questions. They were all formulated within nonmystical Judaism and served as the basis and framework for the setting and solution of problems in the various forms of Jewish mysticism. This mysticism, especially in its "Kabbalistic" form, brought about profound transformations in the concepts of the world, God and "last things" (resurrection, last judgment, messianic kingdom, etc.) set forth in biblical and rabbinical Judaism. Nevertheless, Jewish mysticism's own set of problems about the origins of the universe and of man, of evil and sin, of the meaning of history, of the afterlife and the end of time is rooted in the very ground of Judaism and cannot be conceived outside of an exegesis of revealed Scripture and rabbinical tradition.

MAIN LINES OF DEVELOPMENT

A study of the main lines of Jewish mysticism, following its actual historical development, reveals that during a very long period, from its origins in the 1st century CE to the middle of the 12th century, only the first two of the three types outlined above existed. It was not until the second half of the 12th century that esoterism became clearly discernible; from then on it continued to develop in various forms up to very recent times.

Early stages to the 6th century CE. The centuries that followed the return from the Babylonian Exile in the 6th century BCE witnessed the growth and intensification of reflection on the intermediary beings between man and God, of meditation on the divine appearances whose special place of occurrence had formerly been the most sacred part of the Jerusalem Temple, of speculation on the coming into being and organization of the universe and on the creation of man. None of these themes was absent from the Bible, which was held to be divinely revealed, but each had become the object of a constant ideological readjustment that also involved the infiltration of concepts from outside and reaction against them. The speculative taste of Jewish thinkers between the 2nd century BCE and the 1st century CE took them in many different directions: angelology (doctrine about angels) and its counterpart demonology (doctrine about devils); mythical geography and uranography, description of the heavens; speculation on the divine manifestations—which had as background the Jerusalem Temple worship and the visions of the moving "Throne" (the "Chariot," *Merkava*) in the prophecy of Ezekiel; on the double origin of man, a being formed of the earth but also the "image of God"; on the end of time; on resurrection (a concept that appeared only toward the end of the biblical period); and on rewards and punishments in the afterlife.

The literary crystallization of all this ferment was accomplished in writings, such as the book of Enoch, of which Pharisaic (rabbinical) Judaism—which became the normative Jewish tradition after the Roman conquest of Jerusalem and the destruction of the Second Temple (70 CE)—retained almost nothing and even the vestiges of which it tended to obliterate in its own writings; the Talmud and the Midrash (rabbinical legal and interpretative literature) touched these themes only with great reserve, often unwillingly and more often in a spirit of negative polemic.

As early as the 1st century CE, and probably even before the national calamity of 70, there were certainly sages or teachers recognized by the religious community for whom meditation on the Scriptures—especially the creation narrative, the public revelation of the Torah on Mount Sinai, the *Merkava* vision of Ezekiel, the Song of Solomon—and reflection on the end of time, resurrection, and the afterlife were not only a matter of exegesis and of attaching new ideas to texts recognized to be of divine origin but also a matter of inner experience. It was, however, probably in other circles that speculation on the invisible world was engaged in and where the search for the means of penetrating it was carried out. It is undeniable that there exists a certain continuity between the apocalyptic visions (*i.e.*, of the cataclysmic advent of God's Kingdom) and documents of certain sects (Dead Sea Scrolls) and the writings, preserved in Hebrew, of the "explorers of the supernatural world" (*Yorde Merkava*). The latter comprise ecstatic hymns, descriptions of the "dwellings" (*hekhalot*) located between the visible world and the ever-inaccessible divinity, whose transcendence is paradoxically expressed by anthropomorphic descriptions consisting of inordinate hyperboles (*Shiʿur qoma,* "Divine Dimensions"). In addition, a few documents have been preserved that attest to the existence of methods and practices having to do with the initiation of carefully chosen persons who were made to undergo tests and ordeals in accordance with psychosomatic criteria borrowed from physiognomy (art of determining character from physical, especially facial, traits). Some theurgic efficacy was attributed to these practices, and there was some contamination from Egyptian, Hellenistic, or Mesopotamian magic. (A curious document in this respect, rich in pagan material, is the *Sefer ha-razim,* the "Treatise on Mysteries," which was discovered in 1963.)

Two early sources of mystic writings

In this extrarational domain, there are many similarities between concepts reflected in unquestionably Jewish texts and the documents of contemporary non-Jewish esoterism, to the point that it becomes difficult, sometimes impossible, to distinguish the giver from the receiver. Two facts are certain however. On the one hand Gnosticism never ceases to exploit in its own way biblical themes (such as the tale of creation and speculation on angels and demons) that have passed through Judaism, whatever their original source may have been; on the other hand, though Jewish esoterism may borrow this or that motif from ancient gnosis or syncretism (fusion of various faiths) and may even raise to a very high rank in the hierarchy of being a supernatural entity such as the angel Metatron, also known as "little Adonai" (*i.e.,* little Lord or God), it still remains inflexibly monotheistic and rejects the Gnostic concept of a bad or simply inferior demiurge who is responsible for the creation and governing of the visible world. Finally, it is noteworthy that during the centuries that separate the Talmudic period (2nd to 5th centuries AD) from the full resurgence of Jewish esoterism in the middle of the 12th century, the texts that have been preserved progressively lose their density and affective authenticity and become reduced to the level of literary exercises that are more grandiloquent than substantial.

Sefer Yetzira. In the ancient esoteric literature of Judaism, a special place must be given to the *Sefer Yetzira* ("Book of Creation"), which deals with cosmogony and cosmology (the origin and order of the universe). Creation, it affirms with a clearly anti-Gnostic insistence, is the work of the God of Israel and took place on two different levels: the ideal, immaterial level and the concrete level. This was done according to a complex process that brings in the 10 numbers (*sefirot,* singular *sefira*) of decimal notation and the 22 letters of the Hebrew alphabet. The 10 numbers are not to be taken merely as arithmetical symbols: they are cosmological factors, the first of which is the spirit of God—with all the ambiguities that this term *ruaḥ* has in Hebrew—while the nine others seem to be the archetypes of the three elements (air, water, fire) and the spatial dimensions (up, down, and the four cardinal points). After having been manipulated either in their graphic representation or in combination, the letters of the alphabet, which are considered to be adequate transcriptions of the sounds of the language, are in turn instruments of creation.

The ontological value of speech

The basic idea of all this speculation is that speech (that is, language composed of words, which are in turn composed of letters/sounds) is not only a means of communication but also an operational agent destined to produce being—it has an ontological value. This value, however, does not extend to every form of language; it belongs to the Hebrew language alone.

The universe that is produced by means of the *sefirot* and the letters is constituted according to the law of correspondences between the astral world, the seasons that mark the rhythm of time, and man in his psychosomatic structure.

The "Book of Creation" certainly does not proceed entirely from biblical data and rabbinical reflection upon them; certain Greek influences are discernible, even in the vocabulary. What is important, however, is its influence on later Jewish thought, down to the present time: philosophers and esoterists have vied with one another in commentating it, pulling it in their own direction, and adjusting it to their respective ideologies. Even more important is the fact that Kabbala (see below *The making of the Kabbala*) borrowed a great part of its terminology from it (*sefira,* among others), naturally making semantic adaptations as required.

The speculation traced above developed during the first six centuries of the Common Era, both in Palestine and in Babylonia (later called Iraq); Babylonian Judaism had its own social and ideological characteristics, which put it in opposition to Palestinian Judaism in various aspects, including esoterism as well as other manifestations of the life of the spirit. The joint doctrinal influence of the two centres was to spread during the period from the mid-8th to 11th century among the Jews established in North Africa and Europe; mystical doctrines also filtered in, but very little is known about the circumstances and means of their penetration.

The Arabic-Islāmic influence (7th–13th century). Arabic-Islāmic culture provided another important influence in Jewish mystical development. A considerable part of Jewry, which had fallen under Muslim domination in the 7th and 8th centuries, participated in the new Arabic-Islāmic civilization; the Jews of Asia, Africa, and Spain soon adopted Arabic, the prevailing language of culture and communication. By way of Arabic-language culture, elements of Greek philosophy and Islāmic mysticism penetrated Judaism and contributed to the deepening of certain theological concepts that were Jewish in origin but had become the common property of the three religions of the Book: affirming the divine unity, purging all anthropomorphism from the idea of God, and approaching the divine by progressing on a spiritual path that leads through an ascetic discipline (both physical and intellectual) to a detachment from this world and a freeing of the soul from all that distracts it from God. Greek philosophy and Islāmic mysticism, moreover, raised very serious questions that threatened many traditional beliefs, such as the creation of the world, the providential action of God, miracles, eschatology (doctrines about the resurrection of the body, rewards, and especially material punishments in the hereafter). Even in the Christian West, where cultural contacts between the majority society and the Jewish minority were far from reaching the breadth and intensity of the Judeo-Arab relations, Jewish intellectuals were unable to remain totally impervious to the incursions of the surrounding civilization. Moreover, at the beginning of the 12th century, if not earlier, European Judaism received part of the intellectual Arabic and Judeo-Arabic heritage through translations or adaptations into Hebrew, its only cultural language.

Greek philosophy and Islāmic mysticism

The making of the Kabbala (c. 1150–1250). It was in these circumstances that, starting around 1150, manifestations of markedly theosophic ideologies appeared in the south of present-day France (in the regions of Provence–Languedoc–Roussillon). Two types can be distinguished at the outset, which are very different as to their manner of appearance, their form, and their content.

Sefer ha-bahir. The first type is represented in fragmentary, poorly written, and badly assembled texts that began to circulate in Provence–Languedoc during the third quarter of the 12th century. Their inspiration, however, leaves no doubt as to the community of their origin. They were in the form of a Midrash; that is, an interpretation of Scripture with the help of a particular interpretative method, full of sayings attributed to ancient rabbinical authorities. This whole body of texts, probably imported from the Near East (Syria–Palestine–Iraq), is known as the "Midrash of Rabbi Nehunya ben Haqana" (from the name of a 2nd-century rabbi) or *Sefer ha-bahir,* "Book of Brightness" (from a characteristic word of the first verse of Scripture to be elucidated in the work). The authorities cited are all inauthentic (as was often the case in late works), and the content of this Midrash, even its nonmystical content, is entirely Gnostic; a Gnosticism that tries nevertheless to escape any ontological dualism (and, as a matter of fact, succeeds).

Its object is to present the origin of things and the course of history centred naturally on that of the chosen people, with the vicissitudes caused in turn by obedience to God and by sin, as bound and conditioned by the

Role of the divine powers

manifestation of divine powers. These "powers" are not "attributes" derived and defined by philosophical abstraction, although that is one of the terms used to designate it: they are hypostases (essences or substances). They are inseparable from God, but each one is clothed in its own personality, each operates in its own manner, in the leaning toward severity or mercy, in dynamic correspondence with the behaviour of man, especially of the Jew, in the visible world. They are ranked in a hierarchical order, which is not yet as fixed as it became starting with the second generation of Kabbalists in Languedoc and Catalonia (see below *The school of Gerona* [*Catalonia*]). The rich nomenclature used to designate the "powers" exploits the resources of both the Bible and rabbinical tradition, of the "Book of Creation," of some ritual observances, and also of the letters of the Hebrew alphabet and the signs that can be added to them to indicate the vowels. All of this combines to give a symbolical rendering of the myth, cosmology, sacred history, and eschatology through which an anonymous group of theosophists attempt to formulate their doctrine: a Gnostic myth, except for the adjustments that eliminate the radical depreciation of the visible world.

Thus, according to the *Sefer ha-bahir,* the universe is the manifestation of the hierarchically organized divine powers, and the one that is at the bottom of the hierarchical ladder has special charge of the visible world. This entity is highly complex. Undoubtedly there are survivals of Gnostic speculation on Sophia ("Wisdom"), who is involved, sometimes to her misfortune, in the material world. This power is also the divine "Presence" (*Shekhina*) of rabbinical theology but profoundly transformed: it has become a hypostasis; by a bold innovation, moreover, it is characterized as a feminine being and thus finds itself, while remaining an aspect of the divinity, in the position of a daughter or a wife, who owns nothing herself and receives all from the father or the husband. It is also identified with the "Community of Israel," another radical innovation, but facilitated by ancient speculation based on the allegorical interpretation of the Song of Solomon, which represents the relationship of God to the chosen nation in terms of the marriage bond. Thus a theosophical equality is established between the whole of the people chosen by God, constituted into a kind of mystical body, and an aspect of the divinity, whence the solidarity and linked destiny of the latter and the human group in question. As a matter of fact, a comparable relationship between the "Presence" and Israel was not totally foreign to ancient rabbinical theology. In this light, the obedience or disobedience of Israel to its particular vocation is a determining factor of cosmic harmony or disruption and extends to the inner life of the divinity. This is the essential and definitive contribution of the *Sefer ha-bahir* to Jewish theosophy. In the same document may be seen the resurgence of a notion fought against by the older theologians—that of *metensōmatōsis,* the reincarnation into several successive bodies of a soul that has not attained the required perfection in a previous existence.

School of Isaac the Blind. Parallel to the appearance of the *Sefer ha-bahir* but independent of it, another theosophic tendency unfolded in Languedoc, the second type referred to above. The two movements would take only about thirty years to converge, to constitute what may conveniently, though not quite precisely, be called classical Kabbala. The second school flourished in Languedoc during the last quarter of the 12th century and crossed the Pyrenees into Spain in the first years of the 13th century.

Return to the undifferentiated One

The most eminent spokesman of this school was Isaac ben Abraham, known as Isaac the Blind. For this theosophist, among whose extant works there is in particular a very obscure commentary of the "Book of Creation," the general vision of the universe proceeds, to use the words of Gershom G. Scholem (the eminent 20th-century Kabbala scholar), from the link he discovers between the hierarchical orders of the created world and the roots of all beings implanted in the world of the *sefirot.* One can already see a Neoplatonic influence in the reflections of Isaac; *e.g.,* the proceeding of things from the One and the corresponding return to the heart of the primordial undifferentiatedness, which is the fullness of being and at the same time every conceivable being. This return is not merely eschatological and cosmic but is in some way realized in the life of prayer of the contemplative mystic privileged to have supernatural inspirations, "appearances" of the prophet Elijah, by means of concentration, of orientation of action and thought (*kawwana*), and of "adhesion" (*devequt*), being-with-God, though not, indeed, a transforming union by which the human personality blends completely into the deity or becomes one with it.

The synthesis of the themes of the *Bahir* and the cosmology of the "Book of Creation," accomplished by Isaac or by others in the doctrinal environment inspired by his teachings, is and remains the foundation of Kabbala whatever enrichment, adjustments, even changes of orientation and sometimes radical modifications the composite may have undergone subsequently.

The 10 sefirot. It is also in this environment that the nomenclature of the 10 *sefirot* became more or less fixed; it is important to remember this, whatever variant terminologies and even divergent concepts as to the nature of these entities may exist elsewhere—*e.g.,* as internal powers of the divine organism (Gnostic point of view), as hierarchically ordered intermediaries between the infinite and the finite (Neoplatonic concept), or simply as instruments of the divine activity, neither partaking of the divine substance nor being outside it.

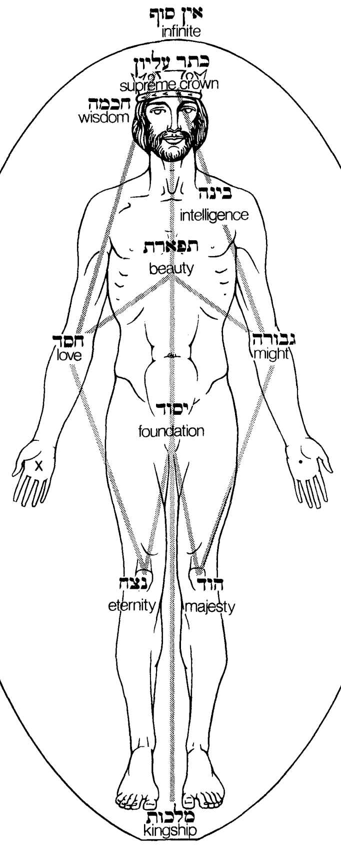
The classical list of the *sefirot* is:

1. *keter ʿElyon*	The Supreme Crown (its identity or nonidentity with the Infinite, *En Sof,* the unknowable deity, remains problematical)
2. *ḥokhma*	Wisdom, the location of primordial ideas in God
3. *bina*	Intelligence, the organizing principle of the universe
4. *ḥesed*	Love, the attribute of goodness
5. *gevura*	Might, the attribute of severity
6. *tif'eret*	Beauty, the mediating principle between the preceding two
7. *netzaḥ*	Eternity
8. *hod*	Majesty
9. *yesod*	Foundation of all the powers active in God
10. *malkhut*	Kingship, identified with the *Shekhina* ("Presence")

The School of Gerona (Catalonia). The double current of the gnosticizing theosophy of the *Sefer ha-bahir* and the contemplative mysticism of the masters of Languedoc became one in the elaborations it was subject to at the hands of the Kabbalists in Catalonia, where the Jewish community of Gerona was, during the first half of the 13th century, a veritable seat of esoterism. These elaborations followed the same overall lines, though they were at the same time highly diversified, depending on the personal inclinations of each writer. To the school of Gerona belong, among others, masters such as Ezra ben Solomon, Azriel of Gerona, Jacob ben Sheshet, Moses ben Naḥman (or Naḥmanides, *c.* 1195–1270, the famous Talmudist, biblical commentator, and mystical philosopher); their influence on the subsequent course of Jewish mysticism is of fundamental importance. None of them has left a complete synthesis of his theosophy; they expressed themselves, with more or less reserve, by means of commentaries, sermons, polemic or apologetic treatises or, at the most, brief summaries for the noninitiated. It is not impossible, however, to discover through these texts their vision of the world and compare it with the views of the Jewish thinkers who attempted to harmonize the biblical-rabbinical tradition with Greco-Arab philosophy, whether of Neoplatonic or Aristotelian inspiration.

The divine "nothing"

At the base of the Kabbalistic view of the world there is an option of faith: it is by a voluntary decision that the unknowable deity—who is "nothing" or "nothingness" (nonfinite) because he is a fullness of being totally inaccessible to any human cogitation—set into motion the process that leads to the visible world. This concept radically separates Kabbala from the determinism from which the philosophy of the period could not, without internal



Sefirot superimposed on the human body (see text).

From C.D. Ginsburg, *Kabbalah: The Essenes;* Samuel Weiser Inc.

contradictions, free the principle of being. In addition it offers a solution consistent with faith to the problem, highly embarrassing for the philosophers, of creation *ex nihilo* (out of nothing): the paradoxical reinterpretation of the concept of the "nothing" eliminates the original matter coeternal with God and solves the opposition between divine transcendence (remoteness from the world) and immanence (presence in the world); issuing from the unfathomable depth of the deity and called to return to it, the world, visible as well as invisible, is radically separated from God, who is at the same time constantly present. The correspondence between the *sefirot,* which are modes of the divine manifestation, and all the degrees of being gives meaning to the structure of the world and to the history of humanity centred on the revelation especially given to the chosen people, a revelation that is a rule of life for this people and, consequently, the criterion of merit and sin, or good and evil. Thus, from the top to the bottom of the ladder, there are but corresponding realities that control one another; contrary to the opinion of the philosophers, evil is also a reality since it is the rupture of the universal harmony. It is also the consequence of this rupture, in the form of punishment, but it is repairable. From this perspective, scrupulous observance of the Torah, the revealed Law (both in the written text and the oral tradition), is the essential factor for the very maintenance of the universe. From that point on, the "rational" motivation of the commandments, which raises insurmountable difficulties for the theologians of philosophical orientation, is in the eyes of the Kabbalists but a false problem; the real problem is the fundamental nature of the Torah. Kabbala brings more than one solution to it, whereas philosophy is not even able to raise the question.

It follows from this general concept that the Jewish faith, with its implications—the conviction of holding the undiluted truth, the faithful preservation of ritual practices, and the eschatological expectation—is safeguarded from all the doubts that either philosophical speculation or the rival religious doctrines of Christianity and Islām could evoke in the minds of Jewish believers. Considered from this point of view, Kabbala, already at the stage it had reached at Gerona, turns out to have been a significant factor in the survival of Judaism, which was exposed everywhere in medieval society to the perils that the history of the period reveals.

Besides the Gerona school and the doctrinal descendants of Isaac the Blind in Languedoc, there was another school of Jewish esoterism in southern Europe during the first half of the 13th century. This school—whose followers preferred to remain anonymous and therefore published their writings, such as the *Sefer ha-ʿiyyun* ("Book of Speculation"), either without giving any author's name or by attributing them to fictitious authorities—directed its speculation both to the highest levels of the divine world, where it discerned further aspects beyond the 10 *sefirot* and attempted to give an idea of them by resorting to the symbolism of light, and to the primordial causes and the archetypes contained in the deity or directly issued from it. The sometimes striking similarity between these speculations and those of John Scotus Erigena, a notable 9th-century Christian philosopher, seems to indicate not only a typological kinship of themes between this Kabbalistic current and Latin-language Christian Neoplatonism but also a concrete influence of the latter upon the former. The same may be true about Isaac the Blind and the school of Gerona, but certain knowledge is lacking.

Sefer ha-temuna. Still another current manifested itself at the same period; it found its literary expression in the *Sefer ha-temuna* ("Book of the Image") of unknown authorship. This very obscure document claims to explain the figures of the letters of the Hebrew alphabet. In fact, the speculation of this treatise bears on two themes that were not foreign to the school of Gerona, but it develops them in a personal manner that decisively influenced the future of Jewish theosophy. On the one hand, it deals with a theory of different cycles through which the world must travel from the time of its emergence to its reabsorption into the primordial unity and, on the other hand, with various readings that correspond to these cycles in the divine manifestation that is constituted by the revealed Scriptures. In other words, the reading, thus the interpretation, and consequently the message of the Torah vary according to the cycles of existence; the passage to a cycle other than that under whose governance humanity is presently living could thus entail the modification, even the abrogation, of the rule of life to which the chosen people are presently subject, an explosive notion that opened the way to an overthrow of the traditional values of Judaism.

The cycles of existence

Medieval German (Ashkenazic) Ḥasidism. The period from *c.* 1150 to 1250, which witnessed the establishment of Kabbala in the south of France and in Spain, is no less important for the shaping of Jewish mysticism in the other branch of European Judaism, in northern France (and England) and in the Rhine and Danube regions of Germany. Unlike medieval Kabbala, which was to experience a broad and varied development starting in the second half of the 13th century, the movement designated somewhat summarily as German (or Ashkenazic, from a biblical place-name conventionally used to designate Germany) Ḥasidism (Pietism), would hardly survive as a living and independent current beyond the second quarter of the 13th century. There was undoubtedly within Franco-German Judaism a certain continuity of mystical tradition, based on the *Sefer Yetzira* and the *Hekhalot* (see above *Sefer Yetzira*); certain elements of theurgy and magic of Babylonian origin had perhaps also reached it through Italy; and it would seem that the gnosticizing current crystallized in the *Sefer ha-bahir* did not pass without leaving traces in Germany. The intellectual atmosphere of Franco-German Judaism, however, differed greatly from that reigning in Spain or even Provence–Languedoc; it was characterized by an almost exclusively Talmudic culture, less intellectual contact with the non-Jewish environment than in the countries of Muslim civilization, and a very limited knowledge of the Jewish theology in Arabic of the Middle East, North Africa, and Spain. This situation would change only in the last third of the 12th century; until then, the "philosophical" equipment of the Franco-German Jewish scholar consisted essentially of a Hebrew paraphrase, dating perhaps from the 11th century, of the treatise *Beliefs and Opinions* by Saʿadia ben Joseph (the great 9th–10th-century Babylonian Jewish scholar and philosopher), and the commentary on the "Book of Creation," written directly in Hebrew (in 946) by the Italian Jew Shabbetai Donnolo. Even when the cultural influence of Spanish Judaism came to be felt more strongly in France–England and Germany, the speculative Kabbala noted above hardly penetrated there. Thinkers within Franco-German Judaism who inclined toward theological speculation had their own problems, which resulted in a mysticism strongly imbued with asceticism, a type of mysticism toward which the general situation of the Jews in those regions contributed, as, especially after the First Crusade, they were severely afflicted by bloody persecutions.

Ashkenazic speculation and asceticism

The main speculative problem was that of the relationship between God in his pure transcendence and total unity and his manifestations in creation, as well as in revelation and communication with inspired men. Reflection on this problem led to the elaboration of various supernatural hierarchies between the inaccessible God and the created universe or the recipient of divine communication; data on angels taken from the Bible and rabbinical and mystical tradition, as well as speculation on the *Shekhina,* were used as material for these hierarchies and also gave a peculiar coloration to liturgical practice. The latter was marked, moreover, by a concern for spiritual concentration by means of fixing the attention on the words and even the letters of the synagogue prayers. Whatever the historical interest of these speculations, they had no great repercussions on the subsequent course of Jewish esoterism; the only exceptions are the mysticism of prayer and demonology, which was sometimes influenced by the beliefs of the Christian environment and fully developed in Ḥasidic circles. On the other hand, the ascetic morality of the movement, which found its literary expression in the work of Eleazar ben Judah of Worms (*c.* 1160–1238) and in the two recensions of the "Book of the Pious" (*Sefer ḥasidim*), was to mark Jewish spirituality, esoteric or not, from then on.

The making of the Zohar (c. 1260–1492). Once the actually marginal episode of German Ḥasidism was finished, almost all of the creative activity in Jewish mysticism was to be situated or would originate in Spain, up to the expulsion of the Jews in 1492.

After the flowering of the schools described above came to an end, around the year 1260, two other currents appeared. The first, in its own manner, resumed relations with Gnosticism in that it placed the problem of evil at the centre of its reflection. The texts that reflect this tendency do not maintain evil in a state of dependence on the "attribute of judgment" within the structure of the *sefirot* set up by the previous Kabbalists but locate it outside the divinity, constructing a parallel system of "left-hand *sefirot,*" with a corresponding development of an exuberant demonology. The second movement, whose main representative was the 13th-century visionary-adventurer Abraham ben Samuel Abulafia, found its justification in inner experiences considered "prophetic" and encouraged by training methods akin to those of Yoga, the Byzantine Hesychasts (mystical, quietist monks), and the Muslim Ṣūfīs (mystics); moreover, an important place was given to speculations on the letters and vocalic signs of the Hebrew script. Unlike the protagonists of other mystical schools of Spain that until then had not sought to spread their ideas outside the circle of initiates, Abulafia applied himself in various places to propaganda and exhibitions that disaffected and worried the leaders of Judaism and caused their initiator to be pursued even by the non-Jewish authorities. The numerous writings that he left were later to stimulate a few minds among the Kabbalists.

The world of Moses de León

The work of Moses ben Shem Tov de León, in the last quarter of the 13th century, marked one of the most important turning points in the development of Jewish mysticism. Moses de León was born in the middle of the 13th century and died in 1305; he was the author of several esoteric works, which he signed with his own name. But at the same time, in order to better spread his ideas and to more effectively combat philosophy, which he considered a mortal danger to the Jewish faith, he turned to the composition of pseudepigrapha (writings ascribed to other authors, usually in past ages) in the form of Midrashim (plural of Midrash) on the Pentateuch, the Song of Solomon, Book of Ruth, and Lamentations, in which Talmudic authorities appeared, of whom only the names were even partially authentic, a procedure already used by the *Sefer ha-bahir* (see above *Sefer ha-bahir*); in its most finished version (for there were several of them), the plot of the tales centred around Rabbi Simeon ben Yoḥai, a doctor of the 2nd century, about whom the Talmud already related some curious anecdotes, most of them semilegendary. Moses de León thus produced, over a period of about 30 years, first a work entitled *Midrash ha-neʿelam* ("The Mystical Midrash") whose method was largely allegorical and whose tongue was mainly Hebrew, and then a larger work, the *Sefer ha-zohar* ("Book of Splendour"), or more briefly the *Zohar,* whose content is theosophic and which was written in artificial Aramaic. The book culminates in a long speech in which Simeon ben Yoḥai, on the day of his death, supposedly exposes the quintessence of his mystical doctrine. The literary hoax of Moses de León was not immediately accepted as authentic by all the esoterists and still less by scholars outside the theosophic movement; it took half a century or more for the *Zohar* and imitations of it to be recognized as authoritative ancient works, and even then it was not without some reluctance. The nearly contemporary imitations of the *Zohar* that were incorporated into it or appended to it were sometimes of a markedly different ideological orientation: the *Raʿya Mehemana* ("Faithful Shepherd"—that is to say, Moses, who is the central figure of this composition, the particular subject of which is the interpretation and theosophic justification of the precepts of the Torah); and the *Tiqqune Zohar,* elaborations in the same vein bearing upon the first word of the book of Genesis (*Bereshit,* "In the beginning"). Although critics were never fully silenced and the authenticity of the *Zohar* was already questioned in the 15th century, the myth created by Moses de León and his imitators became a spiritual reality for the majority of believing Jews; it still retains this character among "Orthodox" Jews. The *Zohar,* believed to be based on supernatural revelations and reinterpreted in diverse ways, would serve as support and reference for all the Jewish theosophies in the centuries ahead.

Doctrine and symbolism of the *Zohar*

As to doctrine, the *Zohar* and its appendixes develop, amplify, and exaggerate speculation and tendencies that already existed, rather than offering any radical innovation. All of the ideas had already been accepted for a long time in Jewish theosophy: the springing forth of being from the depth of the divine "nothing"; the solidarity of the world of the *sefirot* (complicated by the introduction of four ontological levels at each one of which the schema of the 10 *sefirot* is reproduced) with the visible world; the indispensable contribution of man (that is, of the Jew) who observes the biblical and rabbinical precepts in their slightest details, to universal harmony—these emphases remain the main lines of the *Zohar.* But all these themes (the speculations of the *Sefer ha-Temuna,* mentioned above, on the cosmic cycles and the "Prophetic Kabbala" of Abulafia being tacitly set aside) were largely organized and enhanced by the use, or rather the unscrupulous appropriation, of materials taken from rabbinical tradition and ancient esoterism as well as from more recent theological and philosophical currents of thought, despite the lack of esteem that the writers of the Zoharic corpus felt and sought to make others feel toward works created by gentiles.

The method of symbolic representation used by the writings of the Zoharic corpus was supported by a system of interpretation that made use of the originally Christian concept of the fourfold meaning of Scripture: literal, moral, allegorical (philosophical), and mystical. The symbolism thus set up boldly made use of an exuberant anthropomorphic and even erotic imagery whose function was to convey the manifestation of the levels of the *sefirot* to each other and to the extradivine world. The myth of the primordial man (Adam Qadmon), a virtually divine being, reappeared here under a new form, and it was to remain in the subsequent development of Kabbala.

The *Zohar* thus claims to provide a complete explanation of the world, man, history, and the situation of the Jew; on a higher level, to justify the biblical revelation and rabbinical tradition, down to the slightest detail, including the messianic expectation; and thereby to neutralize philosophy. But, while setting itself up as the defender of the traditional religion regulated by the Talmud and its commentaries, in a sense it places itself above tradition, by proclaiming boisterously the incomparable value of the theosophic teaching of "Rabbi Simeon ben Yoḥai" and the superiority of the esoteric doctrine over the Talmudic studies, which were open to all and which, along with the observance of the precepts, were, according to common opinion, supposed to constitute the basic justification of the life of the Jew. There is in this attitude—more accentuated in the *Raʿya Mehemana* (see above) than in the *Zohar* proper—a revolutionary potentiality, a possible threat to the primacy of practice and study of Torah; the future would show that this danger was not completely unreal.

The Lurianic Kabbala. After the establishment of the Zoharic corpus, no major changes took place in Jewish esoterism until the middle of the 16th century, when in Safed (in Upper Galilee, Palestine; present-day Ẓefat, Israel) a religious centre of extreme importance for Judaism was established, which was mainly inspired by teachers coming from families expelled from Spain. Until the expulsion of the Jews from Spain (1492) and during the two generations that followed it, the Kabbalistic literary output had certainly been abundant, in Spain till the expulsion as well as in Italy and the Middle East; but it was primarily a matter of systematizing or even popularizing the *Zohar* or of extending the speculation already developed in the 13th century; there were also some attempts at reconciling philosophy and Kabbala. It should be noted that even the traditionalist theologians adopted a careful and rather reserved attitude toward Kabbala.

The tragedy for Judaism of the expulsion from Spain and of the forced conversions to Christianity that preceded it by a century, and which would become even more extensive in Portugal shortly afterward, deeply marked the victims. These events, accentuating the already existing pessimism in response to the situation of the Jewish people dispersed among the nations, intensified the messianic expectation. This expectation does not seem to have been unrelated to the beginnings of the printed transmission of Kabbala—the first two printed editions of the *Zohar* date from 1558. All these factors, joined with certain internal developments of speculative Kabbala in the 15th century, prepared the ground for the new theosophy inaugurated by the teaching of Isaac ben Solomon Luria, who was born in Jerusalem in 1534, educated in Egypt, and died in Safed in 1572; although his teaching is traditionally associated with Safed, he spent only the last three years of his life there. Luria wrote very little; his doctrine has been transmitted, amplified, and probably somewhat distorted through the works of his disciples, of which the main one was Ḥayyim Vital (1543–1620), who wrote *ʿEtz Ḥayyim* ("Tree of Life"), the standard presentation of Lurianic Kabbala.

The teaching of Isaac Luria

The theosophy of Luria, whose novelty was proclaimed by its creator and perfectly realized by the esoterists who held to the Zoharistic Kabbala (organized and codified precisely in Safed, during the lifetime of Luria, by Moses ben Jacob Cordovero, 1522–70), is of extreme complexity in its details, although basically it is but one more attempt to reconcile divine transcendence with immanence and to bring a solution to the problem of evil, which the believer in the divine unity can recognize neither as a power existing independently of God nor as an integral part of him.

The theosophic vision of Luria is expressed in a vast mythical construct, which is typologically akin to certain Gnostic and Manichaean (3rd-century dualistic) systems but which strives at all costs to avoid dualism. The essential elements of this myth are as follows: the withdrawal (*tzimtzum*) executed by the divine light, which originally filled all things, in order to make room for the extradivine; the sinking, as a result of a catastrophic event that occurred during this process, of luminous particles into matter (*qelippot,* "shells," a term already used in Kabbala to designate the evil powers); whence the necessity of saving these particles and returning them to their origin, by means of "repair" or "restoration" (*tiqqun*). This must be the work of the Jew who not only lives in complete conformity to the religious duties imposed on him by tradition but who also dedicates himself, in the framework of a strict asceticism, to a contemplative life founded on mystical prayer and the directed meditation (*kawwana*) of the liturgy, which is supposed to further the harmony (*yiḥud,* "unification") of the innumerable attributes within the divine life. The successive reincarnations of the soul, a constant theme of Kabbala that Lurianism developed and made more complex, are also invested with an important function in the work of "repair." In short, Lurianism proclaims the absolute requirement of an intense mystical life with, as its negative side, an unceasing struggle against the powers of evil. Thus it presents a myth that symbolizes the origin of the world, its fall, and its redemption; it gives meaning to the existence and to the hopes of the Jew, not merely exhorting him to a patient surrender to God but moving him to a redeeming activism, which is the measure of his sanctity. Obviously, such requirements make the ideal of Lurianism possible only for a small elite; ultimately, it is realizable only through the exceptional personage of the "just"—the ideal holy Jew described above.

Shabbetaianism. During the 60 years that followed the death of the founder, the Kabbala linked to the name of Luria and overlaid with accretions from the other mysticisms of Safed spread through the Jewish Diaspora and deeply permeated its spiritual life, liturgy, and devotional practices. It emphasized the necessity of "repair" of a world in which the uneasiness of the Jew kept growing, for in spite of certain favourable factors—the relative tolerance of the Ottoman Empire and the peaceable establishment of an important Marrano (Iberian Jewish, or Sefardic) community in Amsterdam—there was no overall solution to the problem of the "conversos" (converts) who had remained in the Iberian Peninsula. The other half of the Jewish people, the Ashkenazim, also experienced a serious crisis: its most prosperous and dynamic section, the Jewish population of Poland, was sorely tried, almost totally ruined, and in large part forced to move back toward the west because of the massacres and the destruction that took place during the Cossack uprising of 1648. These ideological and historical data may provide the necessary context for understanding the astonishing though short-lived success of Rabbi Shabbetai Tzevi of Smyrna (1626–76), who proclaimed himself messiah in 1665. The "Messiah" was forcibly converted to Islām in 1666 and ended his life in exile 10 years later, but despite his failure he had faithful followers. A sect was thus born and survived largely thanks to the activity of Nathan of Gaza (*c.* 1644–90), an unwearying propagandist for the "Messiah," who justified the actions of Shabbetai Tzevi, which were contrary to the Law, and his final apostasy by theories that were based on the Lurian theory of "repair": it had to be understood as the descent of the just into the abyss of the "shells" in order to liberate from it the captive particles of divine light. The Shabbetaian crisis lasted nearly a century, some of its aftereffects even longer. It led to the formation of sects whose members were externally converted to Islām—*e.g.,* the Dönme (Turkish, "apostates") of Salonika, whose descendants still live in Turkey—or to Roman Catholicism—*e.g.,* the Polish supporters of Jacob Frank (1726–91), the self-proclaimed Messiah and Catholic convert. In Bohemia–Moravia, however, the Frankists outwardly remained Jews. This crisis did not discredit Kabbala, but it led the spiritual authorities of Judaism to watch over and severely curtail its spread and to exercise rigorous ideological control, by concrete acts of censorship and repression, over anyone, even a person of tested piety and recognized knowledge, who was suspected of Shabbetaian sympathies or of messianic pretensions.

Shabbetai Tzevi, a 17th-century "Messiah"

Modern Ḥasidism. Though it is true that the messianic movement centred around Shabbetai Tzevi could only produce disillusionment and that if it had not been contained it could have led Judaism to its ruin, yet it answered not only the theosophic aspirations of a small number of visionary scholars but also an affective need of the Jewish masses that was left unsatisfied by the dry intellectualism of the Talmudists and the economic and social oppression of the ruling classes (both Jewish and non-Jewish). This was the case especially in Poland, which before the partition of the Polish kingdom (1772–95) included Lithuanian, Belorussian, and Ukrainian territories. It was there that the so-called Ḥasidic movement, in no way connected with medieval German Ḥasidism, originated around the middle of the 18th century—a movement in which the Lurian Kabbala, theoretically maintained as the basis of speculation, underwent adjustments and transformations that continue to the present day.

The Ba'al Shem Ṭov

If modern Ḥasidism may be regarded as a mass movement, having a minimum of organization, using the methods of propaganda and preaching, and forming groups of acknowledged members, then the legend is credible that traces it back to a single founder, Israel ben Eliezer, known as Ba'al Shem Ṭov (Master of the Good Name; that is, a possessor—he was not the only one of his kind—of the secret of the ineffable name of God, which bestows an infallible power to heal and perform other miraculous operations). This man was born about 1700 and died in 1760 in southern Poland. Though relatively untrained according to the norms of the rabbinical Judaism of his time, he was a spiritual personage of exceptional quality and was able to win to his ideas not only the common people but also many representatives of the intellectual elite. The mist of legend that surrounds him is too dense for it to be possible to reconstruct entirely his personal doctrine, which he probably never systematized. Drawing his inspiration from the methods of the itinerant preachers whose activity was becoming more intense in 18th-century eastern European Judaism, he delivered his teaching in the form of homiletic interpretations of sacred texts, having recourse to fables and parables borrowed from daily life and from folklore; this method remained constant in Ḥasidism, but it is undeniably an exaggeration and even an error in perspective to consider, as did Martin Buber (see below *Modern Jewish mysticism*), that the tale and the anecdote are the most authentic expression of the doctrine and the spirituality of Ḥasidism. It is indeed in the doctrinal works, most of them expressed in the form of sermons on the weekly sections of the Pentateuch and other liturgical lessons, that the thought of the Ḥasidic "rabbis" is expressed. It is very diversified thought, for there are as many bodies of doctrine in Ḥasidism as there were creative spirits during the first three generations of the movement. It is, nevertheless, possible to point to a few traits that are fundamental and common to Ḥasidism as a whole.

In theory, it remains rooted in the Lurianic Kabbala—and nothing essential separates it at this point from its most implacable adversaries in the traditional Judaism of eastern Europe. What is unique to it is to have made of *devequt,* "being-with-God," an object of aspiration and even a constant duty for all Jews and in all circumstances of life, even those seemingly most profane; in other words, it demands a total spiritualization of Jewish existence. This requirement entails a reevaluation, less new in its principle than in its concrete application, of the speculative concepts of Kabbala: the emphasis is placed on the inner life of the believer, and it is at this level that the supercosmic drama is played (a drama whose stage was, according to bookish theosophy, in the universe of the *sefirot*); according to several teachers, the same emphasis on inwardness holds for messianic redemption. At the same time, Ḥasidism transforms into social reality a requirement that was also part of the Lurian doctrine of "repair," though it was unfortunately distorted by Shabbetaianism: Ḥasidism puts at the centre of the religious life and organization of the group, as an indispensable guide and unquestioned authority, the inspired leader, endowed with supernatural powers—the "just" (*tzaddiq*), the "miracle-working rabbi" (*Wunder-rebbe*). Ḥasidism thus produced, wherever it triumphed, an undeniable spiritual renewal; the reverse of the medal was the cult of personality, competition between "dynasties" of "rabbis," obstinacy in maintaining the Ḥasidic community apart from the surrounding society, with all the social and economic consequences that this will to isolation entailed and for which it would be false to lay all the blame on the environment, despite its definite hostility toward the Jews.

The *tzaddiq*

From its very beginnings, Ḥasidism was to encounter strong resistance on the part of the official Judaism of the period, which had been sensitized to the anarchism of the Shabbetaians and was at the same time solicitous for the prerogatives of the established community leaders and rabbis, the vigilant guardians over the traditional laws and their application, who were confined to the formal study of the Talmud and its commentaries. The behaviour of the followers of Ḥasidism, though irreproachable in its strict, even rigorous observance of ritual rules, displayed several traits that were distasteful to its adversaries (besides the unconditional submission to the *tzaddiq,* who often doubled as the rabbi of the official congregation): desertion of the general communal synagogues, meetings in small conventicles, modifications of the liturgy, casual dress during prayer, and preference given to mystical meditation rather than to the dialectical study of the Talmud, which requires instead serious intellectual concentration. Nevertheless, the conflict between the Ḥasidim and the "Opponents" (Mitnaggedim) did not finally degenerate into a schism; after three generations, a kind of tacit compromise was established between the two tendencies—Ḥasidic and Talmudic—without the consciousness of differences ever being erased. The compromise was rather to the advantage of Ḥasidism, but not without a few concessions on its part, notably on the question of education.

The strong organization of the Ḥasidic groups allowed them to survive the dislocation of eastern European Judaism as a result of the events of World War II, but its vital centres are today in the United States rather than in Palestine, in part because of economic reasons, in part because of the more or less reserved, and at sometimes frankly hostile, attitude of the Ḥasidic "rabbis" toward political Zionism and the State of Israel. The best known of the U.S.-based groups is the very active Lubavitchers (after Lyubavichi, Russia, seat of a famous school of Ḥasidism), whose headquarters are in the Crown Heights district of Brooklyn, New York.

MODERN JEWISH MYSTICISM

The role played by Kabbala and Ḥasidism in the thought and spirituality of contemporary Judaism is far from being insignificant, though its importance is not as great as in former times. Of course, there is hardly any really living Kabbalistic and Ḥasidic literature, but the personal thought of religious writers such as Abraham Isaac Kook (*c.* 1865–1935), spiritual leader, mystic, and chief rabbi of Palestine, continues to exercise a marked influence. Furthermore, the renewal of religious thought in "westernized" Jewish circles between the two wars received a powerful impulse from the philosopher Martin Buber (1878–1965), whose work is in part devoted to the propagation of Ḥasidic ideology as he understood it. "Neo-Orthodoxy," founded in Germany by Samson Raphael Hirsch (1808–88), was quite indifferent to mysticism at the outset, but it too came to be influenced by it, especially after the rediscovery of living Judaism in Poland during World War I by Western Jewish thinkers. Also significant is the work of Abraham Joshua Heschel (1907–72), a Polish Jewish writer of distinguished Ḥasidic background and double culture—traditional and Western—who emigrated to the United States.

Christian and secular Kabbala

Jewish mysticism also has exercised some influence on thought outside the Jewish community. Kabbala, distorted and deflected from its own intentions, transcended the frontiers of Judaism and helped nourish and stimulate certain currents of thought in Christian society, from the Renaissance to the present: "Christian Kabbala," born in the 15th century under the impetus of Jewish converts from Spain and Italy, claimed to find in the Kabbalistic documents, touched up if necessary or even forged, arguments for the truths of the Christian faith. Thus a certain number of Christian Humanist scholars became interested in Jewish mysticism and several of them acquired a fairly extensive knowledge of it on the basis of authentic texts. Among them were Giovanni Pico della Mirandola (1463–

94) and Gilles of Viterbo (Egidio da Viterbo; *c.* 1465–1532) in Italy, and Johannes Reuchlin (1455–1522), who was responsible for writing one of the principal expositions of Kabbala in a language accessible to the learned non-Jewish public (*De arte Cabbalistica,* 1517), in Germany, while the visionary Guillaume Postel (1510–81) was attracting disciples in France. The occult philosophy of the 16th century, the "natural philosophy" of the 17th and 18th centuries, and the occult and theosophic theories that are cultivated even today and that have coloured the ideology of Freemasonry—all of these focus and continue to make borrowings from Kabbala, though they rarely grasp its spirit and meaning. The same is true of most of the books on Kabbala put out by publishers of occult and theosophic literature today.

The rigorous scholarly study of Jewish mysticism is a very recent phenomenon. The state of mind and the tendencies of the founders of the "science of Judaism" (the scholarly study of Jewish religion, literature, history, etc.) in Germany during the first half of the 19th century were too permeated with rationalism to be favourable to scholarly investigation of a movement judged to be obscurantist and retrograde. Granting some valuable earlier works, research on a large scale and application of the proved methods of philology and history of religions began only with the work of Gershom G. Scholem, who was professor of Kabbala at the Hebrew University, Jerusalem from 1923 to 1965, and has been continued by his disciples, both direct and indirect. This research touched all of the areas of Jewish mysticism that are briefly described in this article; however, the gaps in knowledge remain serious in every area. Critical editions of mystical texts are few in number; unpublished documents are cataloged in a very incomplete manner; and only a few monographs on writers and particular themes exist, though these are indispensable preliminaries to a detailed and thorough synthesis. It is to be hoped that the one outlined by Scholem in 1941, in his *Major Trends in Jewish Mysticism,* though of exceptional value in its time, will be taken up again and completed. (G.V.)

Jewish myth and legend

Jewish myth and legend comprises a vast body of stories transmitted over the past 3,000 years in Hebrew and in vernacular dialects, such as Yiddish (Judeo-German) and Ladino (Judeo-Spanish), spoken by Jews in various parts of the world. These stories have played an important role in the history of Jewish religion and culture.

SIGNIFICANCE AND CHARACTERISTICS

Apart from their intrinsic appeal, Jewish myths and legends claim attention for three special reasons: (1) Those incorporated in the Old Testament now form part and parcel of the cultural heritage of the Western world and have exerted a profound influence on its literature and art. (2) During the Middle Ages Jews were among the principal transmitters of Oriental tales to the West, so that many familiar Eastern stories can be traced to Jewish compilations. (3) Since these stories have been accumulated through centuries of constant migration, they provide an unrivalled body of "clinical" material for studying the process by which popular tales in fact travel and are transformed.

Basic character and role

Not all of the stories are of Jewish origin; many can be readily paralleled elsewhere and are derived from tales the Jews picked up from their non-Jewish neighbours in the lands of their dispersion. Even what is borrowed, however, is usually impressed with a distinctive Jewish stamp, being adapted to point up some precept of the Jewish religion, to illustrate some facet of Jewish life, or to exemplify some trait of Jewish character and temperament. The dominant overall feature of the stories is, indeed, their religious and moral tone; most of them are, in fact, told specifically as part of the homiletic exposition of Scripture. Such stories are taught to Jews from early childhood as a regular part of their religious education. To the tradition-minded Jew, therefore, they are more than mere literary fancies and assume a kind of doctrinal complexion. Biblical characters and events present themselves to him more in the lineaments of later legend than in their original biblical form; while popular notions about heaven and hell, rewards and punishments, the coming of the Messiah, and the resurrection of the dead derive mainly from this source rather than from Scripture itself.

Virtually all the standard types of folktale are represented. Conspicuously absent, however, are pure fairy tales because fairies, elves, and the like are foreign to the Jewish imagination, which prefers to people the otherworld with angels and demons subservient to God.

A distinction must be made, of course, between myth and legend. In common parlance, a myth is a story about gods or otherworldly beings. Judaism, however, is a rigorously monotheistic religion; hence, in this narrower sense, there can be no original Jewish myths. Nevertheless, from the earliest times, Jews have not disdained to borrow those of their pagan neighbours and then adapt them to their own religious outlook. If, however, the term is interpreted in a larger sense, to mean the portrayal of continuous, transtemporal concerns in the context of particular and punctual events, myth is indeed one of the essential vehicles by which Judaism conveys its message; for it is only when historical happenings are translated into this wider dimension that they cease to be mere antiquarian data and acquire continuing relevance. In Judaism, for example, the Exodus from Egypt is projected mythically from something that happened at a particular time into something that is continually happening, and it thus comes to exemplify the situation and experience of all men everywhere—their emergence from the bondage of obscurantism, their individual revelations at their individual Sinais, their trek through a figurative wilderness, even their death in it so that their children or children's children may eventually reach the figurative "promised land." By the same token, the historical destruction of the Temple of Jerusalem is transformed by myth into a paradigm of the continuing mutual estrangement of God and man, their exile from one another.

Legend, on the other hand, implies no more than a fanciful embroidering of purportedly historical fact. Unlike myth, it does not transcend the punctual and local.

SOURCES AND DEVELOPMENT

Myth and legend in the Old Testament. The vast repertoire of Jewish myths and legends begins with the Old Testament. Their overall purpose in Scripture is to illustrate the ways of God with man, as exemplified both in historical events and in personal experience. The stories themselves are often derived from current popular lore and possess abundant parallels in other cultures, both ancient and modern. In each case, however, they are given a peculiar and distinctive twist.

Myths in Genesis

Myths. Old Testament myths are found mainly in the first 11 chapters of Genesis, the first book of the Bible. They are concerned with the creation of the world and of man, the origin of the continuing human condition, the primeval Deluge, the distribution of peoples, and the variation of languages.

The basic stories are derived from the popular lore of the ancient Middle East and can be paralleled in the extant literature of the peoples of the area. The Mesopotamians, for instance, also knew of an earthly paradise such as Eden, and the figure of the cherubim—properly griffins rather than nightgowned angels—was known to the Canaanites. In the Bible, however, this mythical garden of the gods becomes the scene of man's fall and the background of a story designed to account for the natural limitations of human life. Similarly, the Babylonians, too, told of the formation of man from clay, but in the scriptural version his function is to bear rule over all other creatures, whereas in the pagan tale it is to serve as an earthly menial of the gods. Again, the story of the Deluge, including the elements of the ark and the dispatch of the raven and dove, appears already in the Babylonian myths of Gilgamesh and Atrahasis. There, however, the hero is eventually made immortal, whereas in the Bible this detail is omitted because to the Israelite mind no child of woman could receive that status. Lastly, while the story of the Tower of Babel was told originally to account for the stepped temples (ziggurats) of Babylonia, to the Hebrew writer its purpose is simply to inculcate the moral lesson that man should not build beyond his assigned station.

Myths in the Prophets and Holy Writings

Scattered through the Prophets and Holy Writings (the two latter portions of the Hebrew Bible) are allusions to other ancient myths—*e.g.,* to that of a primordial combat between Yahweh and a monster variously named Leviathan (Wriggly), Rahab (Braggart), or simply Sir Sea or Dragon. The Babylonians told likewise of a fight between their god Marduk and the monster Tiamat; the Hittites told of a battle between the weather god and the dragon Illuyankas; while from Ras Shamra (ancient Ugarit), in north Syria, has come a Canaanite poem relating the discomfiture of Sir Sea by the deity Baal and the rout of an opponent named Leviathan. (Originally, this myth probably referred to the annual subjugation of the floods.)

Ancient myths are utilized also in the form of passing allusions or poetic "conceits," much as modern Westerners may speak of Cupid or the Muses. Thus, there are references in the prophetic books to a celestial upstart hurled to Earth on account of his brashness and to the imprisonment of certain rebellious constellations.

The prophets used such myths paradigmatically to illustrate the hand of God in contemporary events or to reinforce their forecasts. Thus, to Isaiah the primeval dragon becomes the symbol of that continuous force of chaos and evil that will again have to be vanquished before the Kingdom of God can be established on Earth. Similarly, for Ezekiel the celestial upstart serves as the prototype of the prince of Tyre, destined for an imminent fall; and Habakkuk sees in the impending rout of certain invaders a repetition on the stage of history of Yahweh's mythical sortie against the monster of the sea.

Legends and other tales. Old Testament legends often embellish the accounts of national heroes with standard motifs drawn from popular lore. Thus, the story (in Genesis) of Joseph and Potiphar's wife recurs substantially (with other characters) in an Egyptian papyrus of the 13th century BCE. The depositing of the infant Moses in the bulrushes (in Exodus) has an earlier counterpart in a Babylonian tale about Sargon, king of Akkad (*c.* 2334–*c.* 2279 BCE), and is paralleled later in legends associated with the Persian Cyrus and with Tu-Küeh, the fabled founder of the Turkish nation. Jephthah's rash vow (in Judges) whereby he is committed to sacrifice his daughter recalls the classical legend of Idomeneus of Crete, who had similarly to slay his own son. The motif of the letter whereby David engineers the death in battle of Bathsheba's husband recurs in Homer's story of Bellerophon and again in the episode of Rosencrantz and Guildenstern in *Hamlet.* The celebrated judgment of Solomon concerning the child claimed by two contending women is told, albeit with variations of detail, about Buddha, Confucius, and other Oriental sages; while the story of how Jonah was swallowed by a "great fish" but subsequently disgorged intact finds a parallel in the Indian tale of the hero Śaktideva, who experienced the same thing during his quest for the Golden City. On the other hand, it should be observed that many of the parallels commonly cited from the folklore of primitive peoples may be, in fact, mere playbacks of biblical material picked up from Christian missionaries.

Etiologic tales

Sometimes, worldwide folktales serve in the Old Testament to account for the names of places in Palestine or for the origins of traditional customs and institutions. Thus, the familiar story of the man who has to struggle with the personified current of a river before he can cross it is localized (in Genesis) at the ford of Jabbok simply because that name suggests the Hebrew word *abḳ* ("struggle"); and Samson's felling of 1,000 Philistines with the jawbone of an ass is placed at Ramath-leḥi because *leḥi* is Hebrew for "jawbone." Similarly, a taboo against eating the sciatic nerve of an animal is validated (in Genesis) by the legend that Jacob was struck in the hip when he tussled with an otherworldly being at Penuel (Face of God); and the custom of annually bewailing the vanished spirit of fertility is rationalized (in Judges) as a lamentation for the hapless daughter of Jephthah.

Besides myths and legends the Old Testament also contains a few examples of fables (didactic tales in which animals or plants play human roles). Thus, the serpent in Eden talks to Eve, and Balaam's ass not only speaks but also "flairs" spirits; while in the celebrated parable of Jotham (in Judges) trees compete for kingship.

Finally, in the Book of Job (38:31) there are allusions to star myths concerning the binding of Orion (called the Fool) and the "chaining" of the Pleiades.

Contemporary interpretations. The tendency to interpret biblical tales and legends as authentic historical records or as allegories, or as the relics of solar, lunar, and astral myths, is now a thing of the past. For the modern folklorist, their primary interest lies in the fact that they push back to remote antiquity several tales and motifs long known from later literature. For the theologian, however, they pose the deeper problem of distinguishing clearly between the permanent message of Scripture and the particular form in which it is conveyed. Such a process of "demythologization" is today one of the central concerns of religious thought. It involves recognition of the fact that the natural language of religious truth is myth so that the continuing relevance of ancient scriptures depends not on a total rejection of that vehicle but rather on a constant expansion and remodelling of it—*i.e.,* on remythologization rather than demythologization. In the final analysis, the traditional portrayal of God himself is simply a mythical representation of ultimate reality, but that reality transcends the particular images in which it happens to be expressed. At the same time, it must be clearly understood that expressions that can be reconciled with modern Western patterns of thought only if taken as metaphors were literal statements of fact to ancient and primitive peoples. Gods, for example, were not merely "personifications" of natural phenomena but rather the effective potencies of the phenomena themselves conceived from the start as personal beings, much as a modern child might conceive of a railroad engine as "Mr. Choochoo."

Angelology, demonology, and eschatology

Myth and legend in the Persian period. When, in 539 BCE, the Jews came under Persian domination, they absorbed a good deal of Iranian folklore about spirits and demons, the eventual dissolution of the world in a fiery ordeal, and its eventual renewal. This introduced a new element into Jewish myth and legend. Hierarchies of angels, archangels such as Michael, Gabriel, and Uriel (modelled loosely upon the six Iranian spiritual entities, the *amesha spentas*), and the demonic figures of Satan, Belial, and Asmodeus (corresponding to the Iranian Angra Mainyu [Ahriman], Druj, and Aēshma daeva) now entered their popular mythology, and there was a preoccupation with apocalyptic visions of heaven and hell and of the Last Days. Unfortunately, no Jewish texts of this genre from the Persian period itself are extant so that these new elements can be recognized only inferentially from their survival in later times, notably in such products of the ensuing Hellenistic age as the Dead Sea Scrolls.

The principal monument of Jewish story in the Persian period is the biblical Book of Esther, and this is basically the Judaized version of a Persian novella about the shrewdness of harem queens. The story was adapted to account for a popular festival named Purim, but this is probably a transmogrification of the Persian New Year. Such leading elements of the tale as the parade of Mordecai through the streets dressed in royal robes, the fight between the Jews and their adversaries, and the hanging of Haman and his sons seem, indeed, to reflect customs associated with that occasion, viz., the ceremonial ride of a common citizen through the capital, the mock combat between two teams representing Old Year and New Year, and the execution of the Old Year in effigy.

Myth and legend in the Hellenistic period. *Historiated Bibles and legendary histories.* When, in 330 BCE, Alexander the Great completed his conquest of the Middle East, Judaism entered a new phase. The dominant features of the ensuing Hellenistic age were an increasing cosmopolitanism and a fusion of Oriental and Greek cultures. These found expression in Jewish myth and legend in the composition (in Greek) of stories designed to link the Bible with general history, to correlate biblical and Greek leg-

"Esther and Ahasuerus," tempera painting of Konrad Witz, 15th century. In the Öffentliche Kunstsammlung Basel. 85.5 × 79.5 cm.

By courtesy of the Offentliche Kunstsammlung Basel, Switzerland

ends, and to claim for the Hebrew patriarchs a major role in the development of the arts and sciences. It was asserted, for instance, that Abraham had taught astrology to the king of Egypt; that his and Keturah's sons had aided Heracles against the giant Antaeus; and that Moses, blithely identified both with the semi-mythical Greek poet Musaeus and with the Egyptian Thoth, had been the teacher of Orpheus (putative founder of one of the then current "mystery cults") and the inventor of navigation, architecture, and the hieroglyphic script. Leading writers in this vein were Artapanus, Eupolemus, and Cleodemus (all *c.* 100 BCE), but their works are known to us only from stray quotations by Eusebius and Clement of Alexandria, early Church Fathers.

Novelistic versions of biblical figures

Furthermore, the Jews followed a current Greek literary fashion of retelling Homeric and other ancient legends in "modernized," novelistic versions, well seasoned with romantic elaborations. Among the Dead Sea Scrolls has been found a paraphrase of Genesis in which the biblical narrative is tricked out with several familiar folklore motifs. Thus, when Noah is born, the house is filled with light, just as it is said elsewhere to have been at the birth of the Roman king Servius Tullius, of Buddha, and (later) of several Christian saints. When Abraham's life is threatened he dreams of a cedar about to be felled—the same omen said to have presaged the deaths of Domitian and Severus Alexander. (True, the parallels are of later date, but they illustrate the persistence of age-old popular traditions.) The same trend toward fanciful elaboration of scriptural tales is manifested also in the *Testaments of the Twelve Patriarchs* ("testaments" meaning last wills), in which the virtues and weaknesses of the sons of Jacob are illustrated by moralistic legends. There is also a lengthy paraphrase of early biblical narratives, mistakenly attributed to Philo, the famous Alexandrian Jewish philosopher of the first century CE.

Apocrypha and Pseudepigrapha. The principal monuments of Jewish literature during the Hellenistic period are the works known collectively as the Apocrypha and Pseudepigrapha. The former are certain later writings excluded by Jews from the canon of the Old Testament but found in the Greek Septuagint version. The latter are other late writings not included in any authorized version of the Scriptures and spuriously attributed to biblical personalities.

Judaized versions of foreign tales

The Apocrypha include several Judaized versions of tales well represented in other cultures. The book of Tobit, for instance, turns largely on the widespread motifs of "The Grateful Dead" and the "Demon in the Bridal Chamber." The former relates how a traveller who gives burial to a dishonoured corpse is subsequently aided by a chance companion who turns out to be the spirit of the deceased. The latter tells how a succession of bridegrooms die on the nuptial night through the presence of a demon beside the bridal bed. Similarly, in Bel and the Dragon (2nd century BCE) occurs the equally familiar motif that fraud (in this case perpetrated in a temple) is detected by the imprint of the culprit's foot on strewn ashes—a motif that reappears later in the French and Celtic romance of Tristan and Iseult. Again, *Susanna and the Elders* (also 2nd century BCE) revolves around the well-worn theme that a charge of unchastity levelled against a beautiful woman is refuted when a clever youngster ("Daniel come to judgment") points out discrepancies in the testimony of her accusers. The story has a close parallel in a Samaritan tale about the daughter of a high priest in the 1st century CE; while the motif of the clever youngster who surpasses seasoned judges recurs later in infancy gospels and in the tale of ʿAlī Khamājah in *The Thousand and One Nights.*

The most interesting folktale in the Pseudepigrapha is that contained in *The Martyrdom of Isaiah* (1st century CE?), which tells how the prophet, fleeing from King Manasseh, hid in a tree that opened miraculously and how he eventually perished when it was sawn asunder. A similar tale is related in the Talmud about a certain Isaac ben Joseph and (later) in the Persian epic *Shāh-nāmeh* (*c.* 1000 CE) about the hero Jamshīd.

Myth and legend in Talmud and Midrash. *Midrash and Haggada.* Toward the end of the 1st century CE, through a process known as "canonization," certain traditional Hebrew writings came to be recognized as an authoritative corpus of divine revelation, later called the Hebrew Bible or Old Testament. The study of them became, henceforth, an essential element of the Jewish religion. This meant that the sacred text had to be subjected to a form of interpretation that would bring out its universal significance and permanent relevance. The process was known as Midrash (literally "searching the Scriptures"), and a leading constituent of it was the spicing of homiletic discourses with elaborative legends—a pedagogic device called Haggada ("storytelling"). Originally transmitted orally, the legends were eventually committed to writing in that vast sea of literature known as the Talmud (the authoritative compendium of early rabbinic law and lore), as well as in later compilations geared to particular books or sections of the Old Testament, to scriptural lessons read in the services of the synagogue, or to specific biblical characters or moral themes (see also above *Torah*).

The range of Haggada is virtually inexhaustible; a few representative examples must suffice. In regard to biblical characters, both Moses and David were born circumcised; Cain had a twin sister; Abraham will sit at the gate of hell to reproach the damned on Judgment Day; Aaron once locked the angel of death in the tabernacle; Solomon understood the language of animals; King Hiram, who supplied materials for the Temple, entered paradise alive; the flesh of Leviathan will feed the righteous in the world to come.

Indebtedness of Haggada to non-Jewish sources

In such fanciful elaborations of Scriptures, Haggada does not disdain to draw on classical tales. The men of Sodom, it is said, subjected itinerant strangers to the ordeal of Procrustes' bed; the Earth opened to rescue newborn Hebrew males from the Pharaoh, as it did for Amphiaraus, the prophet of Argos, when he fled from Periclymenus after the attack on Thebes; Moses spoke at birth, as did Apollo; Solomon's ring, cast into the river, was retrieved from a fish that had swallowed it, as was that of Polycrates, the tyrant of Samos, in the story told by Herodotus; the Queen of Sheba had the feet of an ass, like the child-stealing witch (Onoskelis) of Greek folklore; no rain ever fell on the altar at Jerusalem, just as none was said to have fallen on Mt. Olympus.

Other familiar motifs also appear. Moses qualifies as a husband for Zipporah by alone being able to pluck a rod from Jethro's garden—a variant of the tale told later about the sword Excalibur in the Arthurian legend; David's harp is played at night by the wind, like that of Aeolus; Isaiah, like Achilles and Siegfried, has only one vulnerable spot in his body—his mouth; Job has a magic belt, which relieves his pains.

Legends are developed also from fanciful interpretations of scriptural verses. Thus, Adam is said to have fallen only a few hours after his creation because the Hebrew text of Ps. 49:12 can be literally rendered "Adam does not last the night in glory." Lamech slays the wandering Cain—a fanciful interpretation of his boast in Gen. 4:23–24. Melchizedek is immortal in view of Ps. 110:4: "You are a priest for ever after the order of Melchizedek." The first man is a hermaphrodite (this notion has analogues elsewhere) because Gen. 1:27 says of God's creation, "Male and female he created them."

Fables and animal stories. Midrash also uses fables paralleled in non-Jewish sources. Aesop's fable of the "Lion and the Crane" is quoted by a rabbi of the 1st century CE, and the tales of the "Fox in the Vineyard" and of the "Camel Who Got Slit Ears for Wanting Horns" likewise make their appearance.

Sometimes, too, material is drawn from medieval bestiaries (manuals on animals, real or imaginary, with symbolic or moralistic interpretations). Bears, we are told, lack mother's milk; hares and hyenas can change sex; only one pair of unicorns exists at a time; there is a gigantic bird (*ziz*) that reaches from Earth to sky.

Contribution of Haggada to Christian and Islāmic legends. Several of the stories related in Haggadic literature were later adopted and adapted by Christian writers. Thus, the legend that Adam was created out of virgin soil was taken to prefigure the fact that the second Adam (*i.e.,* Jesus) was likewise born of a virgin; while the story that the soil in question was taken from the site of the future Temple was transformed into the claim that Adam had been molded out of the dust of Calvary. Similarly, the legend that, at the dedication of the Temple, the doors had swung open automatically to admit the ark of the Covenant was transferred to the consecration of a church by St. Basil; and the Talmudic tale that the bronze Nicanor gates of the Temple had floated to Jerusalem when cast overboard for ballast during their shipment from Alexandria was applied to the doors of a sacred edifice erected in honour of St. Giles.

Nor was it only the Christians who absorbed Haggadic legends. The Qurʾān, the sacred book of Islām, likewise incorporates a good deal of such material in its treatment of such biblical characters as Joseph, Moses, David, and Solomon.

Myth and legend in the medieval period. *Jewish contribution to diffusion of folktales.* The Middle Ages was a singularly productive period in the history of Jewish myth and legend. Jews now began to play a prominent role in the transmission of Oriental tales to the West and thereby enhanced their own repertoire with a goodly amount of secular material. Especially in Spain and Italy, Arabic versions of standard collections were translated into Hebrew and thence into Latin, thus spreading the stories to the Christian world. The Indic fables of Bidpai, for example, were rendered into Hebrew from the 8th-century Arabic version of ʿAbd Allāh ibn al-Muqaffaʿ, and from this Hebrew rendering there subsequently developed, in the 12th century, John of Capua's *Directorium humanae vitae* ("Guide for Human Life"), one of the most celebrated repertoires of moralistic tales (*exempla*) used by Christian preachers. So, too, the famous *Senbād-nāmeh* ("Fables of Sinbad"; one of the sources, incidentally, of Boccaccio's *Decameron*) was rendered from Arabic into Hebrew and thence into Latin; while the renowned romance of *Barlaam and Josaphat*—itself a Christian adaptation of tales about the Buddha—found its Jewish counterpart in a compilation entitled *The Prince and the Dervish,* adapted, from an Arabic text, by Abraham ben Samuel ibn Ḥisdai, a leader of Spanish Jewry in the 13th century.

The *Alexander Romance,* Arthurian legend, and Aesop's fables

Hebrew versions of medieval romances. Here, too, however, the traffic moved in both directions: Hebrew translations were also made from Latin and other European languages. There are, for instance, several Hebrew adaptations of the *Alexander Romance,* based mainly (though not exclusively) on Leo of Naples' Latin rendering of the Greek original by Callisthenes. The central theme is, of course, the exploits of the great Macedonian conqueror, and the narrative is spiced with fanciful accounts of his adventures in foreign lands and of the outlandish peoples he encounters. There is likewise a Hebrew reworking of the Arthurian legend, in the form of a secular sermon in which Arthurian and biblical scenes are blithely mixed together. Finally, there is a Hebrew *Ysopet* (the common title for a medieval version of Aesop) that shares several of its fables with the famous collection made by Marie de France in the late 12th century.

Jewish contributions to Christian and Islāmic tales. Moreover, apart from these Hebrew translations of Oriental and European works, a good deal of earlier haggadic material is embodied in the *Disciplina clericalis* of Peter Alfonsi, a baptized Jew of Aragon originally known as Moses Sephardi. This book, composed in the 12th century, is the oldest European collection of novellas and served as a primary source for the celebrated *Gesta Romanorum* ("Deeds of the Romans") of the same period—a major quarry for European storytellers, poets, and dramatists for many centuries.

Haggadic material percolated also to Arabic writers during this period. Not only does the Qurʾān incorporate such material but also the Egyptian recension of *The Thousand and One Nights* seems to have drawn extensively on Jewish sources, as, for instance, in its tales of "The Sultan and His Three Sons," "The Angel of Death," "Alexander and the Pious Man," and the legend of Baliqiyah.

Major medieval Hebrew collections. Between the 11th and 13th century the tendency developed in Europe to compile, both for entertainment and edification, comprehensive collections of tales and fables; standard examples are the British *Gesta Romanorum,* the Spanish *El novellino,* and the aforementioned *Disciplina clericalis.* Among Jews similar collections were made, especially in Morocco as well as in Moorish Spain. Two of the most important are *The Book of Comfort* by Nissim ben Jacob ben Nissim of al-Qayrawān (11th century) and *The Book of Delight* by Joseph ben Meir ibn Zabara of Spain (end of the 12th century). The former, composed in Judeo-Arabic, is a collection of some 60 moralizing tales designed to comfort the author's father-in-law on the loss of a son. It belongs to a well-known genre of Arabic literature, derives mainly from Arabic sources, and is permeated by a preoccupation with divine justice, typical of the Muʿtazilite school of Islāmic theology. It was later translated into Hebrew. *The Book of Delight* consists of 15 tales, largely about the wiles of women, exchanged between two travelling companions—a form of cadre, or "enclosing tale," adopted on a more extensive scale by Chaucer in his *Canterbury Tales,* which dates from the same period. Typical is the tale of the "Silversmith and His Wife," which relates how a craftsman, persuaded by his greedy wife to make a statue of a princess, gets his hands cut off by the king for violating the Islāmic law against making images, while his wife reaps rich rewards from the flattered princess. Although most of the stories are taken from Arabic sources, some indeed find parallels in rabbinic literature To the latter category belongs, for instance, the famous tale of the matron of Ephesus, who, while keeping vigil over her husband's tomb, at the same time engages in an intrigue with a guard posted nearby to watch over the corpses of certain crucified robbers. When, during one of their trysts, one of the corpses is stolen and her lover therefore faces punishment, the shrewd woman exhumes the body of her husband and substitutes it. This tale is found already in the *Satyricon* of Petronius and was later used by Voltaire in his *Zadig* and by the 20th-century English playwright Christopher Fry in his *A Phoenix Too Frequent.*

The story of the matron of Ephesus

Of the same genre but deriving mainly from west European rather than Arabic sources are the *Mishle shuʿalim* ("Fox Fables") of Berechiah ha-Nakdan (the Punctuator), who may have lived in England toward the end of the 12th century. About half of these tales recur in Marie de France's *Ysopet,* and only one of them is of specifically Jewish origin. Berechiah's work was translated into Latin and thence became a favourite repertoire of European storytellers.

Among anonymous compendiums of this type is *The*

Alphabet of Ben Sira, extant in two recensions, probably of the 11th century. This is basically a collection of proverbs attributed to the famous sage of the apocryphal book Ecclesiasticus (Wisdom of Jesus the Son of Sirach). In one of the recensions they are illustrated by appropriate tales. The author is represented as an infant prodigy who performs much the same feats of sapience as are attributed to Jesus in some of the Infancy Gospels.

The *Josippon, Jashar,* and *Jerahmeel* collections

Medieval historiated Bibles and legendary histories. Two other developments mark the history of Jewish myth and legend during the Middle Ages. The first was a revival of the Hellenistic vogue of compiling large-scale compendiums in which the history of the Jews was "integrated," in legendary fashion, with that of the world in general and especially with classical traditions. Two major works of this kind, both composed (apparently) in Italy during the 9th century, are (1) *Josippon,* composed by a certain Ben Gorion, which presents a fanciful record from the creation onward and contains numerous references to foreign nations; and (2) the *Book of Jashar,* a colourful account from Adam to Joshua, named for the ancient book of heroic songs and sagas mentioned in the Bible (Josh. 10:13; II Sam. 1:18). There is also a voluminous *Chronicles of Jerahmeel,* written in the Rhineland in the 14th century. This draws largely on Pseudo-Philo's earlier compilation, mentioned above, and is of special interest because it includes Hebrew and Aramaic versions of certain books of the Apocrypha.

Medieval Haggadic compendiums. The other development was the gathering of Haggadic legends and tales into comprehensive, systematic compendiums. Works of this kind are (1) *Yalquṭ Shimeʿoni* ("The Collection of Simeon"), attributed to a certain Rabbi Simeon of Frankfurt am Main; (2) *Midrash ha-gadol* ("The Great Midrash"), composed after the death of Moses Maimonides (1204), whom it quotes; and (3) the *Midrash of David ha-Nagid,* grandson of Maimonides. About 100 years later appeared a similar work, *Yalquṭ ha-Makiri* ("The Collection of Makhir"), on the Prophets and Holy Writings, compiled by one Makhir ben Abba Mari in Spain (see above *Torah*). It has been suggested that the compilation of such works was spurred by the necessity of providing "ammunition" for the public disputations with Christian ecclesiastics that the church forced upon Jewish scholars in this period.

Myth and legend in the modern period. *Kabbalistic tales.* In the 16th century, Jewish myth and legend took several new directions. The disappointment of messianic expectations through the dismal eclipse of the pretender Shabbetai Tzevi produced, by way of compensation, an increased interest in occult speculation and in the mystical lore of the Kabbala (esoteric Jewish mysticism). Important schools of Kabbala arose in Italy and at Safed, in Palestine, and tales of the miraculous Faust-like powers of such masters as Isaac Luria and Ḥayyim Vital Calabrese began to circulate freely after their deaths.

The legend of Joseph della Reyna

Another reaction to the dashing of messianic hopes is represented by the beautiful story of the Kabbalist Joseph della Reyna and his five disciples, who go journeying through the world to oust Satan and prepare the way for the Deliverer. Warned by the spirits of such worthies as Rabbi Simeon ben Yoḥai and the prophet Elijah, they nevertheless succeed eventually in procuring their blessing and help and are sent on to the angel Metatron. The latter furnishes them with protective spells and spices and advises Joseph to inscribe the ineffable name of God on a metal plate. When, however, they reach the end of their journey Satan and his wife, Lilith, attack them in the form of huge dogs. When the dogs are subdued they beg for food. Moved to pity, Joseph gives them spices to revive them. At once they summon a host of devils. Two of the disciples die of terror; two go mad, and only Joseph and one disciple are left. The Messiah weeps in heaven, and Elijah hides the great horn of salvation. A voice rings out telling Joseph that it is vain to attempt to hasten the footsteps of the Redeemer.

The repertory of Jewish tales and legends was seasoned, however, by other elements. During the 16th century—the age of the great navigators—stories began to circulate about the discovery of the Ten Lost Tribes in remote parts of the world.

Judeo-German (Yiddish) tales. It was at the same period that Judeo-German (Yiddish) came increasingly to replace Hebrew as the language of Jewish tales and legends in Europe, a major factor in this development being the desire to render them accessible to women unschooled in the sacred tongue. Not only were the synagogal lessons from Scripture legendarily embellished in a so-called *Taitsh Humesh* ("Yiddish Pentateuch"), in the more fancifully titled *Tzeʾena u-reʾena* ("Go Forth and See"; *cf.* S. of Sol. 3:11) by Jacob ben Isaac Ashkenazi, and in adaptations of the story of Esther designed for dramatic presentation on the feast of Purim, but the Hebrew *Chronicles of Josippon* also assumed Yiddish dress. More secular productions were a verse rendition of the Arthurian legend, entitled *Artus Hof* ("The Court of King Arthur"), based largely on Gravenberg's medieval *Wigalois,* and the *Bove Buch* by Elijah Levita, which retold the romance of Sir Bevis of Southampton.

These "frivolous" productions were in time offset by collections of moral and ethical tales. The principal of these are (1) the *Brantspiegel,* attributed to a certain Moses Henoch (Prague 1572), and (2) the *Maʿaseh Buch* ("Story Book"), a compendium of 254 tales compiled by Jacob ben Abraham of Meseritz and first published at Basel in 1602. The latter was drawn mainly from the Talmud but was supplemented by later legends about medieval rabbis. Jewish legends also circulated in the form of ephemeral chapbooks, a large selection of which is preserved in the library of the Yiddish Scientific Institute in New York City.

Judeo-Persian and Judeo-Spanish (*Ladino*) *tales.* A similar development, though on a lesser scale, took place among Jews who spoke other vernacular dialects. Major monuments of Judeo-Persian literature are poetic embellishments of biblical narratives composed by a certain Shāhīn of Shīrāz in the 14th century and by Joseph ben Isaac Yahudi (*i.e.,* the Jew) some 300 years later. These, however, are exercises in virtuosity rather than in creative storytelling. In Judeo-Spanish (Ladino) there are versified elaborations of the story of Joseph, entitled *Coplas de Yoçef* ("Song of Joseph"), composed, in 1732, by Abraham de Toledo and embodying a certain amount of traditional haggadic material. From a revival of literary activity in the 18th century comes a comprehensive "legendary Bible" called Me-ʿam Loʿez ("From a People of Strange Tongue"; *cf.* Ps. 114:1), begun by one Jacob Culi and continued by later writers, as well as several renderings of standard Hebrew collections and a number of Purim plays. Until the Nazi holocaust in the 1940s, Judeo-Spanish folktales were still current in Macedonia and Yugoslavia, but these leaned more on Balkan than on Jewish sources.

Ḥasidic tales. The rise of the Ḥasidic sect (a popular pietistic-mystical movement) in eastern Europe at the end of the 18th century begat a host of legends (circulated mainly through chapbooks) concerning the lives, wise sayings, and miracles of such *tzaddiqim,* or masters, as Israel ben Eliezer, "the Besht" (1700–60), and Dov Baer of Meseritz (died 1772). (See also above *Jewish mysticism.*) These, however, are anecdotes rather than formally structured stories and often borrow from non-Jewish sources.

Droll stories. To the popular creativity of the ghetto belong also the droll tales of the Wise Men of Chełm (in Poland)—Jewish counterparts of the German noodles (stupid people; hence "noodle stories") of Schildburg and of the more familiar English Wise Men of Gotham. These, too, were circulated mainly in Yiddish popular prints. Typical of them is the tale of the two "sages" who went for a walk, one carrying an umbrella and the other without one. Suddenly it began to rain. "Open your umbrella," said the one without one. "It won't help," answered the other, "it's full of holes." "Then why did you bring it?" rejoined his friend. "I didn't think it would rain," was the reply.

Modern Israeli folktales. The gathering of Jews from many lands into the modern state of Israel has made that country a happy hunting ground for the student of Jewish folktales. Assiduous work has been undertaken by Dov Noy of Hebrew University in Jerusalem, aided by enthusiastic amateurs throughout the country. Mainly, however, the stories are retellings of traditional material. (T.H.G.)

Judaism in world perspective

RELATION WITH NON-JUDAIC RELIGIONS

Exclusivist and universalist emphases. The biblical tradition out of which Judaism emerged was predominantly exclusivist ("no other gods"). The gods of the nations were regarded as "no gods" and their worshippers as deluded, while the God of Israel was acclaimed as the sole lord of history, and the Creator of heaven and earth. The unexpected universalist implications of this exclusivism are most forcibly expressed in an oft-quoted verse from Amos (9:7):

God's rule over all the nations

> "Are you not like the Ethiopians to me, O people of Israel?" says the Lord. "Did I not bring up Israel from the land of Egypt, and the Philistines from Caphtor and the Syrians from Kir?"

Here the universal rule of the God of Israel is unmistakably proclaimed. Yet in the same book (3:1–2), after referring to the deliverance from Egypt—an act recognized as similar to that occurring in the affairs of other peoples—the prophet, speaking for God, says: "You only have I known of all the families of the earth." Thus the exclusivism has two focuses, one universal, the other particularistic. The ultimate claim of the universalistic position is found in Malachi 1:11: "For from the rising of the sun to its setting my name is great among the nations." This, however, in no way negates the special covenantal relationship between God and his people; indeed, it is this universalistic theme that underscores that special relation. To interpret Judaism's stance toward other religious systems in any other way is to fail to do justice to its inner dialectic. It is neither a bland latitudinarianism that admits any or all viewpoints and practices, nor a fanatical intolerance but rather a subtle interplay of affirmation and rejection. The latter is directed primarily against the worship of finite things or aspects—idolatry—the basic failure of the peoples who are the objects of the same divine solicitude as is Israel. If the religions of the nations are rejected because of their failure fully and truly to know God, the peoples themselves are not. Living under the covenant with Noah (see above), their fulfillment of such responsibilities provides for their acceptance, for they are not expected to live within the realm of Torah (see also *Relations with other religions* below).

Relation to Christianity. Judaism's relation to Christianity is a complicated one because of the close historical interconnections between them. From a Judaic standpoint, Christianity is or was a Jewish "heresy" and as such may be judged somewhat differently than other religions. Its claims over against Judaism as the true fulfillment of the covenant and, thus, as the true Israel have given rise throughout the centuries to polemics of varying intensity. The rise to power of the church and the embodiment of its anti-Judaic sentiments and attitudes in the political structures and processes of Christian nations made sharply negative Jewish responses inevitable. Nevertheless, during the Middle Ages Jewish thinkers attempted to avoid designating Christianity as idolatry and even to argue that, in a special way, being derived from Judaism, it was fulfilling—at least on the moral plane—the divine purpose.

Christian and Jewish polemic and counter-polemic

In modern times the relation has undergone changes necessitated by the newer situations into which the Jewish community has moved. This does not mean that the polemical-apologetic stance has come entirely to an end. The rejection of Judaism as a living religion by Christians continued and continues, argued not so much on dogmatic as on scholarly grounds. The Jewish response to this has often been countercriticism. Beyond this, however, there has been a growing inclination within the Jewish community to respond to the development of an affirmative theology of Judaism in both the Roman Catholic and Protestant churches by providing a theology of Christianity within Jewish thought. Occasional formulations in this direction have appeared, but it is far too early to know exactly what will emerge. At the same time, it must be noted, there are those who see no need for such a movement, arguing that the failure of the Christian churches in recent years to respond adequately to the tragedies of Jewish existence precludes any real engagement of one with the other.

Relation to Islām. The emergence of Islām in Arabia in the 7th century CE brought Judaism face to face with a second religious movement that derived some of its ideas and structures from the older tradition. In this case, as in that of Christianity, the new religion claimed a special relation with Judaism. Muḥammad held that the faith he proclaimed was none other than the pristine religion of Abraham, the father of Ishmael—progenitor of the Arabs—as well as of Isaac, from whom the people of Israel descended. That religion had been distorted both by Judaism and Christianity; and Muḥammad, the "seal" of the prophets, had been called by God to restore it to its purity. The confrontation between Judaism and Islām, as that with Christianity, was coloured by political and social considerations both before and after Islām moved out of Arabia to build a world empire (including the conquest and settlement of Palestine). During the subsequent period, the intellectual development of the Islāmic world and the emergence of theologians and philosophers of the highest order challenged Judaism and had considerable influence on the rise of similar thinkers within that community. Given the strong monotheism and the anti-iconic attitude of Islām, many of the questions that arose between Judaism and trinitarian and iconic Christianity were not an issue between Judaism and Islām. The crucial point of dispute here was the nature of prophecy, given Muḥammad's claim concerning his culminating role in the prophetic tradition. The medieval period thus saw polemics directed against that claim and, as in the case of the theological work of Moses Maimonides, *More nevukhim* (*The Guide of the Perplexed*), an exposition of the nature of prophecy that, without directly dealing with Muḥammad's claim, may be understood to undercut it. Nonetheless, Islām, too, was understood to contribute to the fulfillment of the divine purpose. From the late medieval period onward, the intellectual engagement between the two religions diminished with the general decline in the Turkish Empire that then embraced the Muslim world. In modern times it has not yet been renewed for many reasons. Once the political problems in the eastern Mediterranean between the State of Israel and the Arab world have been meliorated, the contiguity of the two communities suggests an inevitable renewal of conversations on the religious as on many other levels.

Relations with other religions. Judaism's encounters with religions other than Christianity and Islām have been in large measure limited to the past. In the Hellenistic world, it confronted and rejected the varieties of syncretistic cults that grew up. Within the Sāsānian Empire it was forced to deal with Zoroastrianism, but the outlines of its response have not yet been entirely disentangled from the literature of the period. In the modern world, particularly in the most recent period, it has come face to face with the religions of the Middle and Far East, but beyond a few tentative explorations nothing tangible has appeared. What seems certain is that, considering the growing interest in and exchange between East and West, Jewish thinkers will not be able to rest with older formulations concerning the nature of other religious systems. Without compromising its own faith or falling into an uncritical relativism, Judaism may indeed in the future seek a new way of understanding and relating to the varieties of religious systems facing it on the world scene.

Judaism confronts the religions of mankind

THE ROLE OF JUDAISM IN WESTERN CULTURE AND CIVILIZATION

Its historic role. Given the relationship between Judaism and Christianity—the dominant religious force in the development of Western culture—the role of Judaism in that development was significant. Although the church drew from other sources as well, its retention of the sacred Scriptures of the synagogue (the "Old Testament") as an integral part of its Bible—a decision sharply debated in the 2nd century CE—was crucial. Not only was the development of its ideas and doctrines deeply influenced, but it received as well an ethical dynamism that constantly overcame an inclination to withdraw into world-denying isolation. It was, however, not only Judaism's heritage but its persistence that touched Western civilization. The continuing existence of the Jews, even as a pariah people, was both a challenge and a warning; and ultimately, at the beginning of the modern era, their liberation from the shackles of discrimination, segregation, and rejection was understood by many to be the touchstone of all human liberty. Until the final ghettoization of the Jew—it is well to remember that the term "ghetto" belongs in the first instance to Jewish history—at the end of the Middle Ages and the beginning of the Renaissance, intellectual contact between Judaism and Christianity, and thus with Western culture, did not cease. Jerome translated the Hebrew Bible into Latin with the aid of Jewish scholars; Luther, into German with the aid of commentaries beholden to Jewish authors. Jewish thinkers mediated the remarkable intellectual achievements of the Islāmic world to Christian Europe and added their own contributions as well. Even heresies within the church found, on occasion, their inspiration or prototype in Judaism.

Its present role. In the modern world, while the influence of Jews has increased in almost every realm of cultural life, the impact of Judaism has diminished. The reason for this is not difficult to find. The Gentile leaders who extended emancipation to the Jews at the end of the 18th and beginning of the 19th centuries, while eager to grant political equality to the individual Jew, did so with the implicit and explicit requirement that conformity through reforms of Judaism be agreed to. With the transformation of Judaism into an ecclesiastical institution, largely on the model of German Protestant churches, its ideas and structures took on the cast of its environment in a way quite unlike what had ensued in its earlier confrontations with various philosophical systems. Indeed, for some, Judaism and 19th-century European thought were held to be not merely congruent but identical. Thus, while numerous contributors to diverse aspects of Western culture and civilization are to be found among Jews of the 20th century—scientists, politicians, statesmen, scholars, musicians, artists—their activities cannot, except in specific instances, be considered as deriving from Judaism as it has been sketched above.

The two key events: Auschwitz and the establishment of Israel

Future prospects. Two events of the 20th century have, however, confronted Judaism in such ways as to suggest that its wrestling with them and their profound challenge to it may presage a new role and a new influence for Judaism: "Auschwitz" and the establishment of the State of Israel. The premeditated murder of some 6,000,000 European Jews by the Nazis for no other reason than that they were Jews, has shaken Jewish thinkers to their very core. Indeed, so traumatic was this event, that for almost two decades following it, no substantial attempt was made to plumb its meaning. At the same time, the reappearance of the State of Israel, viewed for the most part from outside the Jewish community as nothing more than a political event, has set in motion an entirely different chain of theological inquiry. These two happenings have clearly, but in as yet unpredictable ways, begun to work and to move within the thought of contemporary Judaism. Out of this working and moving there may emerge an inescapable spiritual impact upon Western culture and civilization, which have, as yet, resolutely refused to face the realities these fateful occurrences represent. If contemporary Judaism is able to say what they mean, however haltingly, it will have renewed its potent relationship to the Western world, and, given the nature of contemporary society, established a similar bond with the Eastern world as well. (L.H.S.)

BIBLIOGRAPHY

General history: SALO W. BARON, *A Social and Religious History of the Jews,* 2nd ed., 15 vol. (1952–73), a comprehensive presentation of the intertwined social and religious history with copious bibliographical information critically evaluated; LOUIS FINKELSTEIN (ed.), *The Jews: Their History, Culture and Religion,* 4th ed., 3 vol. (1970–71), critical essays by outstanding authorities on the major aspects of Jewish history and culture; JULIUS GUTTMANN, *Die Philosophie des Judentums* (1933; Eng. trans., *Philosophies of Judaism,* 1964), the best single volume on the history of Jewish thought from ancient times to the present, especially valuable for medieval Jewish philosophy; LEO W. SCHWARZ (ed.), *Great Ages and Ideas of the Jewish People* (1956), interpretive and highly readable essays by six historians on Jewish history, with emphasis on intellectual history, intended primarily for the layman; GERSHOM G. SCHOLEM, *Major Trends in Jewish Mysticism,* 3rd rev. ed. (1954), the classic treatment of mystical doctrines and schools in Judaism; MAX L. MARGOLIS and ALEXANDER MARX, *A History of the Jewish People* (1927, reprinted 1958), an excellent, readable, introductory survey. See also ROBERT M. SELTZER, *Jewish People, Jewish Thought: The Jewish Experience in History* (1980).

Biblical Judaism: (*General reference*): *The Interpreter's Dictionary of the Bible,* 4 vol. (1962). (*Surveys of the culture and religion of ancient Israel*): JOHANNES PEDERSEN, *Israel: Its Life and Culture,* 4 vol. in 2 (1926–40, reprinted 1959); W.F. ALBRIGHT, *From the Stone Age to Christianity,* 2nd ed. (1957); YEHEZKEL KAUFMANN, *The Religion of Israel, from Its Beginnings to the Babylonian Exile* (1960) and *The Babylonian Captivity and Deutero-Isaiah* (1970); ROLAND DE VAUX, *Les Institutions de l'Ancien Testament,* 2 vol. (1958–60; Eng. trans., *Ancient Israel: Its Life and Institutions,* 1961); GERHARD VON RAD, *Theologie des Alten Testaments,* 2nd ed. (1958; Eng. trans., *Old Testament Theology,* 2 vol., 1962–65); HELMER RINGGREN, *Israelitische Religion* (1963; Eng. trans., 1966). (*Annual bibliographic keys*): *Elenchus Bibliographicus Biblicus,* ed. by P. NOBER (1920–); *Book List* of the British Society for Old Testament Study. (*Special topics*): E.A. SPEISER, "The Biblical Idea of History in Its Common Near Eastern Setting," *Oriental and Biblical Studies: Collected Writings of E.A. Speiser,* ed. by J.J. FINKELSTEIN and MOSHE GREENBERG, pp. 187–210 (1967); MENAHEM HARAN, "The Religion of the Patriarchs: Beliefs and Practices," *Patriarchs: The World History of the Jewish People,* vol. 2, pp. 219–245 (1970); MOSHE GREENBERG, "Crimes and Punishments," *Interpreter's Dictionary of the Bible,* vol. 1, pp. 733–744 (1962); "Some Postulates of Biblical Criminal Law," *The Jewish Expression,* ed. by JUDAH GOLDIN, pp. 18–37 (1970); AELRED CODY, *A History of Old Testament Priesthood* (1969); JOHANNES LINDBLOM, *Prophecy in Ancient Israel* (1962); E.J. BICKERMAN, "The Historical Foundations of Postbiblical Judaism," *The Jews: Their History, Culture, and Religion,* ed. by LOUIS FINKELSTEIN, 4th ed., vol. 1, pp. 70–114 (1960); DAN JACOBSON, *The Story of the Stories: The Chosen People and Its God* (1982).

Hellenistic Judaism: (*Bibliographies*): GERHARD DELLING (ed.), *Bibliographie zur jüdisch-hellenistischen und intertestamentarischen Literatur 1900–1970,* 2nd ed. (1975), extremely comprehensive bibliography on religion and literature of Diaspora Judaism, arranged according to topics, with separate bibliographies on every major Hellenistic Jewish author and on each book of the Apocrypha and Pseudepigrapha; RALPH MARCUS, "A Selected Bibliography (1920–1945) of the Jews in the Hellenistic-Roman Period," *Proceedings of the American Academy for Jewish Research,* 16:97–181 (1946–47), covers both Palestine and the Diaspora, helpful for noting works that are useful introductions and works that are indispensable to the specialist; LOUIS H. FELDMAN, *Scholarship on Philo and Josephus, 1937–1962* (1963), critical bibliography, arranged topically, with comments on both Diaspora and Palestinian Judaism generally. (*Papyrological and archaeological sourcebooks*): VICTOR A. TCHERIKOVER, ALEXANDER FUKS, and MENAHEM STERN (eds.), *Corpus Papyrorum Judaicarum,* 3 vol. (1957–64), contains text, translation, bibliography, and commentary on all papyri and inscriptions pertaining to Jews from 323 BCE to 641 CE—thoroughly reliable; ERWIN R. GOODENOUGH, *Jewish Symbols in the Greco-Roman Period,* 13 vol. (1953–68), a magnificent, exhaustive collection of the archaeological findings, with highly insightful, if controversial, commentary. (*Standard scholarly treatments of Hellenistic Judaism*): VICTOR A. TCHERIKOVER, *Hellenistic Civilization and the Jews* (1959; orig. pub. in Hebrew, 1930), extremely meticulous and generally balanced (though with an anti-theological bias), particularly in dealing with the political and social factors in both Palestinian and Diaspora Jewry; ROBERT H. PFEIFFER, *History of New Testament Times, with an Introduction to the Apocrypha* (1949), a sane and useful, if unoriginal, survey of the political, religious, and literary history of Palestinian and especially Diaspora Judaism, from 200 BCE to 200 CE; MOSES HADAS, *Hellenistic Culture: Fusion and Diffusion* (1959), highly suggestive, though often extravagant, treatment of the interaction of Hellenism and other cultures, especially Judaism. (*Introductory popular treatments of Hellenistic Judaism*): RALPH MARCUS, "The Hellenistic Age," in LEO W. SCHWARZ (ed.), *Great Ages and Ideas of the Jewish People,* pp. 93–139 (1956), extremely judicious and readable account by an eminent authority; for a readable guide to the literature, together with representative samples, see his "Hellenistic Jewish Literature," in LOUIS FINKELSTEIN (ed.), *The Jews: Their History, Culture and Religion,* 3rd. ed., vol. 2, pp. 1077–1115 (1960). (*Works on Palestinian Judaism*): GEORGE FOOT

MOORE, *Judaism in the First Centuries of the Christian Era,* 3 vol. (1927–30, reprinted 1966–67), classic treatment based primarily on the Talmudic corpus, though the view of a Pharisaic "normative" Judaism has since been strongly challenged; SOLOMON ZEITLIN, *The Rise and Fall of the Judaean State: A Political, Social and Religious History of the Second Commonwealth,* 2 vol. (1962–67), stimulating and often highly original survey of the period from 332 BCE to 66 CE, though the scholarly dogmatism is occasionally jarring; SAUL LIEBERMAN, *Greek in Jewish Palestine,* 2nd ed. (1965) and *Hellenism in Jewish Palestine,* 2nd ed. (1962), highly significant, ingenious, and learned illustrations of the influence of Greek culture on the language and exegetical format of the Palestinian rabbis. (*Works on Diaspora Judaism*): HARRY A. WOLFSON, *Philo: Foundations of Religious Philosophy in Judaism, Christianity, and Islam,* rev. ed., 2 vol. (1962), a great and seminal, though controversial, work that makes Alexandrian Judaism a collateral branch of Palestinian Pharisaic Judaism; LOUIS H. FELDMAN, "The Orthodoxy of the Jews in Hellenistic Egypt," *Jewish Social Studies,* 22:215–237 (1960), a survey using literature, papyri, and art objects to examine the synthesis of Greek culture and Judaism in the upper and lower classes, respectively, of Hellenistic Egypt. See also JOHN J. COLLINS, *Between Athens and Jerusalem: Jewish Identity in the Hellenistic Diaspora* (1983); and MENAHEM MOR and URIEL RAPPAPORT, *Bibliography of Works on Jewish History in the Hellenistic and Roman Periods* (1982).

Rabbinic Judaism: (*Palestinian Judaism*): SOLOMON SCHECHTER, *Aspects of Rabbinic Theology* (1961), a concise, authoritative, and engaging treatment of classical (*i.e.,* rabbinic) Judaism. (*Babylonian Judaism*): JACOB NEUSNER, *A History of the Jews in Babylonia,* 5 vol. (1965–70; vol. 1, rev. 1969), the most comprehensive treatment of Babylonian Jewry during the Tannaitic and Amoraic periods. (*Judeo-Arabic culture*): S.D. GOITEIN, *Jews and Arabs: Their Contacts Through the Ages,* 3rd rev. ed. (1974), a popular work by the ranking authority on all aspects of Jewish–Arabic symbiosis, particularly valuable for the medieval period; ABRAHAM IBN DAUD, *Sefer ha-Qabbalah* (*The Book of Tradition*), ed. and trans. by GERSON D. COHEN (1967), the classic medieval Hebrew chronicle with analytic essays on Spanish Jewry's "golden age." (*Jews of medieval Europe*): ISRAEL ABRAHAMS, *Jewish Life in the Middle Ages,* new ed. rev. by CECIL ROTH (1932), delightful and erudite studies of social life and institutions in medieval Europe; MORITZ GÜDEMANN, *Geschichte des Erziehungswesens und der Cultur der Juden,* 3 vol. (1880–88), a social and intellectual history of medieval Ashkenazic Jewry (France, Germany, Italy); CECIL ROTH, *The Jews in the Renaissance* (1959), lucid and informative, but with little critical analysis, valuable on Jewish contact with Christian men of letters; GERSHOM SCHOLEM, *Shabbethai Zevi,* 2 vol. (1957), a penetrating and comprehensive study (in Hebrew) of the great false Messiah as well as of his religious antecedents and legacy; STEPHEN SHAROT, *Messianism, Mysticism, and Magic: A Sociological Analysis of Jewish Religious Movements* (1982), an informative scholarly study.

Contemporary Judaism: The most convenient summary for the study of modern Jewish history is HOWARD MORLEY SACHAR, *The Course of Modern Jewish History,* updated and expanded ed. (1977). Modern Jewish thought and movements are covered in a useful manual by JOSEPH BLAU, *Modern Varieties of Judaism* (1966). NATHAN ROTENSTREICH, *Jewish Philosophy in Modern Times* (1968); and the last (modern) section of JULIUS GUTTMAN, *Philosophies of Judaism* (1964), are more advanced. The very best book on Jewish religion in the modern age is unfortunately still untranslated: MAX WIENER, *Jüdische Religion im Zeitalter der Emanzipation* (1933). For Zionism, see the only attempt at a comprehensive reader in English, ARTHUR HERTZBERG (ed.), *The Zionist Idea* (1959). There are two excellent expositions of Judaism from a Reform–Liberal point of view: LEON ROTH, *Judaism: A Portrait* (1960); and LEO BAECK, *The Essence of Judaism* (1961). Conservative Judaism is well described in a book about its early history, MOSHE DAVIS, *The Emergence of Conservative Judaism* (1963); and by JACOB AGUS, *Dialogue and Tradition* (1969), the essays of a distinguished Conservative thinker. Reconstructionism is best understood in the words of its founder, MORDECAI M. KAPLAN, *Judaism As a Civilization,* 2nd ed. (1957). The standard modern single volume about Orthodox Judaism is ISIDORE EPSTEIN, *Judaism* (1935). Neo-Ḥasidism was best described by its greatest exponent, MARTIN BUBER, *The Origin and Meaning of Hasidism* (1960). Later studies include EUGENE B. BOROWITZ, *Choices in Modern Jewish Thought: A Partisan Guide* (1983); PAUL R. MENDES-FLOHR and JEHUDA REINHARZ (eds.), *The Jew in the Modern World* (1980); DAN ROSS, *Acts of Faith: A Journey to the Fringes of Jewish Identity* (1982).

Judaic literature: B. GERHARDSSON, *Memory and Manuscript* (1961), contains a description of the methods and techniques by which the oral tradition was transmitted. H.L. STRACK, *Introduction to the Talmud and Midrash* (1931); and M. MIELZINER, *Introduction to the Talmud,* 4th ed. (1968), are still the best introductions for the general reader. The latter is particularly helpful in explaining Talmudic dialectic terminology and debate. J. BOWKER, *The Targums and Rabbinic Literature* (1969); and E. DEUTSCH, *The Talmud* (1895), are both descriptions of the Talmud, the former concentrating upon Talmudic literary compilations and the latter upon Talmudic content. The introduction of J. GOLDIN, *The Living Talmud* (1957), contains a vivid description of Talmudic debate. C. ALBECK, *Introduction to the Talmud,* in Hebrew (1969); and Z.H. CHAJES, *The Student's Guide Through the Talmud* (Eng. trans. 1952), are more advanced introductions, the former analytical and scientific and the latter representing the traditional view. J. NEUSNER (ed.), *The Formation of the Babylonian Talmud* (1970), contains summaries of some research by modern scholars on the question of how the Talmud was formed. L. GINZBERG, "Introduction to the Talmud," in *A Commentary on the Palestinian Talmud,* vol. 1 (1941), is the only introduction to the Palestinian Talmud available in the English language. L. FINKELSTEIN, *Akiba* (1962), is a historical and sociological approach to the development of Halakha. See SAUL LIEBERMAN, *op. cit.,* for the interrelationship between the ancient rabbinic world and its Gentile environment. B. COHEN, *Jewish and Roman Law,* 2 vol. (1966); and I. HERZOG, *The Main Institutions of Jewish Law,* 2 vol. (1966–67), are the best English descriptions of Jewish law. J. Z. LAUTERBACH, *Rabbinic Essays* (1951); and E.E. URBACH, *The Sages: Their Concepts and Beliefs,* in Hebrew (1969), cover the major aspects of rabbinic theology. I. HEINEMANN, *Paths of the Aggadah,* in Hebrew (1970); M. KADUSHIN, *The Rabbinic Mind,* 3rd ed. (1972); and D. BEN AMOS, *Narrative Forms of the Haggadah: Structural Analysis* (1969), discuss Haggadic methods, forms, concepts, and thinking. L. ZUNZ, *Die Gottesdienstlichen Vorträge der Juden* (1892; updated Hebrew translation, 1950), is a thorough historical survey of Haggadic literature. D. NOY, *Motif-Index of the Talmudic-Midrashic literature* (1954); J.J. SLOTKI, *Index Volume to the Soncino Talmud* (1952); M. GASTER, *The Exempla of the Rabbis* (1924, rev. ed. 1968); and C.G. MONTEFIORE and H. LOEWE, *A Rabbinic Anthology* (1938), are very helpful as reference guides; while W.O.E. OESTERLEY, H. LOEWE, and E.I.J. ROSENTHAL, *Judaism and Christianity,* rev. ed. (1969); and C. MERCHAVIA, *The Church Versus Talmudic and Midrashic Literature,* in Hebrew (1970), describe the relationship between the church and rabbinic Judaism. E.R. BEVAN and C. SINGER (eds.), *The Legacy of Israel* (1927), deals with the influence of Judaism on world culture. Extensive bibliographies may be found in the works of Gerhardsson, Mielziner, Bowker, and Ben Amos. JACOB NEUSNER, *Judaism: The Evidence of the Mishnah* (1981), introduces new methods of textual criticism.

Basic beliefs and doctrines: KAUFMANN KOHLER, *Jewish Theology: Systematically and Historically Considered* (1918, reprinted 1928); YEHEZKEL KAUFMANN, *The Religion of Israel from Its Beginnings to the Babylonian Exile* (1960; orig. pub. in Hebrew, 1937–56), an abridgement and translation of the work of one of the most influential Jewish biblical scholars of modern times; GEORGE FOOT MOORE, *Judaism in the First Centuries of the Christian Era,* 3 vol. (1927–30), a masterful work by a distinguished Christian student of Judaism; SOLOMON SCHECHTER, *Some Aspects of Rabbinic Theology* (1909, reprinted 1936 and 1961), an insightful presentation of the basic doctrines; CLAUDE G. MONTEFIORE and HERBERT LOEWE, *A Rabbinic Anthology* (1960), a collection of materials from rabbinic sources, arranged under theologic headings, with ample notes and discussions; JULIUS GUTTMANN, *Philosophies of Judaism: The History of Jewish Philosophy from Biblical Times to Franz Rosenzweig* (1964), a philosophy of Judaism in the form of a history of philosophy in Judaism; ARTHUR HERTZBERG (ed.), *Judaism* (1961); JACOB NEUSNER, *The Way of Torah: An Introduction to Judaism,* 3rd ed. (1979), a very helpful statement using a history-of-religions approach; LEO BAECK, *Dieses Volk; jüdische Existenz,* 2 vol. (1955–57; Eng. trans., *This People Israel: The Meaning of Jewish Existence,* 1964), a masterful interpretation of Jewish affirmations set within an historical context; ABRAHAM E. MILLGRAM (ed.), *Great Jewish Ideas* (1964), a collection of essays by various scholars setting forth classic positions, covering a wide range of theological ideas; JACOB B. AGUS, *The Jewish Quest* (1983), a collection of essays on basic concepts of Jewish theology.

Ethics and society: SIMON BERNFELD (comp.), *The Foundations of Jewish Ethics,* 2nd rev. ed. (1968), source materials covering Jewish history, with brief introductions by several notable scholars; MORITZ LAZARUS, *Die Ethik des Judenthums,* 2 vol. (1898–1911; Eng. trans., *The Ethics of Judaism,* 1900), a classic presentation from a 19th-century perspective; SAMUEL S. COHON, *Judaism: A Way of Life* (1948), written from the Reform point of view but deeply sympathetic to a wide range of ideas. Jewish ethics is also explored in ANNE ROIPHE, *Generation Without Memory: A Jewish Journey in Christian Amer-*

ica (1981); and illustrated with literary examples in FRANCINE KLAGSBRUN (comp.), *Voices of Wisdom: Jewish Ideals and Ethics for Everyday Living* (1980).

Basic practices and institutions: LEWIS N. DEMBITZ, *Jewish Services in Synagogue and Home* (1898); HAYYIM SCHAUSS, *The Jewish Festivals* (1938; orig. pub. in Hebrew, 1933), and *The Lifetime of a Jew Throughout the Ages of Jewish History* (1950); ABRAHAM Z. IDELSOHN, *Jewish Liturgy and Its Development* (1932, reprinted 1967); MARK L. RAPHAEL (ed.), *Jews and Judaism in the United States: A Documentary History* (1983); SHALOM LILKER, *Kibbutz Judaism: A New Tradition in the Making* (1982).

Art and iconography: FRANZ LANDSBERGER, *A History of Jewish Art* (1946); ABRAHAM Z. IDELSOHN, *Jewish Music in Its Historical Development* (1929, reprinted 1967); CECIL ROTH (ed.), *Jewish Art: An Illustrated History* (1961).

Relation with non-Judaic religions: LEO BAECK, *Judaism and Christianity* (1958); SAMUEL SANDMEL, *We Jews and You Christians: An Inquiry into Attitudes* (1967); JACOB KATZ, *From Prejudice to Destruction: Anti-Semitism* (1980); FRANCIS E. PETERS, *Children of Abraham: Judaism, Christianity, Islam* (1982).

The role of Judaism in Western culture and civilization: CECIL ROTH, *The Jewish Contribution to Civilization* 3rd ed. (1956); DENNIS B. KLEIN, *Jewish Origins of the Psychoanalytic Movement* (1981).

The present-day forms of Judaism: LEON D. STITSKIN (ed.), *Studies in Torah Judaism* (1969); ALFRED JOSPE (ed.), *Tradition and Contemporary Experience: Essays on Jewish Thought and Life* (1970); ARNOLD JACOB WOLF (ed.), *Rediscovering Judaism: Reflections on a New Theology* (1965); *The Condition of Jewish Belief: A Symposium, Compiled by the Editors of Commentary Magazine* (1966); MORDECAI M. KAPLAN, *Judaism as a Civilization,* 2nd ed. (1957); SOLOMON POLL, *The Hasidic Community of Williamsburg: A Study in the Sociology of Religion* (1969); BERNARD MARTIN (ed.), *Contemporary Reform Jewish Thought* (1968); CHARLES LIEBMAN, "Orthodoxy in American Jewish Life," *American Jewish Yearbook* (1965); MORDECAI WAXMAN (ed.), *Tradition and Change: The Development of Conservative Judaism* (1958); JOSHUA ROTHENBERG, *The Jewish Religion in the Soviet Union* (1972).

The religious year: ROLAND DE VAUX, *op. cit.,* summarizes the contemporary state of biblical scholarship regarding the origin and development of the Jewish calendar, sabbath, and festivals. Other relevant works include: THEODOR H. GASTER, *Festivals of the Jewish Year* (1953), an anthropological, comparative, and often speculative approach to the sabbath and festivals; SHLOMO YOSEF ZEVIN, *ha-Moʿadim ba-halakhah* (1944), a modern classic (in Hebrew) treating talmudic and post-talmudic developments in the festival observances; and MENAHEM M. KASHER, *Torah Shelemah,* vol. 13 (1949), a comprehensive history of the Jewish calendar, also in Hebrew.

General introductions to Jewish philosophy: JULIUS GUTTMANN, *Die Philosophie des Judentums* (1933; Eng. trans., *Philosophies of Judaism,* 1964), the best general treatment of Jewish philosophy from the ancient to the modern period, ending with Franz Rosenzweig; NATHAN ROTENSTREICH, *Jewish Philosophy in Modern Times: From Mendelssohn to Rosenzweig* (1968), a good philosophical discussion of the major thinkers in modern Jewish philosophy, especially perceptive on Hermann Cohen. ALEXANDER ALTMANN, *Essays in Jewish Intellectual History* (1981), a collection of insightful scholarly essays on many aspects of Jewish philosophy.

Hellenistic philosophy: NAHUM GLATZER (ed.), *The Essential Philo* (1971), lengthy selections from the major works of Philo, with notes; HARRY A. WOLFSON, *Philo,* rev. ed., 2 vol. (1962), the most comprehensive study of Philo in any language, with special emphasis upon Philo's influence upon later philosophy.

Medieval philosophy: ISAAC HUSIK, *A History of Medieval Jewish Philosophy* (1940, reprinted 1969), a thorough examination of each of the major medieval Jewish philosophers from Isaac Israeli through Joseph Albo, with good bibliography; *Three Jewish Philosophers: Philo, Saadya Gaon, Jehuda Halevi,* trans. and ed. by HANS LEWY, ALEXANDER ALTMANN, and ISAAK HEINEMANN (1965), a good introductory anthology containing representative selections from these classical figures with perceptive introductions and explanatory notes; GEORGES VAJDA, *Introduction à la pensée juive du moyen âge* (1947), a good general survey organized around the major philosophical traditions in medieval Jewish philosophy, with extensive bibliography.

Jewish Kalām: SAʿADIA BEN JOSEPH, *The Book of Beliefs and Opinions,* trans. by SAMUEL ROSENBLATT (1948), Saʿadia's important philosophical work dealing with the major topics in Jewish theology such as faith and reason, creation, God, and reward and punishment; *A Karaite Anthology,* trans. and ed. by LEON NEMOY (1952), an excellent collection of the more accessible Karaite materials, covering a wide range of themes, writers, and periods; ISRAEL EFROS, "Medieval Jewish Philosophy" (1967), a collection of essays (in Hebrew) containing a full-length study of Saʿadia's philosophy, in addition to short essays on Judah ha-Levi and Maimonides.

Jewish Neo-Platonism: ISAAC ISRAELI, *Works,* ed. and trans. by ALEXANDER ALTMANN and S.M. STERN (1958), contains complete translations of three of Israeli's works and excerpts from another, together with a comprehensive essay on his philosophy; SOLOMON IBN GABIROL, "Fountain of Life," trans. from the Latin into Hebrew by J. BLUWSTEIN (1950), in addition to a complete vocalized translation of the text this edition contains the Hebrew summary made by Shem Tov ibn Falaquera in the 13th century.

Judah ha-Levi: The Kuzari, trans. by HARTWIG HIRSCHFELD (1964), a complete English translation of the text and notes, with an introductory essay by HENRY SLONIMSKY; LEO STRAUSS, "The Law of Reason in the Kuzari," in *Persecution and the Art of Writing* (1952), in addition to this perceptive and provocative essay on ha-Levi's attitude towards philosophy and rationalistic ethics, the book contains important essays on Maimonides and Spinoza; BAHYA IBN PAQUDA, *The Duties of the Heart* (Hebrew ed. by A. ZIFRONI, 1928; Eng. trans. by EDWIN COLLINS, 1904), one of the more widely read medieval classics of Jewish philosophy, concentrating upon ethics and personal piety; ABRAHAM BAR HIYYA, *The Meditation of the Sad Soul,* trans. with introduction by G. WIGODER (1971), one of the more accessible philosophical works of this important medieval astronomer and mathematician, devoted primarily to ethical and religious themes.

Jewish Aristotelianism: ABRAHAM IBN DAUD, *Das Buch Emunah Ramah,* ed. and trans. into German by S. WEIL (1852), one of the first manifestations of the Jewish assimilation of Aristotelian philosophy, containing a critique of Gabirol's neo-Platonism.

Maimonides: The Guide of the Perplexed, trans. with introductory essay by SHLOMO PINES (1963), an accurate and clear translation, with an excellent essay by Pines and a valuable prefatory essay by LEO STRAUSS; *A Maimonides Reader,* ed. with introduction and notes by I. TWERSKY (1972), a fine anthology containing important material from Maimonides' *Mishne Torah* and other legal writings, as well as from his shorter philosophical-theological essays and the *Guide;* HARRY A. WOLFSON, "Maimonides on Negative Attributes," in the *Louis Ginzberg Jubilee Volume,* pp. 411–446 (1945), the best historical and analytical study of Maimonides' doctrine of divine attributes; and "Halevi and Maimonides on Prophecy," *Jewish Quarterly Review,* n.s., 32:345–370, 33:49–82 (1942), an excellent historical analysis of the sources of doctrines of both philosophers on prophecy and a clear analysis of the differences between them. ISADORE TWERSKY, *Studies in Jewish Law and Philosophy* (1982), is a collection of previously published articles, some of them in Hebrew.

Averroists: (*Isaac Albalag*): GEORGES VAJDA, *Isaac Albalag, Averroïste juif, traducteur et annotateur d'Al-Ghazâlî* (1960), a translation of and commentary on Albalag's main philosophical work, al-Ghazālī's "Inconsistencies of the Philosophers," containing valuable citations from the classical Arabic philosophical texts. (*Levi ben Gerson*): "The Wars of the Lord" (1866), Gersonides' major philosophical work (in Hebrew) dealing with the most difficult and controversial topics in medieval philosophy and science; SEYMOUR FELDMAN, "Gersonides' Proofs for the Creation of the Universe," *Proceedings of the American Academy for Jewish Research,* 35:113–137 (1967), an analytical study of Gersonides' theory of creation and of his criticism of Aristotle's theory of eternity of the universe. (*Hasdai Crescas*): "The Light of the Lord" (1861), Crescas' most original critique (in Hebrew) of Aristotelian philosophy and his vigorous defense of traditional Judaism; HARRY A. WOLFSON, *Crescas' Critique of Aristotle* (1929), the most important study of medieval Jewish and Arabic philosophy so far written, containing a translation and critical Hebrew text of part 1 of the treatise with comprehensive, detailed, and most valuable notes and introductory essay; JOSEPH ALBO, *Book of Principles,* trans. with critical text by ISAAC HUSIK, 4 vol. (1946), a fine translation with helpful notes of Albo's treatise in Jewish dogmatics.

Modern Jewish philosophy: (*Iberian-Dutch philosophers*): I.S. RÉVAH, *Spinoza et le dr. Juan de Prado* (1959), a study of the cultural background of Spinoza's Amsterdam, especially of the heterodox elements in Sefardic Judaism, containing valuable material pertaining to the excommunication of Spinoza and the ideas of Uriel da Costa; BARUCH SPINOZA, *Theologico-Political Treatise,* trans. by R.H.M. ELWES (1883, reprinted 1951), Spinoza's critique of the Bible and the Jewish religion; LEO STRAUSS, *Spinoza's Critique of Religion* (1965), an excellent philosophical study of Spinoza's *Treatise,* its relation to Maimonides, Uriel da Costa, and other Sefardic heterodox thinkers; HARRY A. WOLFSON, *The*

Philosophy of Spinoza, 2 vol. (1934, reprinted 1969), a most detailed commentary on Spinoza's *Ethics,* containing valuable references to Spinoza's medieval sources such as Maimonides, Gersonides, and Crescas.

German philosophers: (*Moses Mendelssohn*): *Jerusalem and Other Jewish Writings,* trans. and ed. by ALFRED JOSPE (1969), a complete translation of *Jerusalem* and other miscellaneous writings, pertaining to questions on the Jewish religion, with a brief, informative, acute introduction by the editor; JULIUS GUTTMANN, "Mendelssohn's Jerusalem and Spinoza's *Theologico-Politico Treatise*" (in Hebrew) in his *Religion and Knowledge* (1956), a collection of Guttmann's essays that also includes important essays on ha-Levi, Maimonides, Gersonides, and Crescas, as well as on themes in modern philosophy of Judaism; NACHMAN KROCHMAL, "Collected Works" (1961, in Hebrew), containing his "Guide for the Perplexed of Our Times," ed. by S. RAWIDOWICZ and Rawidowicz' comprehensive analytical study of Krochmal's philosophy of Jewish history and its relationship to Hegelian philosophy. (*Hermann Cohen*): *Die Religion der Vernunft aus den Quellen des Judentums* (1919, reprinted 1966), Cohen's major work dealing with his philosophy of Judaism, organized according to the main themes of the Jewish religion; *Reason and Hope: Selections from the Jewish Writings of Hermann Cohen,* trans. and ed. by EVA JOSPE (1971), a useful collection from Cohen's miscellaneous writings on Jewish themes. (*Franz Rosenzweig*): *The Star of Redemption,* trans. by WILLIAM HALLO (1971), Rosenzweig's early but impressive statement of his "re-conversion" to Judaism, which has been influential in contemporary Jewish and non-Jewish theology; NAHUM GLATZER, *Franz Rosenzweig: His Life and Thought,* 2nd ed. (1961), an excellent anthology of Rosenzweig's various writings, containing many letters that reveal the more important and intimate episodes in Rosenzweig's career. (*Martin Buber*): See bibliography to article BUBER, MARTIN in the *Micropædia.* Interpretive studies include EMIL L. FACKENHEIM, *To Mend the World: Foundations of Future Jewish Thought* (1982); ARTHUR A. COHEN, *The Tremendum: A Theological Interpretation of the Holocaust* (1981); ALFRED JOSPE (ed.), *Studies in Jewish Thought: An Anthology of German Jewish Scholarship* (1981).

Jewish mysticism: G.G. SCHOLEM, *Major Trends in Jewish Mysticism,* rev. ed. (1961), the standard survey of the subject, with a chapter-by-chapter bibliography, *On the Kabbalah and its Symbolism* (Eng. trans. 1965), several studies on some of the great themes of Jewish mysticism; A.E. WAITE, *The Holy Kabbalah: A Study of Secret Tradition in Israel* (1929), a theosophical view of Jewish mysticism. HUGO ODEBERG, *3 Enoch; or the Hebrew Book of Enoch* (1928); ISIDOR KALISCH, *A Book on Creation* (1877); J. ABELSON, *The Immanence of God in Rabbinical Literature* (1912); G.G. SCHOLEM, *Jewish Gnosticism, Merkabah Mysticism, and Talmudic Tradition* (1960); JOSHUA TRACHTENBERG, *Jewish Magic and Superstition* (1939, paperback 1961). G.G. SCHOLEM, *Ursprung and Anfänge der Kabbala* (1962; French trans., *Les Origines de la Kabbale,* 1966); *Le commentaire d'Ezra de Gérone sur le Cantique des Cantiques,* trans. by G. VAJDA (1969); *The Zohar,* trans. by H. SPERLING and M. SIMON, 5 vol. (1931–34; paperback, sel. and ed. by G.G. SCHOLEM, 1963). MOSES CORDOVERO, *The Palm Tree of Deborah,* trans. from the Hebrew by L. JACOBS, 3rd ed. (1981); RAPHAEL J. ZWI WERBLOWSKY, *Joseph Karo: Lawyer and Mystic* (1962). L.I. NEWMAN, *The Hasidic Anthology* (Eng. trans. 1963); DOB BAER OF LUBAVITCH, *Tract on Ecstasy,* trans. from the Hebrew by L. JACOBS (1963). MARTIN BUBER, *The Origin and Meaning of Hasidism* (Eng. trans. 1960); S.H. DRESNER, *The Zaddik: The Doctrine of the Zaddik According to the Writings of Rabbi Yaakov Yosef of Polnoy* (1960); S. POLL, *The Hasidic Community of Williamsburg* (1962). J.L. BLAU, *The Christian Interpretation of the Cabala in the Renaissance* (1944); F. SECRET, *Les Kabbalistes chrétiens de la Renaissance* (1964); BEN ZION BOKSER, *The Jewish Mystical Tradition* (1981).

Jewish myth and legend: T.H. GASTER, *Myth, Legend, and Custom in the Old Testament* (1969); R.H. CHARLES (ed.), *The Apocrypha and Pseudepigrapha of the Old Testament in English,* 2 vol. (1913, reprinted 1964); M.R. JAMES, *The Lost Apocrypha of the Old Testament, Their Titles and Fragments* (1920). Myth and legend of the Hellenistic period is treated in W.N. STEARNS (ed.), *Fragments from Graeco-Jewish Writers* (1908); C.C. TORREY (ed. and trans.), *Lives of the Prophets* (1946); T.H. GASTER (trans.), *The Dead Sea Scriptures in English Translation,* pp. 256–267, 2nd ed., (1964). The Legends of the Talmud and Midrash are digested and annotated in LOUIS GINZBERG'S classic, *Legends of the Jews,* 7 vol. (1909–39), also available in a one-volume abridgment (1961). The *Midrash Rabbah* has been translated and edited by HARRY FREEDMAN and MAURICE SIMON, 13 vol. in 5, 3rd ed. (1983); and the *Midrash on Psalms,* by W.G. BRAUDE, 2 vol. (1959). Principal editions and translators are listed in H.L. STRACK, *Einleitung in Talmud und Midraš,* 5th ed. (1921; Eng. trans., *Introduction to the Talmud and Midrash,* 1931). Compilations and studies of medieval myth and legend include *The Book of Jashar,* trans. by M.M. NOAH (1840); M.R. JAMES (trans.), *The Biblical Antiquities of Philo* (1917); MOSES GASTER (trans.), *The Chronicles of Jerahmeel* (1899); I.J. KAZIS (ed. and trans.), *The Book of the Gests of Alexander of Macedon* (in Hebrew; 1962); CURT LEVIANT (ed. and trans.), *King Artus* (1969); MOSES HADAS (trans.), *The Book of Delight* (1932). On the diffusion of medieval Jewish tales, see J. JACOBS, *Jewish Ideals, and Other Essays,* pp. 135–161 (1896); LOUIS GINZBERG, "Jewish Folklore: East and West," *Independence, Convergence, and Borrowing in Institutions, Thought, and Art,* Harvard Tercentenary Conference of Arts and Sciences, pp. 89–108 (1937). Judeo-German (Yiddish) works on the subject include MOSES GASTER, *Ma'aseh Book,* 2 vol. (1934); N.C. GORE (ed. and trans.), *Tzeenah u-Reenah: A Jewish Commentary on the Book of Exodus* (1965). Judeo-Spanish (Ladino) works include IGNACIO GONZALEZ LLUBERA (ed.), *Coplas de Yocef* (1935); C. CREWS, "Judaeo-Spanish Folktales in Macedonia," *Folk-Lore,* 43:193–225 (1932). For a treatment of Hasidic legend, see MARTIN BUBER (ed.), *Tales of the Hasidim,* 2 vol. (1947–48). Myths and legends of the Holy Land are treated in DOV NOY (ed.), *Folktales of Israel* (1963); ZEV VILNAY, *Legends of Palestine* (1932; orig. pub. in Hebrew, 1929), and his *The Sacred Land,* 3 vol. (1973–78).

Judicial and Arbitrational Systems

This article deals primarily with the operations of the judicial branch of government. It explores some of the fundamental relationships of this branch with legislative and executive branches and analyzes the functions, the structure and organization, and, finally, the key personnel of courts, such as judges and juries. This article also treats arbitration, another legal means of resolving disputes.

The approach is comparative, contrasting and comparing the systems of the two predominant legal traditions of the contemporary world: first, that of the common law, represented by England, the United States, Canada, Australia, and other nations deriving their legal systems from the English model; and, second, that of the civil law, as represented by nations of western Europe and Latin America and certain Asian and African nations that have modelled their legal systems on western European patterns. Reference is made to the legal institutions in the Soviet Union and other Communist nations that display distinctive characteristics different from those of the civil-law tradition, from which, basically, they developed. A separate section deals more specifically with systems in Communist countries.

The article is divided into the following sections:

Functions of courts

KEEPING PEACE

The primary function of any court system in any nation—to help keep domestic peace—is so obvious that it is rarely considered or mentioned. If there were no agency to decide impartially and authoritatively whether a person had committed a crime and, if so, what should be done with him, other persons offended by his conduct would take the law into their own hands and proceed to punish him according to their uncontrolled discretion. If there were no agency empowered to decide private disputes impartially and authoritatively, self-help, quickly degenerating into physical violence, would prevail and anarchy would result. Not even a primitive society could survive under such conditions. All social order would be destroyed. In this most basic sense, courts constitute an essential element in society's machinery for keeping peace.

DECIDING CONTROVERSIES

In the course of helping to keep the peace, courts are called upon to decide controversies. If, in a criminal case, the defendant denies committing the acts charged against him, the court must choose between his version of the facts and the prosecution's; and if he asserts that his conduct did not constitute a crime, the court must decide whether his view of the law or the prosecution's is correct. In a civil case, if the defendant disputes the plaintiff's account of what happened between them—for example, whether they entered into a certain agreement—or if he disputes the plaintiff's view of the legal significance of whatever occurred—for example, whether the agreement was legally binding—the court again must choose between the contentions of the parties. The issues presented to, and decided by, the court may be either factual, legal, or both.

It would be a mistake, however, to assume that courts spend all of their time deciding controversies. Many cases brought before them are not contested. They represent potential, rather than actual, controversies in which the court's role is more administrative than adjudicatory. The mere existence of a court renders unnecessary any very frequent exercise of its powers. The fact that it operates by known rules and with reasonably predictable results leads those who might otherwise engage in controversy to compose their differences.

The court's administrative role

Most people arrested and charged with crime in the common-law world plead guilty. If they do so understandingly and without coercion of any sort, there is no need to determine guilt, for the sole question is whether the defendant should go to jail, pay a fine, or be subjected to other corrective treatment. In civil-law countries some judicial inquiry into the question of guilt or innocence is required even after a confession. But the inquiry is brief and tends to be perfunctory. The main problem to be resolved, usually without contest, is what sentence should be imposed.

The vast majority of civil cases are also uncontested or, at least, are settled before trial. The court keeps the calendar moving, sometimes encouraging settlement, and decides such questions of law or fact as are presented by the parties; but the number of cases actually tried is small compared to the number settled.

Most divorce cases are uncontested, both parties usually being anxious to terminate the marriage and often agreeing on related questions concerning support and the custody of children. All the court does in such cases is to review what the parties have agreed upon and give its official approval.

Many other uncontested matters come before courts, such as the adoption of children, the distribution of assets in trusts and estates, and the setting up of corporations. Occasionally questions of law or fact arise that have to be decided by the court, but normally all that is required is judicial supervision and approval.

JUDICIAL LAWMAKING

As courts decide controversies they create an important by-product beyond the peaceful settlement of disputes, that is, the development of rules for future cases. Law is thus made not only by legislatures but also by the courts.

To an extent that varies greatly between common-law and civil-law nations, all courts apply preexisting rules formulated by legislative bodies. In the course of doing so, they interpret those rules, sometimes distorting them, sometimes transforming them from generalities to specifics, sometimes filling gaps to cover situations never considered by the original lawmakers. The judicial decisions embodying these interpretations then become controlling for future cases, sometimes to the extent of virtually supplanting the legislative enactments themselves.

The uses of experience in the law: stare decisis

This is one aspect of the doctrine of precedent, or, as it is sometimes called, stare decisis (literally, "to stand by decided matters"). Judges follow earlier decisions, not only to save themselves the effort of working out fresh solutions for the same problems each time they recur but also, and primarily, because their goal is to render uniform and stable justice. If one individual is dealt with in a certain way today, the theory is that another individual engaging in substantially identical conduct under substantially identical conditions tomorrow or a month or year hence should be dealt with in the same way. This, reduced to its essentials, is all that precedent means.

In civil-law nations all judicial decisions are, in theory, based upon legislative enactments, and the doctrine of judicial precedent does not apply. Practice, however, departs from theory. While there are comprehensive legislative codes in these countries, supposedly covering almost every aspect of human conduct and supplying ready-made answers for all problems that can arise, in fact many of the provisions are exceedingly vague and are sometimes almost meaningless until applied to concrete situations, when judicial interpretation gives them specific meaning. Furthermore, the legislative codes cannot anticipate all situations that may arise and come before the courts. The gaps in legislative schemes must be and are filled by judicial decisions, for no court in any nation is likely to refuse to decide a case on the ground that it has not been told in advance the answers to the questions presented to it. Decisions dealing with circumstances unforeseen by the codes and giving specific meaning to vague legislative provisions are published in most civil-law countries and are frequently referred to by lawyers and relied upon by judges. They are not considered "binding," but neither are they forgotten or disregarded. In actual practice, they have almost as much influence as statutory interpretations in nations that formally adhere to the doctrine of stare decisis.

Precedent and common law

It remains true that in common-law countries judicial lawmaking is more pervasive and more frankly acknowledged than in civil-law countries. In addition to rendering decisions that authoritatively interpret statutes, the courts of these nations have created a vast body of law without any statutory foundation whatever. Centuries ago, when there was no legislation to guide them, judges began to decide cases in accordance with their own conceptions of justice. Later judges followed them, deciding like cases in the same manner but distinguishing earlier cases when dissimilar factors were discovered in the cases before them. The later cases also became precedents to be followed in still later cases presenting substantially similar fact patterns. So the process has continued over centuries and is still continuing. The total accumulation of all these judicial decisions is what constitutes "the common law"—the by-product of judges deciding cases and setting forth their reasons. In the common-law nations, legislation is, as a result, more limited in scope than in the civil-law countries. It does not purport to provide for all possibilities but leaves large areas of conduct to be governed solely by judge-made law.

To speak of precedent as "binding" even in common-law systems is misleading. As already noted, earlier decisions can be and are distinguished when judges conclude that they are based upon situations different from those before the court in later cases. Even more significant, earlier decisions can be overruled by the courts that rendered them (not by courts lower in the judicial hierarchy) when the judges conclude that they have proved to be so erroneous or unwise as to be unsuited for current or future application. The Supreme Court of the United States has overruled many of its own earlier decisions, to the consternation of those who yearn for a rigid separation of powers and who are unable to accept the inevitability of judicial lawmaking. Many of these overrulings are in the field of constitutional law, in which legislative correction of an erroneous judicial interpretation of the Constitution is impossible and in which the only alternative is the exceedingly slow, cumbersome, costly, and difficult process of constitutional amendment. Nevertheless, the power to overrule decisions is not restricted to constitutional interpretations. It extends to areas of purely statutory and purely judge-made law as well, areas in which legislative action would be equally capable of accomplishing needed changes. Even in England, which has no written constitution and which has traditionally followed a far more rigid doctrine of stare decisis than the United States, the House of Lords, in its role as the highest court, has announced its intention of departing from precedent "in appropriate cases."

Conflicting views of the court's role

The desirability of judicial lawmaking has long been the subject of lively debate in both civil- and common-law countries. That courts should not arrogate to themselves unrestricted legislative power is universally accepted. But when existing statutes and precedents are outmoded or barbarous as applied to specific cases before the courts, should not judges be able to change the law in order to achieve what they conceive to be just results or, stated differently, to avoid what they consider unjust results?

The extent to which the judges should be bound by statutes and case precedents as against their own ethical ideas and concepts of social, political, and economic policy is an important question, as is the matter of which should prevail when justice and law appear to the judges to be out of alignment with each other. These are questions upon which reasonable persons disagree vigorously even when they are in basic agreement on the proposition that some degree of judicial lawmaking is inevitable. What is mainly at issue is the proper tempo and scope of judicial change. How quickly should judges act to remedy injustice and when should they consider an existing rule to be so established that its alteration calls for constitutional amendment or legislative enactment rather than judicial decision? As many dissenting opinions attest, judges themselves disagree on the answers to these questions, even when they are sitting on the same bench hearing the same case.

CONSTITUTIONAL DECISIONS

Marbury v. *Madison*

In some nations courts not only interpret legislation but determine its validity and in so doing sometimes render statutes inoperative. This happens only in nations that have written constitutions and have developed a doctrine of "judicial supremacy." The prime example is the United States, and the classic statement of the doctrine is the Supreme Court's decision in *Marbury* v. *Madison* (1803), in which Chief Justice Marshall said:

> The powers of the legislature are defined and limited; and that those limits may not be mistaken, or forgotten, the Constitution is written. To what purpose are powers limited, and to what purpose is that limitation committed to writing, if these limits may, at any time, be passed by those intended to be restrained? The distinction between a government with limited and unlimited powers, is abolished, if those limits do not confine the persons on whom they are imposed, and if acts prohibited and acts allowed, are of equal obligation. It is a proposition too plain to be contested, that the Constitution controls any legislative act repugnant to it. . . . It is emphatically the province and duty of the judicial department to say what the law is. Those who apply the rule to particular cases, must of necessity expound and interpret that rule. If two laws conflict with each other, the courts must decide on the operation of each.

Armed with the authority asserted at this early date, the Supreme Court of the United States has held many statutes, federal as well as state, unconstitutional and has also invalidated executive actions that violated the Constitution. Even more surprising is the fact that lower courts also possess and exercise the same powers. Whenever a question arises in any U.S. court at any level as to the constitutionality of a statute or executive action, that court is obligated to determine its validity in the course of deciding the case before it. The case may have been brought for the sole and express purpose of testing the constitutionality of the statute or it may be an ordinary civil or criminal case, in which a constitutional question incidental to the main purpose of the proceeding is raised. Of course, when a lower court decides a constitutional question, its decision is subject to appellate review, sometimes at more than one level. When a state statute is challenged as violating the state constitution, the final authority is the supreme court of that state; when a federal or state statute or a state constitutional provision is challenged as violating the Constitution of the United States, the ultimate arbiter is the Supreme Court of the United States.

In a few American states, questions as to the constitutional validity of a statute may be referred in abstract form to the state's highest court by the chief executive or the legislature for an advisory opinion. This, however, is unusual and, in any event, supplementary to the normal procedure of raising and deciding constitutional questions. The normal pattern is for a constitutional question to be raised at the trial-court level in the context of a genuine controversy and to be decided finally on appellate review of the trial-court decision.

Other methods of constitutional adjudication

The U.S. pattern of constitutional adjudication is not followed in all nations that have written constitutions. In some, such as Germany, there is a special court at the highest level of government that handles only constitutional questions and to which all such questions are referred as soon as they arise. A constitutional question may be referred to the special court in abstract form for a declaratory opinion by a procedure similar to that prevailing in the minority of U.S. states that allow advisory opinions.

In other nations, written constitutions may be in effect but not accompanied by any conception that their authoritative interpretation is a judicial function. Legislative bodies, rather than courts, act as the guardians and interpreters of the constitution, being guided by their provisions but not bound by them in any realistic sense.

Finally, there are some nations, such as England, that have no written constitutions. Here parliamentary supremacy clearly prevails. The courts have no power to invalidate statutes, although they can and do interpret them.

PROCEDURAL RULE MAKING

Distinct from the type of lawmaking just described is a more conscious and explicit type of judicial legislation and one that is less controversial. It is directed toward the rules of procedure by which the courts operate. This is a technical area in which expert knowledge of the type possessed by judges and lawyers is needed; in which constant attention to detail is required; and in which major problems of social, economic, or political policy are seldom encountered. Some legislative bodies, able or willing to devote only sporadic attention to the day-to-day problems of the management of litigation, have delegated the power to regulate procedure to the courts themselves. This is not ad hoc judicial lawmaking as a by-product of deciding cases but openly acknowledged promulgation of general rules for the future, in legislative form, by courts rather than legislatures.

An outstanding example of judicial rule making is found in the United States, where Congress has delegated to the Supreme Court broad power to formulate rules of civil, criminal, and appellate procedure for the federal courts. The Supreme Court also has and exercises the power to amend the rules from time to time as experience indicates that changes are desirable. Congress reserves the power to veto the rules so promulgated but has felt no need to exercise it.

Other legislative bodies, including those of some American states and most of the nations of continental Europe, have been unwilling to repose equal trust in the courts and have retained for themselves the power to regulate procedure. The results have been varied. Courts sometimes become so immersed in day-to-day decision making that they fail to pay adequate attention to the proper functioning of the judicial machinery and so perpetuate rules that are unduly rigid, unrealistic, and unsuited to the needs of litigants, which was the case in England and the American colonies during the 18th and first part of the 19th century. When such a condition occurs, reform through legislative action is indicated. Apart from the occasional necessity of major sweeping changes, however, experience in the common-law countries, at least, indicates that procedural rule making is better vested in the courts than in legislative bodies.

REVIEW OF ADMINISTRATIVE DECISIONS

Existing alongside the courts in any nation are administrative agencies of various kinds. Some do substantially the same kind of work as is done by courts and in substantially the same manner; some have quite different functions such as the issuing of licenses and the payment of welfare benefits.

The relationship between such agencies and regular courts differs markedly between common-law countries and civil-law countries. In common-law countries the actions of administrative agencies are subject to review in the ordinary courts. If the agency is one that decides controversies in substantially the same manner as a court, but in a different and more limited area, judicial control takes much the same form of appellate review as is provided for the decisions of lower courts. The objective of reviewing the record of proceedings is to determine whether the administrative agency acted within the scope of its jurisdiction, whether there was any evidence to support its conclusion, and whether the governing law was correctly interpreted and applied. Administrative decisions are seldom upset by the courts because of a belief on the part of most judges that administrative agencies have special expertise in the area of their specialty. However, they can be and occasionally are upset, thus underscoring the large degree of judicial control over other agencies of government that characterizes common-law systems. If the administrative agency does not engage in formal adjudication, it produces no record of proceedings for judicial review. Nevertheless, its action can be challenged in court by way of trial rather than appeal. The same problems are presented for judicial determination: did the agency act within its jurisdiction, did it correctly follow the law, and was there any rational or factual basis for its action?

In many civil-law countries, the ordinary courts have no control over administrative agencies. Their decisions are reviewed by a special tribunal that is engaged exclusively in that work and that has nothing to do with cases of the type that come into the courts. Its function is solely appellate and solely within the specialized areas entrusted to the administrative agencies. The prototype of this type of tribunal is the Conseil d'État of France.

ENFORCEMENT OF JUDICIAL DECISIONS

The method of enforcing a judicial decision depends upon its nature. If it does nothing more than declare legal rights, as is true of a simple divorce decree (merely severing marital ties, not awarding alimony or the custody of children), or a declaratory judgment (for example, interpreting a contract or a statute), no enforcement is needed. If a judgment orders a party to do or refrain from doing a certain act, as happens when an injunction is issued, the court itself takes the first step in enforcing the judgment by holding in contempt anyone who refuses to obey its order and sentencing him to pay a fine or go to jail. Thereafter, enforcement is in the hands of the executive branch of government, acting through its correctional authorities.

In routine criminal cases and in civil cases that result in the award of money damages, courts have little to do with the enforcement of their judgments. That is the function of the executive branch of government, acting through sheriffs, marshals, jailers, and similar officials. The courts themselves have no machinery for enforcement.

The limits of judicial power

Some judgments are extremely controversial, as was the case with the decision of the Supreme Court of the United States ordering racial desegregation of the schools. When voluntary compliance with such a judgment is refused, forcible methods of enforcement are necessary, sometimes extending to the deployment of armed forces under the control of the executive branch of the government. The withdrawal of executive support seldom occurs, even when decisions are directed against the executive branch itself; but when such executive support is withheld, the courts are rendered impotent. Judges, being aware of their limited power, seldom render decisions that they know to be so lacking in support that they will not be enforced.

Court structure and organization

TYPES OF COURTS

There are many different types of courts and many ways to classify and describe them. Basic distinctions must be made between civil and criminal courts, between courts of general jurisdiction and those of limited jurisdiction, and between trial and appellate courts.

Criminal courts. Criminal courts deal with persons accused of crime, deciding whether they are guilty and, if so, determining the consequences they shall suffer. Prosecution is on behalf of the public, represented by some official such as a district attorney, procurator, or a police officer. Courts are also public agencies, but in this instance they stand neutral between the prosecution and the defense, their objective being to decide between the two in accordance with law.

In civil-law countries a more active role is assigned to the judge and a more passive role to counsel than in common-law countries. In the common-law courts, in which the "adversary" procedure prevails, the lawyers for both sides bear responsibility for producing evidence and they do most of the questioning of witnesses. In civil-law countries, "inquisitorial" procedure prevails, with judges doing most of the questioning of witnesses and having an independent responsibility to discover the facts. This difference pertains more to procedure rather than function.

If a person has been found guilty, he is sentenced, again according to law and within limits fixed by legislation. The objective is not so much to wreak vengeance upon the offender as to rehabilitate him and deter others from following his example. Hence the most common sentences are fines, short terms of imprisonment, and probation (which allows the offender to remain at large but under supervision). In extremely serious cases, the goal may be to prevent the offender from committing further crimes, which may call for a long term of imprisonment or even capital punishment. The death penalty, however, is gradually disappearing from the criminal codes of civilized nations.

Criminal proceedings in any nation inevitably have some educational impact on defendants and upon members of the general public. In Communist nations education is a conscious and primary goal. A basic provision of Soviet law declares:

> By all its activities the court shall educate the citizens of the U.S.S.R. in the spirit of devotion to the Motherland and the cause of communism in the spirit of strict and undeviating observance of Soviet laws, of care for socialist property, of labor discipline, of honesty toward public and social duty, of respect for the rights, honor and dignity of citizens, for the rules of socialist common-life.

Civil courts. Civil courts deal with "private" controversies, as where two individuals (or corporations) are in dispute over the terms of a contract or over who shall bear responsibility for an auto accident. Ordinarily the public is not a party as in criminal proceedings, for it has no interest beyond providing just rules for decision and a forum where the dispute can be impartially and peacefully resolved.

It is possible, however, for the government to be involved in civil litigation if it stands in the same relation to a private party as another individual might stand. Thus, if a postal truck should run down a pedestrian, the government might be sued civilly by the injured person; or if the government contracted to purchase supplies that turned out to be defective, it might sue the dealer for damages in a civil court.

Distinction between civil and criminal actions

The objective of a civil action is not punishment or correction of the defendant or the setting of an example to others but rather to restore the parties so far as possible to the positions they would have occupied had no legal wrong been committed. The most common civil remedy is a judgment for money damages, but there are others, such as an injunction ordering the defendant to do or refrain from doing a certain act or a judgment restoring property to its rightful owner.

Civil claims do not ordinarily arise out of criminal acts. A person who breaks his contract with another or who causes him a physical injury through negligence may have committed no crime but only a civil wrong for which he may not be prosecuted criminally by the public.

There are, however, areas of overlap, for a single incident may give rise to both civil liability and criminal prosecution. In some nations, such as France, both types of responsibility can be determined in a single proceeding under a concept known as adhesion by which the injured party is allowed to assert his civil claim in the criminal prosecution, agreeing to abide by its outcome. This removes the necessity of two separate trials. In common-law countries there is no such procedure, even though civil and criminal jurisdiction may be merged in a single court. Two separate actions must be brought, independent of each other.

Courts of general jurisdiction. Although there are some courts that handle only criminal cases and others that handle only civil cases, a more common pattern is for a single court to be vested with both civil and criminal jurisdiction. Such is the High Court of England and such are many of the trial courts found in U.S. states. Often these tribunals are called courts of general jurisdiction, signifying that they can deal with almost any type of controversy, although in fact they may not have jurisdiction over certain types of cases assigned to specialized tribunals. Often such courts are also described as superior courts, because they are empowered to handle serious criminal cases and important civil cases involving large amounts of money.

Even if a court possesses general or very broad jurisdiction, it may nevertheless be organized into specialized branches, one handling criminal cases, another handling civil cases, another handling juvenile cases, and so forth. The advantage of such an arrangement is that judges can be transferred from one type of work to another, and cases do not fail to be heard for having been instituted in the wrong branch since they can be transferred administratively with relatively little effort.

Courts of limited jurisdiction. Specialized tribunals of many kinds exist, varying from nation to nation. Some deal only with the administration of the estates of deceased persons (probate courts), some only with disputes between merchants (commercial courts), some only with disputes between employers and employees (labour courts). All are courts of limited jurisdiction. Deserving of special mention because of their importance are juvenile courts, empowered to deal with misconduct by children and sometimes also with the neglect or maltreatment of children. Their procedure is much more informal than that of adult criminal courts, and the facilities available to them for the pretrial detention of children and for their incarceration, if necessary after trial, are different. The emphasis is on salvaging children, not punishing them.

Traffic courts also deserve mention because they are so common. They process motor vehicle offenses such as speeding and improper parking. Their procedure is summary and their volume of cases heavy. Contested trials are relatively infrequent.

Finally, in most jurisdictions there are what are called, unfortunately and for want of a better term, "inferior" courts. These are often manned by part-time judges who are not trained in the law. They handle minor civil cases involving small sums of money, such as bill collections, and minor criminal cases carrying light penalties, such as simple assaults. In addition to finally disposing of minor criminal cases, such courts may handle the early phases of more serious criminal cases—fixing bail, advising defendants of their rights, appointing counsel, and conducting preliminary hearings to determine whether the evidence is sufficient to justify holding defendants for trial in higher "superior" courts.

Appellate courts. The tribunals described thus far are trial courts or "courts of first instance." They see the parties, hear the witnesses, receive the evidence, find the facts, apply the law, and determine the outcome.

Above them, to review their work and correct their errors, are appellate courts. These are usually collegiate bodies, consisting of several judges instead of the single judge who usually presides over a trial court. The jurisdiction of the appellate courts is usually general; specialized appellate tribunals handling, for example, only criminal appeals or only civil appeals are rare, although not unknown. The Conseil d'État of France and the Federal Constitutional Court of Germany have already been mentioned as examples of specialization.

Appellate review is not automatic. It must be sought by some party aggrieved by the judgment in the court below. For that reason, and because an appeal may be both expensive and useless, there are far fewer appeals than trials and, if successive appeals are available, as is often the case, far fewer second appeals than original appeals. Judicial systems are organized on a hierarchical basis: at the bottom are numerous trial courts scattered throughout the nation; above them are a smaller number of first level appellate courts, usually organized on a regional basis; and at the apex is a single court of last resort.

Varieties of appellate review

There are three basic types of appellate review. The first consists of a retrial of the case, with the appellate court hearing the evidence for the second time, making fresh findings of fact, and in general proceeding in much the same manner as the court that originally rendered the judgment under attack. This "trial de novo" is used in common-law countries for the first stage of review but only when the trial in the first instance was conducted by an "inferior" court—one typically manned by a part-time judge or two or more such judges, empowered to try only minor cases and keeping no adequate record of its proceedings.

The second type of review is based in part on a "dossier," which is a record compiled in the court below of the evidence received and the findings made there. The reviewing court has the power to rehear the same witnesses again or to supplement their testimony by taking additional evidence, but it need not and frequently does not do so, being content to rely on the record already made in reaching its own findings of fact and conclusions of law. This type of proceeding prevails generally in civil-law countries for the first stage of appellate review, even when the original trial was conducted in a superior court, staffed by professional judges, and empowered to try important or serious cases.

The third type of review is based solely on a written record of proceedings in the court or courts below. The reviewing court does not itself receive evidence directly but concentrates its effort on discovering from the record whether any errors were committed of such a serious nature as to require reversal or modification of the judgment under attack or a new trial in the court below. The emphasis is on questions of law (both procedural and substantive) rather than on questions of fact. This type of review prevails both in civil-law nations and common-law nations at the highest appellate level. It is also used in common-law nations at lower levels when the judgment of a superior court is under attack. The purpose of this type of review is not merely to assure that correct results are reached in individual cases but also to clarify and expound the law in the manner described earlier. Lower courts have little to do with the development of the law, for they ordinarily do not write or publish opinions. The highest appellate courts do, and it is their opinions that become the guidelines for future cases.

Courts in federal systems. Many nations, such as England, France, and Japan, have unitary judicial systems with all courts (that is, regular courts as distinguished from administrative bodies) fitting into a single national hierarchy of tribunals along the lines just described. Other nations, organized on a federal basis, tend to have more complicated court structures, reflecting the fragmentation of governmental powers between the central authority and the local authorities. In the United States, for example, there are 51 separate judicial systems, one for each state and another for the federal government. To a limited extent, the jurisdiction of the federal courts is exclusive of that exercised by the state courts, but there are large areas of overlap and duplication. At the top level is the Supreme Court of the United States, hearing appeals not only from the lower federal courts but also from state courts insofar as they present federal questions arising under the Constitution of the United States or under federal statutes or treaties. If a case in a state court involves only a question of state law—for example, the interpretation of a state statute—the ultimate authority is the state supreme court, and no appeal is possible to the Supreme Court of the United States.

Court structure in a federal form of government need not be as complicated as that in the United States. It is possible to have only one set of courts for the nation, operated by the central government and handling all cases that arise under state law as well as federal law.

Another possibility is for each state or province to have its own system of courts, handling all questions of federal as well as state law, and for the central government to maintain only a single supreme court to decide questions as to the relationship of the central authority and the local authorities or as to the relationship among the local authorities themselves. This is the pattern in Canada and Australia.

Conflict of laws problems

Another complication resulting from a federal form of government is that questions involving conflict of laws arise with great frequency. Such questions concern the choice to be made between the law of one jurisdiction and another as the rule for decision in a particular case. Even in a unitary system, such problems cannot be avoided, for an English court may be called upon to try a case arising from a transaction that took place in France and to decide whether English or French law should govern. Such problems arise much more often, however, in federal systems, where laws differ from state to state and people move about very freely. Their activities in one state sometimes become the subject of a lawsuit in another, requiring the court to decide which law should apply.

JUDGES

A court is a complex institution whose functioning depends upon many people: not only the judge but also the parties, their lawyers, witnesses, clerks, bailiffs, probation officers, administrators, and many others, including, in certain types of cases, jurors. Nevertheless, the central figure in any court is the judge.

Judges vary enormously, not only from nation to nation but often within a single nation. For example, a rural justice of the peace in the United States—untrained in the law, serving part-time, sitting alone in work clothes in a makeshift courtroom, collecting small fees or receiving a pittance for salary, trying a succession of routine traffic cases and little else—obviously bears little resemblance to a justice of the Supreme Court of the United States—a full-time, well-paid, black-robed professional, assisted by law clerks and secretaries, sitting in a marble palace with eight colleagues and deciding at the highest appellate level only questions of profound national importance. Yet both persons are judges.

The levels of legal training

Lay judges. In some civil-law countries, judges at all levels are professionally trained in the law, but in many other nations they are not. In England, part-time lay judges outnumber full-time professional judges by about 60 to 1. Called magistrates or justices of the peace, they dispose of about 97 percent of all criminal cases in that nation and do so with general public satisfaction and the approbation of most lawyers. Professional judges deal only with the most serious crimes, which are relatively few in number; most of their time is devoted to civil cases. England places unusually heavy reliance on lay judges, but they are far from unknown in the courts of many other nations, particularly at the lowest trial level. This is as true in the U.S.S.R. as it is in the United States. There is considerable diversity in the way laymen are chosen and used in judicial work. In the U.S.S.R. and the United States, for example, lay judges are popularly elected for limited terms, whereas in England they are appointed by the lord chancellor to serve until retirement or removal. In the U.S.S.R. and England the lay judges serve intermittently in panels on a rotating basis for short periods, whereas in the United States they sit alone and continuously. In the U.S.S.R. lay judges (who are called assessors) always sit with professional judges; in England, they sometimes do; and in the United States, they never do. In some underdeveloped nations, few judges at any level are legally trained. They are more likely to be priests, for the law they administer is mainly derived from religious teaching, and religion and secular government are often not sharply differentiated. The vast majority of nations that use lay judges at the lowest trial level, however, insist upon professionally trained judges at higher levels: in trial courts of general jurisdiction and in appellate courts.

Professional judges in the civil-law tradition. Professional judges in civil-law countries are markedly different in background and outlook from professional judges in common-law countries. Both are law trained and both perform substantially the same functions, but there the similarities cease. In a typical civil-law country, a person graduating from law school makes a choice between a judicial career and a career as a private lawyer. If he chooses the former and is able to pass an examination, he is appointed to the judiciary by the minister of justice (a political officer) and enters service in his early 20s. His first assignment is to a low-level court; thereafter, he works his way up the judicial ladder as far as he can until his retirement on a pension. His promotions and assignments depend upon the way his performance is regarded by a council of senior judges, or sometimes upon the judgment of the minister of justice, who may or may not exercise his powers disinterestedly and on the basis of merit. The civil-law judge, in short, is a civil servant.

Professional judges in the common-law tradition. In common-law nations, the path to judicial office is quite different. Upon completion of his formal education, a person typically spends 15, 20, or 25 years in the private practice of law or, less commonly, in law teaching or governmental legal service; then, at about age 50, becomes a judge. He takes no competitive examination but is appointed or elected to office. In England the appointive system prevails for all levels of judges, including even lay magistrates. Appointments are primarily under the control of the lord chancellor, who, although a cabinet officer, is also the highest judge of the realm. They are kept surprisingly free from party politics. In the United States, the appointive method is used in federal courts and some state courts, but it tends to be highly political. Appointments are made by the chief executive of the nation or state and are frequently subject to legislative approval. In many states, judges are popularly elected, sometimes on nonpartisan ballots, sometimes on partisan ballots with all the trappings of traditional political contests. In an attempt to de-emphasize political considerations and yet maintain some measure of popular control over the selection of judges, a third method of judicial selection has been devised and is slowly growing in popularity. Called the Missouri Plan, it involves the creation of a nominating commission that screens judicial candidates and submits to the appointing authority a limited number of names of persons considered qualified. The appointing authority must make his choice from the list submitted. The person chosen as judge then assumes office for a limited time, and, after the conclusion of this probationary period, he stands for "election" for a much longer term. He does not run against any other candidate but only "against his own record."

Political status of the judiciary

In common-law countries, a person does not necessarily enter the judiciary at a low level; he may be appointed or elected to his nation's highest court or to one of its intermediate courts. He does not look forward to any regular pattern of promotion, nor is he necessarily assured of long tenure with ultimate retirement on a pension. In some courts, life tenure is provided, usually subject to mandatory retirement at a fixed age. In others, tenure is limited to a stated term of years. At the conclusion of his term, if not mandatorily retired earlier, the judge must be reelected or reappointed if he is to continue.

While in office, the common-law judge enjoys greater power and prestige and more independence than his civil-law counterpart. He occupies a position to which most members of his profession aspire. He is not subject to outside supervision and inspection by any council of judges or by a minister of justice; nor is he liable to be transferred by action of such an official from court to court or place to place. The only administrative control over him is that exercised by judicial colleagues, whose powers of management are generally slight, being limited to such matters as requiring periodical reports of pending cases and arranging for temporary (and usually consensual) transfers of judges between courts when factors such as illness or congested calendars require them. Only if a judge misbehaves very badly is he in danger of disciplinary sanctions and then usually only by way of criminal prosecution for his misdeeds or legislative impeachment and trial, resulting in removal from office—a very cumbersome, slow, ill-defined, inflexible, ineffective, and seldom used procedure. In parts of the United States, newer and more expeditious methods of judicial discipline are developing in which senior judges are vested with power to impose sanctions ranging from reprimand to removal from office of erring colleagues. They are also vested with power to retire judges who have become physically or mentally unfit to discharge their duties.

Discipline of judges

Except at the very highest appellate level, common-law judges are no less subject than their civil-law counterparts to appellate reversals of their judgments. But appellate review cannot fairly be regarded as discipline. It is designed to protect the rights of litigants; to clarify, expound, and develop the law; and to help and guide rather than reprimand lower court judges. (D.K.)

JURIES

The jury is a historic legal institution in which a group of laymen participate in a major way in deciding cases brought to trial. Its exact characteristics and powers depend on the laws and practices of the countries, provinces, or states in which it is found, and there is considerable variation. Basically, however, it recruits laymen at random from the widest population for the trial of a particular case and allows them to deliberate in secrecy, to reach a decision by other than majority vote, and to make it public without giving reasons.

History and use. The jury's origin is lost in the past. It may have been indigenous to England or have been brought there by the Norman invaders in 1066. Originally, the jurors were neighbourhood witnesses who passed judgment based on what they themselves knew. But the breakdown of medieval society and the growth of the towns changed this; the jury was called upon to determine the facts of the case, based upon the evidence presented in court. The availability of the jury in the king's courts may have been a key factor in centralizing the nation's courts under the king and in creating the common law. By the 15th century, nonrational modes of trial such as ordeal, in which the defendant was subjected to various tortures that, if successfully endured, proved his innocence, were replaced by the jury trial, which became the established form of trial for both criminal and civil cases at common law.

Two forces moved the jury abroad. One was the expansion of the British Empire, which brought the jury to Asia, Africa, and the American continent. The other was the French Revolution and its aftermath, which brought it, as a symbol of popular government, to the European continent: first to France itself, then, through Napoleon, to the Rhineland, later to Belgium, most of the remaining German states, Austria-Hungary, Russia, Italy, Switzerland, Holland, and Luxembourg, although the last two abolished it immediately after Napoleon's defeat. In each of these countries, use of the jury was from the outset limited to trials of major crimes and of political crimes against the state.

Beginning in the mid-19th century, the jury was weakened in a variety of ways: in 1850, Prussia, for example, removed treason from its jurisdiction; in 1851, the duchy of Nassau removed all political crimes; in 1923, Czechoslovakia removed treason and, one year later, libel; in 1919, Hungary suspended jury trial entirely and never restored it. Germany abandoned the jury in 1924. Both the Soviet bloc and the fascist states abolished it outright; France never restored the jury abolished during the German occupation in the 1940s, and Japan did away with its short-lived jury courts in 1943. After World War II, Austria reintroduced the jury in a weakened form.

United States as major user of jury system

Thus, there are three important points about the history and development of the jury as a legal institution: first, the effort to introduce it outside the Anglo-American legal orbit has failed; further, in England itself, its use was limited by statute to a small category of cases; and thus, the United States has emerged today as the home of the jury system for both criminal and civil cases. Some 120,000 jury trials are conducted there annually, more than 90 percent of all jury trials in the world.

Use of the jury in the United States depends on two factors: the degree to which it is available as a matter of right, and the degree to which the parties themselves choose to use it. The laws as to its availability have varied from state to state, but in 1968, in *Duncan* v. *Louisiana,* the United States Supreme Court declared that a jury trial is a constitutional right in all criminal cases in which the penalty may exceed six months' imprisonment. In civil cases its constitutional status is less clear, but in general jury trial is available. The practice of allowing the parties to waive a jury trial also varies widely from region to region, and, as a result, the number of jury trials per year also varies widely. The annual number of criminal jury trials per 100,000 population ranges between 3 for Connecticut to 144 for Georgia.

Jury procedures. *Selection.* Historically, there were some minimum requirements of property and competence for jury service. More recently, the idea of genuine random selection from the population, to achieve a cross section of the community, has been gaining ground. Since 1969, it has been the principle of selection in the federal courts. Most jurisdictions exempt some groups from jury service: policemen, lawyers, doctors, and so on. Some jurisdictions exempt women entirely. All jurisdictions excuse jurors if the service imposes undue hardship.

The voir dire

The commitment of important decisions to a random group of laymen has been moderated, particularly in the United States, by an elaborate screening, voir dire, conducted by trial counsel at the inception of a trial. The law permits counsel to challenge prospective jurors either for cause (if there is specific likelihood of bias) or, for a limited number, "peremptorily"—that is, without having to give a reason. American trial tradition attaches a great deal of significance to the strategies of juror selection, and, in celebrated cases, the lawyers' voir dire examination has extended for several weeks.

Size and unanimity. Traditionally, the jury had 12 members and was required to reach its decisions with unanimity, a striking arrangement in Anglo-American countries that make all other decisions by majority vote. Over the years some modifications have been made. Some jurisdictions prescribe or allow in minor cases a jury of six. Oregon allows 10:2 verdicts—that is, a majority of 10—in all criminal cases, except capital ones; and in 1968 Great Britain followed the Oregon example. A few Southern states in the United States allow majority verdicts in misdemeanour trials. In civil cases, many states now allow 10:2 verdicts. When the required number (12 or 10) of jurors cannot agree on a verdict (termed a hung jury in the United States), the judge declares a mistrial, which means the case, unless it is withdrawn, must be tried anew. It is somewhat remarkable that "hung" juries occur with relative infrequency even when unanimity is required. In Europe juries operate under a different principle. Unless at least two-thirds of all the jurors vote guilty, the defendant must be acquitted. The U.S. Army court-martial jury also operates under this principle.

Sentencing. Although in civil cases the jury decides both issues of liability and amount of damages, in criminal cases it has been restricted generally to the issues of guilt, with punishment left to the judge. In some Southern U.S. states, however, the jury also decides the sentence within a range that the law provides. And in all jurisdictions that have retained the death penalty, if the jury finds the defendant guilty of the capital crime, it decides, or at least expresses an opinion, as to whether the death penalty is to be imposed. In most jurisdictions decisions on guilt and sentence are rendered simultaneously, but California has introduced the so-called second trial in capital cases, which occurs after a guilty verdict. At such a "second trial" pleas and evidence are presented for and against the imposition of the death penalty in the specific case, and only then is the jury asked to determine the sentence.

Relation of judge and jury

Control. Trial by jury is, of course, trial by jury under the supervision of a judge. The formula for sharing power between judge and jury is complex. First, the judge decides what the jury may or may not hear under the rules of evidence. Second, if the judge finds that the evidence presented leaves no factual issue to be resolved, he may withdraw the issue from the jury and direct the jury to acquit a defendant or, in a civil trial, find for either plaintiff or defendant; he cannot, however, direct a guilty verdict in a criminal trial. Third, in some jurisdictions the judge may, and often will, summarize the evidence or even discuss its weight. Fourth, the judge instructs the jury as to the law it should apply in reaching the verdict. Finally, if the judge finds the jury's verdict to be manifestly against the weight of the evidence, he may with one exception set it aside and order a new trial. The only exception is in a criminal case in which the jury renders an acquittal; under Anglo-American law (though not under European continental law) the jury's acquittal is always final.

The jury normally renders a general verdict, that is, a yes or no answer to liability or guilt, and does not give reasons for its decision. At times, however, courts employ "special verdicts" or "special interrogatories" in which the jurors are asked to decide a series of specific factual issues that bear on the overall verdict.

The controversy over the jury. The jury has been enmeshed in a perennial debate as to its merits, a debate that has recruited some of the great names in law and political philosophy, from Montesquieu, William Blackstone, and Thomas Jefferson to present-day theorists and practitioners, and has centred on three issues. First, there is the debate about collateral aspects: there are favourable contentions that the jury provides an important civic experience, that it makes tolerable the stringency of certain decisions, that it acts as a sort of lightning rod for animosity that otherwise might centre on the more permanent judge, and that the jury is a guarantor of integrity since it is said to be more difficult to bribe 12 people than one. Against this it has been urged that jury duty disenchants the citizen, that it imposes an unfair burden, that the jury is expensive, and that it makes it difficult to do away with the often interminable delays that exist in civil litigation.

The jury's competence

Second, there is the issue of the jury's competence. It is argued that the judge, by training, discipline, experience, and superior intelligence, is better able to understand law and facts than laymen drawn from a broad range of levels of intelligence, without experience and without durable official responsibility. But it is also argued that 12 heads are better than one, that the jury as a group has wisdom and strength beyond that of its individual members, that it makes up in common sense and experience what it lacks in training, and that its very inexperience is an asset because it secures a fresh perception of each trial, avoiding the stereotypes that may infect the judicial eye.

Finally, there is the question of the jury's interpretation of the law. The critics complain that the jury will not follow the law, either because it does not understand it or because it does not like it, and hence will administer justice unevenly and that the jury produces a government by men and not by rule of law, against which Anglo-American political tradition is so steadfastly set. The jury's champions offer this very flexibility as its most endearing characteristic. They see the jury as a remarkable device for ensuring that the rigidity of the general rule can be shaped to justice in a particular case, with government by the spirit of the law and not by its letter.

Performance. In a recent survey of some 7,000 jury trials, the presiding judges were requested to reveal how they would have decided without a jury; the results of the survey provided some major insights into the actual performance of the contemporary American jury. In both civil and criminal trials, judge and jury agreed in 78 percent of all verdicts. In civil cases the disagreement in the remaining cases was symmetrically split; in 19 percent of the criminal cases, however, the judge would have convicted, whereas the jury acquitted. The letter of the law confines the jury to "finding the facts," but the deviations from the judge are mostly due to the jury's subtle, and not always conscious, injecting its sense of justice into a case that might go either way. This sense of justice may be concerned with the person of the accused, with the threat of too harsh a punishment, or with the content of the criminal law rules. Thus, close study of the jury has revealed it as a highly sensitive institution, subtle and discerning, moved by factors far beyond gross sympathy for the defendant. On the whole, the system tolerates and even appreciates these deviations of the jury from the judge, even if in rare cases they reflect what the national community experiences as intolerable local prejudice.

(H.Ka./H.Z./Ed.)

OTHER JUDICIAL OFFICIALS

In most countries there are other officials who serve the court. Court clerks, who are responsible for case records and documents, and bailiffs, who are in charge of keeping order, are found in most judicial systems. Also prevalent are officers who prosecute cases in the government's name: states attorneys and district attorneys in the United States, procurators-general in the U.S.S.R., and *procureurs généraux* in France.

Probation officers are found in many countries including the U.S. and Japan. Notaries in France, Italy, and the U.S.S.R. have greater powers than their counterparts in the U.S. In fact, they perform many services carried out by lawyers in the common-law system, such as drafting and verifying wills and contracts and preparing petitions for presentation in court.

Certain countries have officials that are particularly indigenous to their country or legal system. France, for example, has a *juge d'instruction,* who is responsible for the preliminary investigative proceedings prior to a criminal trial.

THE STRUCTURE AND STATUS OF THE JUDICIARY UNDER COMMUNISM

Although the essential legal institutions of the Soviet Union and other Communist countries are based on the civil-law system, certain features are unique. These characteristics are partly the result of the Soviet Union's attitudes toward law that antedate the Soviet system, but mostly they result from the attempt to reconcile Marxist theory with the institutional needs of a modern society.

According to Marx and his followers, the legal system, like all other governmental structures and instruments of class oppression, would "wither away" in a Communist society; thus the courts that existed after the Revolution were considered temporary institutions, required only during the transition to Communism. The ordinary and traditional business of the courts was carried on by the so-called people's courts, while "revolutionary tribunals" dealt with individuals the government considered to be political opponents. A nonjudicial body in the hands of the secret police (at first called Cheka, later OGPU and NKVD), operating in the style of an administrative agency, also heard cases and handed out sentences—usually of the severest kind.

The Soviet three-level court system

In 1921 some capitalist measures were temporarily introduced to revive the economy, and this necessitated some stabilization of the legal system and its institutions. A three-level system of courts with civil and criminal jurisdiction was established in 1922 for the Russian Republic, which in the same year formed a federation with the other soviet republics under its jurisdiction, making up the Union of Soviet Socialist Republics. A new constitution created a federal court—the U.S.S.R. Supreme Court—and a federal judiciary act of 1924 established uniform principles for the judiciary throughout the republics, patterned largely after the system adopted by the Russian Republic. The basic structure of the courts laid down at that time has remained essentially the same to the present, with some minor changes and reforms.

The "People's Courts" on the local level are courts of original jurisdiction for minor criminal cases and a large number of civil cases. The next level, the provincial courts, receive appeals from the people's courts and have original jurisdiction over political and serious civil and criminal cases. The highest level in each republic is its supreme court, which hears appeals from the provincial courts, disciplines lower courts, and has some original jurisdiction over extremely serious cases.

On all three levels, appellate cases are tried by a court consisting of three full-time judges, whereas one judge and two lay judges, or assessors, preside over cases on first hearing. Judges of the people's courts are popularly elected every five years, and judges on the provincial and supreme court levels are "elected" by soviets (bodies combining legislative and executive functions) of the corresponding levels of government. All judges may be recalled before the expiration of their terms by those who elected them.

The federal court system is twofold. There are courts called military tribunals that deal with charges against men in the armed forces and with charges of espionage brought against civilians. The other federal body is the U.S.S.R. Supreme Court—the highest judicial body—which has original jurisdiction in a few special cases relating to the survival of the regime, appellate power over the decisions of the supreme courts of the republics or decisions of the military tribunals, and the right to issue directives to all inferior courts in matters of administration of justice on the basis of a series of its decisions. Although Soviet legal theory is patterned after that of civil-law countries in that it does not recognize judicial lawmaking, these Supreme Court directives function as a source of law and are binding on all courts. The status of the judiciary in the Soviet Union has undergone some changes that parallel the institutional changes since the early days of the Revolution. The system organized in 1922 had the stated purpose of safeguarding the conquests of the Revolution and establishing the dictatorship of the proletariat. Judges were called upon to use their "revolutionary conscience" in deciding cases, and the doctrine of impartiality and independence of the judiciary was repudiated. With the passage of time, however, the Soviet rulers found the need for legal institutions of a stable nature increasing rather than decreasing, and the goal of the legal system was changed from protection of a particular class to protection of the socialist order and the rights of all citizens. Although the role of the judiciary is still conceived of as a political task, there is some acceptance of the idea that judges should be independent and impartial. Marxist philosophy notwithstanding, the Soviet Union and other Socialist countries are confronted with a growing need for legal institutions to fulfill many of the same functions as those in the West. One attempt to fill this need in recent years has been the appearance of "social organizations," such as the "comrades' courts," which are described as voluntary organizations using persuasion and social influence to deal with matters that would otherwise come before a court. But these organizations are party controlled, have only limited power to impose sanctions, and do not appear at present to offer an effective alternative to the type of legal institutions that have been developing within Soviet society.

The status of judges

The other Communist countries, both in eastern Europe and Asia, adopted legal institutions patterned largely after the Soviet model. Since Stalin's death, however, there have been some modifications in the eastern European countries, coinciding with the reforms in the civil and criminal codes adopted by the Soviet Union in the late 1950s and early 1960s. Chinese leaders, however, have resisted efforts to codify their laws, preferring flexibility in their courts, and they have abandoned the policy of copying Soviet legal patterns.

(Ed.)

Arbitration

Arbitration is a nonjudicial, legal technique for resolving disputes by referring them to a third party for a binding decision, or "award," as an arbitrator's findings are usually described. The arbitrator may be a single person or an arbitration board, usually of three members. Arbitration is most commonly resorted to for the resolution of commercial disputes and must be distinguished from mediation and conciliation, which are common in the settlement of labour disputes between management and labour unions. In such cases, the parties resort to a third person to offer a recommendation for a settlement or to help them to reach a compromise. Such intervention by a third party, which also occurs in international disputes between states in the form of diplomatic intervention and good offices, has no binding force upon the disputants, as has the arbitrator's decision, the award.

COMMERCIAL ARBITRATION

Commercial arbitration is a means of settling disputes by referring them to a third person, an arbitrator, selected by the parties for a decision based on the evidence and arguments presented to the arbitration tribunal. The parties agree in advance that the decision will be accepted as final and binding upon them.

Historically, commercial arbitration was used in resolving controversies between medieval merchants, in fairs and marketplaces in England and on the European continent, and in the Mediterranean and Baltic sea trade. The increased use of commercial arbitration became possible after courts were empowered to enforce the parties' agreement to arbitrate. The first such statute was the English Arbitration Act of 1889, now consolidated into an act of 1950 and adopted by arbitration statutes in most countries of the Commonwealth. It was followed in the United States by an arbitration statute of the state of New York in 1920 and a Federal Arbitration Act of 1925. Codified in 1940, the latter deals with the enforcement in federal courts of arbitration agreements and awards in maritime transactions and those involving interstate and foreign commerce. Most states of the United States adopted, sometimes with minor changes, the Uniform Arbitration Act of 1955, as amended in 1956, which had been promoted by the Commissioners on Uniform State Laws and recommended by the American Bar Association. This act provides for the judicial enforcement of an agreement to arbitrate existing and future disputes, thereby making the arbitration agreement no longer revocable, as it had been under common law. It also provides for the substitution of arbitrators in the event of a party failing to select an arbitrator and for a suspension of any court action instituted in contravention of a voluntary arbitration agreement. The courts thereby play an important role in implementing arbitration agreements, making judicial assistance available against a recalcitrant party. This concept of modern arbitration law, which recognizes the irrevocability of arbitration agreements and the enforceability of awards prevails also in the arbitration statutes of nearly all countries of Europe and Asia. Latin American procedural laws generally provide only for court enforcement of agreements to arbitrate existing disputes and do not provide for the enforcement of subsequent disputes that may arise under the arbitration agreement.

Statutory development of arbitration

Function and scope. Arbitration has been used customarily for the settlement of disputes between members of trade associations and between different exchanges in the securities and commodities trade. Form contracts contain a standard arbitration clause referring to the arbitration rules of the respective organization. Numerous arrangements between parties in industry and commerce also provide for arbitration of controversies arising out of contracts for the sale of manufactured goods, for terms of service of employment, for construction and engineering projects, for financial operations, for agency and distribution arrangements, and for many other undertakings.

The usefulness and significance of arbitration is demonstrated by its increasing use by the business community and the legal profession in many countries of the world. The primary advantage is the speed with which controversies can be resolved by arbitration, compared with the long delays of ordinary court procedure. The expert knowledge of arbitrators of the customs and usages of a specific trade makes testimony by others and much documentation unnecessary, thereby eliminating expenses connected with court procedures. The privacy of the arbitration procedure is also much valued by parties to the controversy; situations unfavourable to the party's credit or deficiencies in manufactured goods revealed in arbitration proceedings do not become known to outsiders. There are, however, also disadvantages in the arbitration process. The fact that in Anglo-American practice no reasons are given by the arbitrator to accompany his award prevents the development of a guideline for the further conduct of business relations. This uncertainty resulting from lack of reasoned precedents, moreover, makes the arbitral decision less predictable. Further obstacles to the wider use of commercial arbitration are the divergencies in municipal laws and court decisions that result in different interpretations of similar arbitration questions and the fact that awards are not published: publication of awards, even without identification of the parties, might assist in the establishment of precedents useful in discouraging future disputes on similar issues in a specific branch of industry or commerce.

Advantages of arbitration over court procedure

Procedure. The method of selecting arbitrators is an important aspect of the arbitration process, for the arbitrator's ability and fairness is the decisive element in any arbitration. The general practice is for both parties to select an arbitrator at the time a conflict arises or at the time the arbitration agreement is concluded. The two arbitrators then select a chairman, forming a tribunal. Selection of arbitrators is also often made by agencies administering commercial arbitration under preestablished rules of procedure. These organizations—various trade associations, produce exchanges, and chambers of commerce in many countries—maintain panels of expert arbitrators. The parties may either make their own selection or entrust the appointment of the arbitrators to the organization.

Challenges to the arbitration process are not uncommon. A party may claim, for example, that no valid arbitration agreement came into existence because the person signing the agreement had no authority to do so or that a condition precedent to arbitration had not been fulfilled. More often, arbitration is contested on the ground that the specific controversy is not covered by the agreement. In such cases, the issue of whether the arbitrator has authority to deal with the conflict is usually determined by a court. Further challenges to the arbitration process may be directed against an arbitrator, on grounds, for example, of alleged lack of impartiality. Any such challenge can generally be maintained only after the arbitration has been concluded, as courts are reluctant to interfere with the arbitration process before an award has been rendered.

The arbitration process is governed by the rules to which the parties referred in their agreement; otherwise, the procedure will be determined by the arbitrators. The arbitration proceedings must be conducted so as to afford the parties a fair hearing on the basis of equality. The arbitrator generally has the authority to request the parties and third persons to produce documents and books and to enforce such a request by issuing subpoenas. If a party fails to appear at a properly convened hearing, without showing a legitimate cause, the arbitrator in most instances will proceed in the absence of that party and render an award after investigation of the matter in dispute.

Under the law and arbitration practice of most countries, the award is valid and binding upon the parties when rendered by a majority of the arbitrators, unless the parties expressly request a unanimous decision of the arbitrators, which they seldom do. The statutory law of various countries and the rules of agencies administering commercial arbitration contain provisions on the form, certification, notification, and delivery of the award, with which requirements the arbitrator has to comply.

A much-disputed question in commercial arbitration concerns the law to be applied by the arbitrators. Generally, the award must be based upon the law as determined by the parties in their agreement. This failing, the arbitrator must apply the law he considers proper in accordance with the rules of conflict of laws. In both cases, the arbitrator will have to take account of the terms of the contract and the usage of the specific trade. If, during any arbitration proceeding, a compromise is reached by the parties, that compromise may be recorded as an award by the arbitrator.

Appeal of arbitration decisions to the courts

Appeals to the courts from the award cannot be excluded by agreement of the parties, since the fairness of the arbitration process as a quasi-judicial procedure has to be maintained. Any court control is, however, confined to specific matters, usually enumerated in the arbitration statutes, such as misconduct of the arbitrator in denying a party the full presentation of its claim or not granting a postponement of the hearing for good cause. A review of the award by a court generally will not deal with the arbitrator's decisions as to facts or with his application of the law. The competence of the courts is restricted in order not to make the arbitration process the beginning of litigation instead of its end. Recognition of an award and its enforcement will be denied when it appears to be contrary to public policy, as might be the case, for example, in cases involving trusts (monopolies), industrial property rights, or violation of foreign-currency restrictions. An arbitration award has the authority of a court decision and may be enforced by summary court action according to the procedural law of the country in which execution is being sought.

International commercial arbitration. International commercial arbitration between traders of different countries has long been recognized by the business community and the legal profession as a suitable means of settling trade controversies out of court. The procedure in international commercial arbitration is basically the same as in domestic arbitration. In order to establish more uniformity in procedure and to make access to arbitration facilities more easily available, the United Nations economic commissions in 1966 published new rules applying to international arbitration. Those for Europe are contained in the "Arbitration Rules of the United Nations Economic Commission for Europe" and for Asia and the Far East in the "Economic Commission for Asia and the Far East Rules for International Commercial Arbitration."

The development of international commercial arbitration has been furthered by uniform arbitration legislation prepared by the United Nations Conference on International Commercial Arbitration in 1958 and by the Council of Europe and the Inter-American Juridical Committee of the Organization of American States. One particularly difficult problem of international commercial arbitration is the enforcement of awards in a country other than the one in which they were rendered. Statutory municipal laws do not usually contain provisions for the enforcement of foreign awards, and parties are faced with uncertainty about the law and practice of enforcement procedure in a country other than their own.

The aforementioned international agreements, to which most of the trading nations of the world adhere, facilitate the enforcement of foreign awards to the extent that no further action is necessary in the country in which the award was rendered: the opposing debtor must establish that the award has been set aside or that its effects have been suspended by a competent authority, thus shifting the burden of proof of the nonbinding character of the award to the losing party.

Further development of international commercial arbitration is encouraged by the United Nations Commission on International Trade Law, which aims at promoting the harmonization and unification of laws in the field of international commercial arbitration.

LABOUR ARBITRATION

Labour arbitration—the reference of disputes between management and labour unions to an impartial third party for a final resolution—is usually the last step under a collective-bargaining agreement after all other measures to achieve a settlement have been exhausted. Labour arbitration is not, as is commercial arbitration, an auxiliary avenue of justice and thereby a substitute for ordinary court procedure but a technique also for settling or avoiding strikes.

Two major aspects of labour arbitration

Two major aspects of labour arbitration are usually distinguished: arbitration of rights and arbitration of interests. Arbitration of rights refers to the arbitration of an existing labour contract when a dispute over the application of that contract arises between labour and management. Arbitration of interests refers to arbitration between labour and management during the negotiation of a new labour contract.

Arbitration of rights. Arbitration of rights under the terms of a collective-bargaining agreement is employed in the United States far more than in most other countries. Outside the United States, labour courts, industrial courts, or conciliation and arbitration commissions perform the function of arbitrating rights. These bodies are usually appointed by the government, and recourse to them is frequently compulsory.

More than 90 percent of the collective-bargaining agreements in the United States provide for arbitration as a last step in the grievance procedure. Employees, for example, through their union, may present for arbitration complaints concerning such matters as discipline, discharge, and violations of working conditions. Other issues frequently submitted to arbitration customarily concern premium payments and incentive rates, overtime and vacations, Christmas bonuses, seniority rights, and fringe benefits, such as pension and welfare plans.

The arbitrator's decision must be based on the collective-bargaining agreement, which provides for the application of an existing contract to the grievance presented. The question of whether the various steps in the grievance procedure have been complied with before the initiation of the arbitration is usually left to the determination of the arbitrator and not of a court. The question, however, of whether the disputed issue is covered by the collective-bargaining agreement has to be determined by a court and not by the arbitrator. This authority of the courts was upheld by the Supreme Court of the United States in 1960.

Selection of labour arbitrators

The choice of arbitrator is made either by naming him in the agreement or, more often, by leaving the choice open until a dispute has arisen. Frequently, only a single arbitrator is appointed—usually an expert in the field of industrial relations. Alternatively, tripartite arbitration boards are established, both parties appointing their own arbitrator, who acts somewhat as advocate. A neutral chairman is selected either by the parties or by the two party-appointed arbitrators.

A further technique of arbitration of rights is the appointment of a single permanent arbitrator, or "umpire," to resolve disputes for the duration of the collective-bargaining agreement. The umpire will be intimately acquainted with the various economic, financial, and other aspects of the particular industry and will be familiar with the relationship between management and union developed in the past. He sometimes follows precedents, especially those established by his predecessor. This permanent umpire system originated in the United States in the anthracite-coal industry at the beginning of the 20th century and has been employed in such other important industries as newspaper printing and clothing.

Labour arbitrators render binding decisions on the disputes submitted to them. They are not bound by strict rules of court procedure, especially as regards burden of proof and the presentation of evidence. As arbitrators, they have the power to subpoena persons and written evidence. Labour arbitrators tend to evaluate factual evidence rather freely and often reduce penalties imposed upon employees by the management for breach of the labour contract. Even minor questions, such as the use of company time by employees for washups or coffee breaks, are submitted to arbitration, in order to establish precedents in the operation of the plant. Generally, however, arbitrators are not bound to follow previous decisions.

Decisions of labour arbitrators are seldom reviewed by the courts, as awards are usually fully complied with by both parties.

Arbitration of interests. Arbitration of the terms of a new contract, referred to as arbitration of interests, may be instituted if management and the labour union are unable to agree on a new contract. In some industries, such as hotels, printing, transit, and utilities, such disputes are submitted to arbitration. In most countries, however, management and union are seldom inclined to forgo resort to lockouts and strikes in an attempt to obtain favourable new contracts, and interest arbitration is thus seldom used.

Compulsory arbitration, directed by legislative fiat, has been a controversial issue in the settlement of industrial disputes. It has been favoured in disputes in the transportation industry, which may involve great public inconvenience, and in disputes in the public-utilities sector when an immediate danger to public health and safety might occur. Compulsory arbitration has been declared unconstitutional in some states of the United States. More recently, however, it has been adopted as a regular procedure for the settlement of disputes with municipal employees in some U.S. cities.

INTERNATIONAL ARBITRATION

Controversies between sovereign states that are not settled by diplomatic negotiation or conciliation are often referred, by agreement of both parties, to the decision of a third disinterested party, who arbitrates the dispute with binding force upon the disputant parties. Such arbitration between states has a long history: it was used between city-states in ancient Greece and also in the Middle Ages, when the pope often acted as the sole arbitrator.

Historical development. The modern development of international arbitration can be traced to the Jay Treaty of 1794 between Great Britain and the United States, which established three arbitral commissions to settle questions and claims arising out of the American Revolution. In the 19th century many arbitral agreements were concluded by which arbitration tribunals were established ad hoc to deal with a specific case or to handle a great number of claims. Most significant was the "Alabama" claim arbitration under the Treaty of Washington (1871), by which the United States and Great Britain agreed to settle claims arising from the failure of Great Britain to maintain its neutrality during the U.S. Civil War.

Commissions consisting of members drawn from both disputant countries ("mixed arbitral commissions") were often used in the 19th century to settle pecuniary claims for compensation of injuries to aliens for which justice could not be obtained in foreign courts. Such was the purpose of a convention of 1868 between the United States and Mexico, by which claims of citizens of each country arising from the Civil War were settled. Boundary disputes between states were also often settled by arbitration.

Development of international arbitration procedures

International arbitration was given a more permanent basis by the Hague Conference of 1899, which adopted a convention on the pacific settlement of international disputes, revised by a Conference of 1907. The convention stated that:

> International arbitration has for its object the settlement of disputes between States by judges of their own choice and on the basis of respect for law. Recourse to arbitration implies an engagement to submit in good faith to the award.

A Permanent Court of Arbitration, composed of a panel of jurists appointed by the member governments, from which the litigant governments may select the arbitrators, was established at The Hague in 1899.

Twenty cases were arbitrated between 1902 and 1932, but, from that year until 1972, only five cases were dealt with. This was largely because the importance of the Permanent Court of Arbitration was lessened by the Permanent Court of Justice (established in 1922) and its successor, the International Court of Justice. More recently, in 1960, the court, which was originally devised for the settlement of disputes between states, has offered its services for the arbitration of controversies between states and individuals or corporations. Such a dispute was arbitrated in 1970 between a British company and the government of the Democratic Republic of Sudan. The case concerned the repudiation of a contract for building houses in the irrigation zone of the Khashm Al Qirbah Dam in The Sudan.

Arbitration provisions of international treaties. There are several multilateral treaties that provide for the pacific settlement of international disputes by arbitration, including the Geneva General Act for the Settlement of Disputes of 1928, adopted by the League of Nations and reactivated by the General Assembly of the United Nations in 1949, which provides for the settlement of various disputes, after unsuccessful efforts at conciliation, by an arbitral tribunal of five members. Other such treaties include the General Treaty of Inter-American Arbitration, signed in Washington in 1929, and the American Treaty on Pacific Settlement of Disputes, signed in Bogotá in 1948. More recently, the Council of Europe adopted the European Convention for the Peaceful Settlement of Disputes (1957).

Arbitration is also mentioned as a proper method of settling disputes between countries in the Charter of the United Nations, as it was in the Covenant of the League of Nations.

Model rules of the International Law Commission

The International Law Commission of the United Nations submitted to the General Assembly in 1955 a Convention on Arbitral Procedure. Its model rules would not become binding on any member-state of the United Nations unless they were accepted by a state in an arbitration treaty or in a special arbitral agreement. The model rules, however, have not yet been adopted in any arbitration arrangement between disputant governments, although in 1958 the General Assembly recommended the model rules for use by member-states when appropriate. It seems clear that states prefer flexibility in the resolution of their disputes by arranging the rules and proceedings of an arbitration according to circumstances.

Great impediments, indeed, still exist in the acceptance of international arbitration, especially in cases in which disputes between governments and foreign private parties are involved. In such cases, the state will often insist that its own local remedies—administrative and court proceedings—be exhausted. Generally, the government of the national who advances a claim against a foreign government will require evidence that the injured party has pursued all remedies in the foreign country before it presses a claim for international negotiation and adjudication, if, indeed, it decides to take up the case at all. Contracting parties may agree in their contract that they need not exhaust local remedies before resorting to arbitration, and one 1965 instrument, the Convention on the Settlement of Investment Disputes, states:

> Consent of the parties to arbitration under this Convention shall, unless otherwise stated, be deemed consent to such arbitration to the exclusion of any other remedy. A Contracting State may require the exhaustion of local administrative or judicial remedies as a condition of its consent to arbitration under this Convention.

The arbitration agreement in a general multilateral treaty, a bilateral convention, or in a specific contractual arrangement between two states often does not deal with the selection of the arbitrators and the appointment of an umpire, the procedure to be followed in the arbitration, the subject matter of the dispute, the specific issues to be submitted, the presentation of evidence, the place of the hearings, the law to be applied by the arbitrators, and the time when the award has to be rendered. These questions are usually dealt with in an agreement between the parties to the dispute known as compromis. If the compromis fails in some particular—to define the applicable law, for example—it is generally understood that the arbitrator shall apply international law.

An award rendered by an arbitral tribunal is customarily complied with by states: it is, in fact, invariably the case that unless a state is prepared to comply with an adverse decision, it will not submit the dispute to arbitration. The difficulties in the use of international arbitration thus consist less in the enforcement of arbitral awards than in persuading states involved in disputes to submit them to a third party, an arbitrator, or an arbitration tribunal.

(M.Do.)

BIBLIOGRAPHY. S. BEDFORD, *The Faces of Justice* (1961), discussion of how cases are handled in England, Germany, Austria, Switzerland, and France; J.H. MERRYMAN, *The Civil Law*

Tradition: An Introduction to the Legal Systems of Western Europe and Latin America (1969), a summary of the principles and institutions in civil-law countries; D. KARLEN *et al., Anglo-American Criminal Justice* (1967), comparison of two judicial systems in the area of criminal law; OLIVER WENDELL HOLMES, *The Common Law*, ed. by M. DEWOLFE HOWE (1963), classic treatment of the growth of law through the judicial decisions; B.N. CARDOZO, *The Nature of the Judicial Process* (1921), explanation by a distinguished judge of how an appellate court reaches its decisions; and ROSCOE POUND, *Organization of Courts* (1940), detailed treatment of court structure in the U.S. W.R. CORNISH, *The Jury* (1968), is a comprehensive British essay on the jury, combining traditional learning with new empirical material.

Overviews of legal institutions within specific countries include R.M. JACKSON, *The Machinery of Justice in England*, 5th ed. (1967); L. MAYERS, *The American Legal System*, rev. ed. (1964); H.J. BERMAN, *Justice in the U.S.S.R.*, rev. ed. (1963); M. CAPPELLETTI *et al., The Italian Legal System* (1967); and H.P. DUBEY, *A Short History of the Judicial Systems of India and Some Foreign Countries* (1968). See also JOHN W. JOHNSON, *American Legal Culture: 1908–1940* (1981).

The leading treatises and handbooks on commercial arbitration in English are FRANCIS RUSSELL, *On the Law of Arbitration*, 18th ed. by ANTHONY WALTON (1970); MARTIN DOMKE, *The Law and Practice of Commercial Arbitration* (1968); NRISINHADAS BASU, *The [Indian] Arbitration Act*, 5th ed. by S.K. BOSE (1965); and the INTERNATIONAL ASSOCIATION OF LAWYERS, *International Commercial Arbitration: A World Handbook*, ed. by PIETER SANDERS, 3 vol. (1956–65). Labour arbitration is treated in FRANK and EDNA ELKOURI, *How Arbitration Works*, rev. ed. (1960); and CLARENCE M. UPDEGRAFF and WHITLEY P. MCCOY, *Arbitration of Labor Disputes*, 2nd ed. (1961), are good accounts of the law and practice of labour arbitration. The historic development is described by E.E. WITTE, *Historical Survey of Labor Arbitration* (1952). Various aspects of international arbitration are dealt with by J.L. SIMPSON and HAZEL FOX, *International Arbitration: Law and Practice* (1959); in a *Report of a Study Group on the Pacific Settlement of International Disputes* (David Davies Memorial Institute 1966); in *International Arbitration: Liber Amicorum for Martin Domke*, ed. by PIETER SANDERS (1967); and in SAMUEL B. BACHARACH, *Bargaining: Power, Tactics, and Outcomes* (1981).

Kant and Kantianism

Immanuel Kant was the foremost thinker of the Enlightenment and one of the great philosophers of all time, in whom were subsumed new trends that had begun with the Rationalism (stressing reason) of René Descartes and the Empiricism (stressing experience) of Francis Bacon. He inaugurated a new era in the development of philosophical thought. His comprehensive and systematic work in theory of knowledge, ethics, and aesthetics greatly influenced all subsequent philosophy, especially the various German schools of Kantianism and Idealism.

This article deals with the man, his achievements, and the subsequent history of Kantianism. It is divided into the following sections:

Life and works

BACKGROUND AND EARLY YEARS

Kant was born on April 22, 1724, at Königsberg in East Prussia (since 1946 a part of the Soviet Union) and lived in that remote province for his entire life. His father, a saddler, was, according to Kant, a descendant of a Scottish immigrant, although scholars have found no basis for this claim; his mother, an uneducated German woman, was remarkable for her character and natural intelligence. Both parents were devoted followers of the Pietist branch of the Lutheran Church, which taught that religion belongs to the inner life expressed in simplicity and obedience to the moral law. The influence of their pastor made it possible for Kant—the fourth of nine children, but the eldest surviving child—to obtain an education.

Pietist rearing and schooling

At the age of eight Kant entered the Pietist school that his pastor directed. This was a Latin school, and it was presumably during the eight-and-a-half years he was there that Kant acquired his life-long love for the Latin classics, especially for the naturalistic poet Lucretius. In 1740 he enrolled in the University of Königsberg as a theological student. But, although he attended courses in theology and even preached on a few occasions, he was principally attracted to mathematics and physics. Aided by a young professor who had studied Christian Wolff, a systematizer of Rationalist philosophy, and who was also an enthusiast for the science of Sir Isaac Newton, Kant began reading the work of the English physicist and, in 1744, started his first book, dealing with a problem concerning kinetic forces. Though by that time he had decided to pursue an academic career, the death of his father in 1746 and his failure to obtain the post of undertutor in one of the schools attached to the university compelled him to withdraw and seek a means of supporting himself.

Tutor and Privatdozent. He found employment as a family tutor and, during the nine years that he gave to it, worked for three different families. With them he was introduced to the influential society of the city, acquired

Kant, pencil portrait by Hans Veit Schnoor von Carolsfeld (1764–1841). In the Kupferstichkabinett, Dresden, East Germany.

Marburg—Art Reference Bureau

social grace, and made his farthest travels from his native city—some 60 miles (96 kilometres) away to the town of Arnsdorf. In 1755, aided by the kindness of a friend, he was able to complete his degree at the university and take up the position of *Privatdozent,* or lecturer.

The three early dissertations

Three dissertations that he presented on obtaining this post indicate the interest and direction of his thought at this time. In one, *De Igne* (On Fire), he argued that bodies operate on one another through the medium of a uniformly diffused elastic and subtle matter that is the underlying substance of both heat and light. His first teaching was in mathematics and physics, and he was never to lose his interest in scientific developments. That it was more than an amateur interest is shown by his publication within the next few years of several scientific works dealing with the different races of men, the nature of winds, the causes of earthquakes, and the general theory of the heavens.

At this period Newtonian physics was important for Kant as much for its philosophical implications as for its scientific content. A second dissertation, the *Monodologia physica* (1756), contrasted the Newtonian methods of thinking with those employed in the philosophy then prevailing in German universities. This was the philosophy of Gottfried Wilhelm Leibniz, a universal scholar, as systematized and popularized by Wolff and by Alexander Gottlieb Baumgarten, author of a widely used text, the *Metaphysica* (1739). Leibniz' works as they are now known were not fully available to these writers; and the Leibnizian philosophy that they presented was extravagantly Rationalistic, abstract, and cut-and-dried. Yet it remained a powerful force, and the main efforts of independent thinkers in Germany at the time were devoted to examining its ideas.

In a third dissertation, *Principiorum Primorum Cognitionis Metaphysicae Nova Dilucidato* (1755), on the first principles of metaphysics, Kant analyzed especially the principle of sufficient reason, which, in Wolff's formulation, asserts that for everything there is a sufficient reason why it should be rather than not be. Although critical, Kant was cautious and still a long way from challenging the assumptions of Leibnizian metaphysics.

During his 15 years as a *Privatdozent,* Kant's fame as a teacher and writer steadily increased. Soon he was lecturing on many subjects besides physics and mathematics—including logic, metaphysics, and moral philosophy. He even lectured on fireworks and fortifications and every summer for 30 years gave a popular course on physical geography. He enjoyed great success as a lecturer; his style, which differed markedly from that of his books, was humorous and vivid, enlivened by many examples from his reading in English and French literature, and in travel and geography, science and philosophy.

Although he twice failed to obtain a professorship at Königsberg, he refused to accept offers that would have taken him elsewhere—including the professorship of poetry at Berlin that would have brought greater prestige. He preferred the peace and quiet of his native city in which to develop and mature his own philosophy.

Critic of Leibnizian Rationalism. During the 1760s he became increasingly critical of Leibnizianism. According to one of his students, Kant was then attacking Leibniz, Wolff, and Baumgarten, was a declared follower of Newton, and expressed great admiration for the moral philosophy of the Romanticist Jean-Jacques Rousseau.

The attack on Leibniz

His principal work of this period was *Untersuchung über die Deutlichkeit der Grundsätze der natürlichen Theologie und der Moral* (1764; "An Inquiry into the Distinctness of the Fundamental Principles of Natural Theology and Morals"). In this work he attacked the claim of Leibnizian philosophy that philosophy should model itself on mathematics and aim at constructing a chain of demonstrated truths based on self-evident premises. Kant argued that mathematics proceeds from definitions that are arbitrary, by means of operations that are clearly and sharply defined, upon concepts that can be exhibited in concrete form. In contrast with this method, he argued that philosophy must begin with concepts that are already given, "though confusedly or insufficiently determined," so that philosophers cannot begin with definitions without thereby shutting themselves up within a circle of words. Philosophy cannot, like mathematics, proceed synthetically; it must analyze and clarify. The importance of the moral order, which he had learned from Rousseau, reinforced the conviction received from his study of Newton that a synthetic philosophy is empty and false.

Besides attacking the methods of the Leibnizians, he also began criticizing their leading ideas. In an essay *Versuch, den Begriff der negativen Grössen in die Weltweisheit einzuführen* (1763), he argued that physical opposition as encountered in things cannot be reduced to logical contradiction, in which the same predicate is both affirmed and denied, and, hence, that it is pointless to reduce causality to the logical relation of antecedent and consequent. In an essay of the same year, *Der einzig mögliche Beweisgrund zu einer Demonstration des Daseyns Gottes,* he sharply criticized the Leibnizian concept of Being by charging that the so-called ontological argument, which would prove the existence of God by logic alone, is fallacious because it confuses existential with attributive statements: existence, he declared, is not a predicate of attribution. Moreover, with regard to the nature of space, Kant sided with Newton in his confrontation with Leibniz. Leibniz' view that space is "an order of co-existences" and that spatial differences can be stated in conceptual terms, he concluded to be untenable.

Investigation of the world of spirits

Some indication of a possible alternative of Kant's own to the Leibnizian position can be gathered from his curious *Träume eines Geistersehers erläutert durch Träume der Metaphysik* (1766). This work is an examination of the whole notion of a world of spirits, in the context of an inquiry into the spiritualist claims of Emanuel Swedenborg, a scientist and biblical scholar. Kant's position at first seems to have been completely skeptical, and the influence of the Scottish Skeptic David Hume is more apparent here than in any previous work; it was Hume, he later claimed, who first awoke him from his dogmatic slumbers. Yet Kant was not so much arguing that the notion of a world of spirits is illusory as insisting that men have no insight into the nature of such a world, a conclusion that has devastating implications for metaphysics as the Leibnizians conceived it. Metaphysicians can dream as well as spiritualists, but this is not to say that their dreams are necessarily empty; there are already hints that moral experience can give content to the ideal of an "intelligible world." Rousseau thus acted upon Kant here as a counterinfluence to Hume.

Early years of the professorship at Königsberg. Finally, in 1770, after serving for 15 years as a *Privatdozent,* Kant was appointed to the chair of logic and metaphysics, a position in which he remained active until a few years before his death. In this period—usually called his critical period, because in it he wrote his great *Critiques*—he published an astounding series of original works on a wide variety of topics, in which he elaborated and expounded his philosophy.

"Inaugural Dissertation" of 1770

The *Inaugural Dissertation* of 1770 that he delivered on assuming his new position already contained many of the important elements of his mature philosophy. As indicated in its title, *De Mundi Sensibilis atque Intelligibilis Forma et Principiis: Dissertatio,* the implicit dualism of the *Träume* is made explicit; and it is made so on the basis of a wholly un-Leibnizian interpretation of the distinction between sense and understanding. Sense is not, as Leibniz had supposed, a confused form of thinking but a source of knowledge in its own right, although the objects so known are still only "appearances"—the term that Leibniz also used. They are appearances because all sensing is conditioned by the presence, in sensibility, of the forms of time and space, which are not objective characteristics or frameworks of things but "pure intuitions." But though all knowledge of things sensible is thus of phenomena, it does not follow that nothing is known of things as they are in themselves. Certainly, man has no intuition, or direct insight, into an intelligible world; but the presence in him of certain "pure intellectual concepts, such as those of possibility, existence, necessity, substance, cause, enables him to have some descriptive knowledge of it. By means of these concepts he can arrive at an exemplar that provides him with "the common measure of all other things as far as real." This exemplar gives man an idea of perfection for both the theoretical and practical orders: in the first, it is that of the Supreme Being, God; in the latter, that of moral perfection.

After the *Dissertation,* Kant published virtually nothing for 11 years. Yet, in submitting the *Dissertation* to a friend at the time of its publication, he wrote:

> About a year since I attained that concept which I do not fear ever to be obliged to alter, though I may have to widen it, and by which all sorts of metaphysical questions can be tested in accordance with entirely safe and easy criteria, and a sure decision reached as to whether they are soluble or insoluble.

PERIOD OF THE THREE "CRITIQUES"

In 1781 the *Kritik der reinen Vernunft* (spelled "Critik" in the first edition; *Critique of Pure Reason*) was published, followed for the next nine years by great and original works that in a short time brought a revolution in philosophical thought and established the new direction in which it was to go in the years to come.

The Critique of Pure Reason. The *Critique of Pure Reason* was the result of some 10 years of thinking and meditation. Yet, even so, Kant published the first edition only reluctantly after many postponements; for although convinced of the truth of its doctrine, he was uncertain and doubtful about its exposition. His misgivings proved well-founded, and Kant complained that interpreters and critics of the work were badly misunderstanding it. To correct these wrong interpretations of his thought he wrote the *Prolegomena zu einer jeden künftigen Metaphysik die als Wissenschaft wird auftreten können* (1783) and brought out a second and revised edition of the first "critique" in 1787. Controversy still continues regarding the merits of the two editions: readers with a preference for an Idealistic interpretation usually prefer the first edition, whereas those with a Realistic view adhere to the second. But with regard to difficulty and ease of reading and understanding, it is generally agreed that there is little to choose between them. Anyone on first opening either book finds it overwhelmingly difficult and impenetrably obscure.

The cause for this difficulty can be traced in part to the works that Kant took as his models for philosophical writing. He was the first great modern philosopher to spend all of his time and efforts as a university professor of the subject. Regulations required that in all lecturing a certain set of books be used, with the result that all of Kant's teaching in philosophy had been based on such handbooks as those of Wolff and Baumgarten, which abounded in technical jargon, artificial and schematic divisions, and great claims to completeness. Following their example, Kant accordingly provided a highly artificial, rigid, and by no means immediately illuminating scaffolding for all three of his *Critiques.*

The *Critique of Pure Reason,* after an introduction, is divided into two parts, of very different lengths: A "Transcendental Doctrine of Elements," running to almost 400 pages in a typical edition, followed by a "Transcendental Doctrine of Method," which reaches scarcely 80 pages. The ". . . Elements" deals with the sources of human knowledge, whereas the ". . . Method" draws up a methodology for the use of "pure reason" and its a priori ideas. Both are "transcendental," in that they are presumed to analyze the roots of all knowledge and the conditions of all possible experience. The "Elements" is divided, in turn, into a "Transcendental Aesthetic," a "Transcendental Analytic," and a "Transcendental Dialectic."

The simplest way of describing the contents of the *Critique* is to say that it is a treatise about metaphysics: it seeks to show the impossibility of one sort of metaphysics and to lay the foundations for another. The Leibnizian metaphysics, the object of his attack, is criticized for assuming that the human mind can arrive, by pure thought, at truths about entities, which, by their very nature, can never be objects of experience, such as God, human freedom, and immortality. Kant maintained, however, that the mind has no such power and that the vaunted metaphysics is thus a sham.

As Kant saw it, the problem of metaphysics, as indeed of any science, is to explain how, on the one hand, its principles can be necessary and universal (such being a condition for any knowledge that is scientific) and yet, on the other hand, involve also a knowledge of the real and so provide the investigator with the possibility of more knowledge than is analytically contained in what he already knows; *i.e.,* than is implicit in the meaning alone. To meet these two conditions, Kant maintained, knowledge must rest on judgments that are a priori, for it is only as they are separate from the contingencies of experience that they could be necessary and yet also synthetic; *i.e.,* so that the predicate term contains something more than is analytically contained in the subject. Thus, for example, the proposition that all bodies are extended is not synthetic but analytic because the notion of extension is contained in the very notion of body; whereas the proposition that all bodies are heavy is synthetic because weight supposes, in addition to the notion of body, that of bodies in relation to one another. Hence, the basic problem, as Kant formulated it, is to determine "How [*i.e.,* under what conditions] are synthetic a priori judgments possible?"

The synthetic a priori in mathematics and physics

This problem arises, according to Kant, in three fields, viz., in mathematics, physics, and metaphysics; and the three main divisions of the first part of the *Critique* deal respectively with these. In the "Transcendental Aesthetic," Kant argued that mathematics necessarily deals with space and time and then claimed that these are both a priori forms of human sensibility that condition whatever is apprehended through the senses. In the "Transcendental Analytic," the most crucial as well as the most difficult part of the book, he maintained that physics is a priori and synthetic because in its ordering of experience it uses concepts of a special sort. These concepts—"categories," he called them—are not so much read out of experience as read into it and, hence, are a priori, or pure, as opposed to empirical. But they differ from empirical concepts in something more than their origin: their whole role in knowledge is different; for, whereas empirical concepts serve to correlate particular experiences and so to bring out in a detailed way how experience is ordered, the categories have the function of prescribing the general form that this detailed order must take. They belong, as it were, to the very framework of knowledge. But although they are indispensable for objective knowledge, the sole knowledge that they can give is of objects of possible experience; they yield valid and real knowledge only when they are ordering what is given through sense in space and time.

The synthetic a priori in metaphysics

In the "Transcendental Dialectic" Kant turned to consideration of a priori synthetic judgments in metaphysics. Here, he claimed, the situation is just the reverse from what it was in mathematics and physics. Metaphysics cuts itself off from sense experience in attempting to go beyond it and, for this very reason, fails to attain a single true a priori synthetic judgment. To justify this claim, Kant analyzed the use that metaphysics makes of the concept of the unconditioned. Reason, according to Kant, seeks for the unconditioned or absolute in three distinct spheres: (1) in philosophical psychology it seeks for an absolute subject of knowledge; (2) in the sphere of cosmology, it seeks for an absolute beginning of things in time, for an absolute limit to them in space, and for an absolute limit to their divisibility; and (3) in the sphere of theology, it seeks for an absolute condition for all things. In each case, Kant claimed to show that the attempt is doomed to failure by leading to an antinomy in which equally good reasons can be given for both the affirmative and the negative position. The metaphysical "sciences" of rational psychology, rational cosmology, and natural theology, familiar to Kant from the text of Baumgarten, on which he had to comment in his lectures, thus turn out to be without foundation.

With this work, Kant proudly asserted that he had accomplished a Copernican revolution in philosophy. Just as the founder of modern astronomy, Nicolaus Copernicus, had explained the apparent movements of the stars by ascribing them partly to the movement of the observers, so Kant had accounted for the application of the mind's a priori principles to objects by showing that the objects conform to the mind: in knowing, it is not the mind that conforms to things but things that conform to the mind.

The Critique of Practical Reason. Because of his insistence on the need for an empirical component in knowledge and his antipathy to speculative metaphysics, Kant is sometimes presented as a Positivist before his time; and his attack upon metaphysics was held by many in his own day to bring both religion and morality down with it. Such, however, was certainly far from Kant's intention. Not only did he propose to put metaphysics "on the sure path of science," he was prepared also to say that he "inevitably" believed in the existence of God and in a future life. It is also true that his original conception of his critical philosophy anticipated the preparation of a critique of moral philosophy. The *Kritik der praktischen Vernunft* (1788, spelled "Critik" and "practischen"; *Critique of Practical Reason*), the result of this intention, is the standard source book for his ethical doctrines. The earlier *Grundlegung zur Metaphysik der Sitten* (1785) is a shorter and, despite its title, more readily comprehensible treatment of the same general topic. Both differ from *Die Metaphysik der Sitten* (1797) in that they deal with pure ethics and try to elucidate basic principles; whereas the later work is concerned with applying what they establish in the concrete, a process that involved the consideration of virtues and vices and the foundations of law and politics.

Similarity between his ethics and epistemology

There are many points of similarity between Kant's ethics and his epistemology, or theory of knowledge. He used the same scaffolding for both—a "Doctrine of Elements," including an "Analytic" and a "Dialectic," followed by a "Methodology"; but the second *Critique* is far shorter and much less complicated. Just as the distinction between sense and intelligence was fundamental for the former, so is that between the inclinations and moral reason for the latter. And just as the nature of the human cognitive situation was elucidated in the first *Critique* by reference to the hypothetical notion of an intuitive understanding, so is that of the human moral situation clarified by reference to the notion of a "holy will." For a will of this kind there would be no distinction between reason and inclination; a being possessed of a holy will would always act as it ought. It would not, however, have the concepts of duty and moral obligation, which enter only when reason and desire find themselves opposed. In the case of human beings, the opposition is continuous, for man is at the same time both flesh and spirit; it is here that the influence of Kant's religious background is most prominent. Hence, the moral life is a continuing struggle in which morality appears to the potential delinquent in the form of a law that demands to be obeyed for its own sake—a law, however, the commands of which are not issued by some alien authority but represent the voice of reason, which the moral subject can recognize as his own.

In the "Dialectic," Kant took up again the ideas of God, freedom, and immortality. Dismissed in the first *Critique* as objects that men can never know because they transcend human sense experience, he now argued that they are essential postulates for the moral life. Though not reachable in metaphysics, they are absolutely essential for moral philosophy.

Kant is often described as an ethical Rationalist, and the description is not wholly inappropriate. He never espoused, however, the radical Rationalism of some of his contemporaries nor of more recent philosophers for whom reason is held to have direct insight into a world of values or the power to intuit the rightness of this or that moral principle. Thus, practical, like theoretical, reason was for him formal rather than material—a framework of formative principles rather than a content of actual rules. This is why he put such stress on his first formulation of the categorical imperative: "Act only on that maxim through which you can at the same time will that it should become a universal law." Lacking any insight into the moral realm, men can only ask themselves whether what they are proposing to do has the formal character of law—the character, namely, of being the same for all persons similarly circumstanced.

The Critique of Judgment. The *Kritik der Urteilskraft* (1790: spelled "Critik")—one of the most original and instructive of all of Kant's writings—was not foreseen in his original conception of the critical philosophy. Thus it is perhaps best regarded as a series of appendixes to the other two *Critiques*. The work falls into two main parts, called respectively "Critique of Aesthetic Judgment" and "Critique of Teleological Judgment." In the first of these, after an introduction in which he discussed "logical purposiveness," he analyzed the notion of "aesthetic purposiveness" in judgments that ascribe beauty to something. Such a judgment, according to him, unlike a mere expression of taste, lays claim to general validity; yet it cannot be said to be cognitive because it rests on feeling, not on argument. The explanation lies in the fact that, when a person contemplates an object and finds it beautiful, there is a certain harmony between his imagination and his understanding, of which he is aware from the immediate delight that he takes in the object. Imagination grasps the object and yet is not restricted to any definite concept; whereas a person imputes the delight that he feels to others because it springs from the free play of his cognitive faculties, which are the same in all men.

Critiques of aesthetics and natural teleology

In the second part, Kant turned to consider teleology in nature as it is posed by the existence in organic bodies of things of which the parts are reciprocally means and ends to each other. In dealing with these bodies, one cannot be content with merely mechanical principles. Yet if mechanism is abandoned and the notion of a purpose or end of nature is taken literally, this seems to imply that the things to which it applies must be the work of some supernatural designer; but this would mean a passing from the sensible to the suprasensible, a step proved in the first *Critique* to be impossible. Kant answered this objection by admitting that teleological language cannot be avoided in taking account of natural phenomena; but it must be understood as meaning only that organisms must be thought of "as if" they were the product of design, and that is by no means the same as saying that they are deliberately produced.

LAST YEARS

The critical philosophy was soon being taught in every important German-speaking university, and young men flocked to Königsberg as a shrine of philosophy. In some cases, the Prussian government even undertook the expense of their support. Kant came to be consulted as an oracle on all kinds of questions, including such subjects as the lawfulness of vaccination. Such homage did not interrupt Kant's regular habits. Scarcely five feet tall, with a deformed chest, and suffering from weak health, he maintained throughout his life a severe regimen. It was arranged with such regularity that people set their clocks according to his daily walk along the street named for him, "The Philosopher's Walk." Until old age prevented him, he is said to have missed this regular appearance only on the occasion when Rousseau's *Émile* so engrossed him that for several days he stayed at home.

Works on religion and the ban on them

With the publication of the third *Critique,* Kant's main philosophical work was done. From 1790 his health began to decline seriously. He still had many literary projects but found it impossible to write more than a few hours a day. The writings that he then completed consist partly of an elaboration of subjects not previously treated in any detail, partly of replies to criticisms and to the clarification of misunderstandings. With the publication in 1793 of his work *Die Religion innerhalb der Grenzen der blossen Vernunft,* Kant became involved in a dispute with Prussian authorities on the right to express religious opinions. The book was found to be altogether too Rationalistic for orthordox taste; he was charged with misusing his philosophy to the "distortion and depreciation of many leading and fundamental doctrines of sacred Scripture and Christianity" and was required by the government not to lecture or write anything further on religious subjects. Kant agreed but privately interpreted the ban as a personal promise to the King, from which he felt himself to be released on the latter's death in 1797. At any rate, he returned to the forbidden subject in his last major essay, *Der Streit der Fakultäten* (1798; "The Conflict of the Faculties").

The large work at which he laboured until his death—the fragments of which fill the two final volumes of the great Berlin edition of his works—was evidently intended to be a major contribution to his critical philosophy. What remains, however, is not so much an unfinished work as a series of notes for a work that was never written. Its original title was *Übergang von den metaphysische Anfangsgründe der Naturwissenschaft zur Physik* ("Transition from the Metaphysical Foundations of Natural Science to Physics"), and it may have been his intention to carry further the argument advanced in the *Metaphysische Anfangsgründe der Naturwissenschaft* (1786) by showing that it is possible to construct a priori not merely the general outline of a science of nature but a good many of its details as well. But judging from the extant fragments, however numerous they are, it remains conjectural whether its completion would have constituted a major addition to his philosophy and its reputation.

After a gradual decline that was painful to his friends as well as to himself, Kant died in Königsberg, February 12, 1804. His last words were "Es ist gut" ("It is good"). His tomb in the cathedral was inscribed with the words (in German) "The starry heavens above me and the moral law within me," the two things that he declared in the conclusion of the second *Critique* "fill the mind with ever new and increasing admiration and awe, the oftener and the more steadily we reflect on." (O.A.B.)

Kantianism

As a philosophical designation, Kantianism can signify either the system of thought contained in the writings of the epoch-making 18th-century philosopher Immanuel Kant or those later philosophies that arose from the study of Kant's writings and drew their inspiration from his principles. Only the latter is the concern of this section.

NATURE AND TYPES OF KANTIANISM

Diverse trends in Kantianism

The Kantian movement comprises a loose assemblage of rather diverse philosophies that share Kant's concern with exploring the nature, and especially the limits, of human knowledge in the hope of raising philosophy to the level of a science in some sense similar to mathematics and physics. Participating in the critical spirit and method of Kant, these philosophies are thus opposed to dogmatism, to expansive speculative naturalism (such as that of Baruch Spinoza, the Jewish Rationalist), and, usually, to irrationalism. The various submovements of Kantianism are characterized by their sharing of certain "family resemblances"; *i.e.,* by the preoccupation of each with its own selection of concerns from among the many developments of Kant's philosophy: a concern, for example, with the nature of empirical knowledge; with the way in which the mind imposes its own categorial structure upon experience, and, in particular, with the nature of the structure that renders man's knowledge and moral action possible, a structure considered to be a priori (logically independent of experience); with the status of the *Ding an sich* ("thing-in-itself"), that more ultimate reality that presumably lurks behind man's apprehension of an object; or with the relationship between knowledge and morality. A brief exposition of Kant's philosophical system may be found above.

A system such as the critical philosophy of Kant freely lends itself to reconstructions of its synthesis according to whatever preferences the private philosophical inclinations of the reader may impose or suggest. Kant's system was a syncretism, or union, of British Empiricism (as in John Locke, George Berkeley, and David Hume) that stressed the role of experience in the rise of knowledge; of the scientific methodology of Isaac Newton; and of the metaphysical apriorism (or Rationalism) of Christian Wolff, who systematized the philosophy of Gottfried Leibniz, with its emphasis on mind. Thus it constituted a synthesis of elements very different in origin and nature, which tempted the student to read his own presuppositions into it.

Varieties of Kantianism

The critical philosophy has been subjected to a variety of approaches and methods of interpretation. These can be reduced to three fundamental types: those that conceive of the critical philosophy as an epistemology or a pure theory of (scientific) knowledge and methodology; those that conceive of it as a critical theory of metaphysics or the nature of Being (ultimate reality); and those that conceive of it as a theory of normative or valuational reflection parallel to that of ethics (in the field of action). Each of these types—known, respectively, as epistemological, metaphysical, and axiological Kantianism—can, in turn, be subdivided into several secondary approaches. Historically, epistemological Kantianism included such different attitudes as empirical Kantianism, rooted either in physiological or psychological inquiries; the logistic Kantianism of the Marburg school, which stressed essences and the use of logic; and the realistic Kantianism of the Austrian Alois Riehl. Metaphysical Kantianism developed from the transcendental Idealism of German Romanticism to Realism, a course followed by many speculative thinkers, who—like nearly all contemporary Kantians—saw in the critical philosophy the foundations of an essentially inductive metaphysics, in accordance with the results of the modern sciences. Finally, axiological Kantianism—concerned with value theory—branched, first, into an axiological approach (properly so-called), which interpreted the methods of all three of Kant's *Critiques* (*i.e., Critique of Pure Reason, Critique of Practical Reason,* and *Critique of Judgment*) as normative disciplines of thought; and, second, into an eclectic or relativistic Kantianism, which regarded the critical philosophy as a system of thought dependent upon social, cultural, and historical conditions.

The chief representatives of these submovements are identified in the historical sections below.

It is essential to distinguish clearly between two periods within the Kantian movement: first, the period from 1790 to 1831 (the death of Hegel); and, second, the period from 1860 to the present—separated by a time when an antiphilosophical Positivism, a type of thought that supplanted metaphysics with science, was predominant. The first period began with the thorough study and emendation of Kant's chief theoretical work, *Kritik der reinen Vernunft* (2nd ed., 1787; the *Critique of Pure Reason,* 1929); but it soon became intermingled with the romantic tendencies in German Idealism. The second period, called specifically Neo-Kantianism, was, first of all, a conscious reappraisal, in whole or in part, of the theoretical *Critique,* but also, as a total system, a reaction against Positivism. Earlier Neo-Kantianism reduced philosophy to the theory of knowledge and scientific methodology; systematic Neo-Kantianism, arising at the beginning of the 20th century, expressed itself in attempts at building metaphysical structures.

EARLY KANTIANISM: 1790–1835

Kant himself

According to Immanuel Kant, his major work, the *Critique of Pure Reason,* comprised a treatise on methodology, a preliminary investigation prerequisite to the study of science, which placed the Newtonian method (induction, inference, and generalization) over against that of Descartes and Wolff (deduction from intuitions asserted to be self-evident). The result was a critique of metaphysics, the value of which lay not in science but in a realm of being accessible only to the pure intellect. In exploring this "noumenal" realm, as he called it, Kant placed his *Critique* in a positive role. Recalling the revolution that occurred in astronomy when Nicolaus Copernicus discerned, in the apparent motions of the planets, reflections of the earth's own motion, Kant inaugurated a Copernican revolution in philosophy, which claimed that the subject doing the knowing constitutes, to a considerable extent, the object; *i.e.,* that knowledge is in part constituted by a priori or transcendental factors (contributed by the mind itself), which the mind imposes upon the data of experience. Far from being a description of an external reality, knowledge is, to Kant, the product of the knowing subject. When the data are those of sense experience, the transcendental (mental) apparatus constitutes man's experience or his science, or makes it to be such. These transcendental elements are of three different orders: at the lowest level are the forms of space and time (technically called intuitions); above these are the categories and principles of man's intelligence (among them substance, causality, and necessity); and at the uppermost level of abstraction the ideas of reason—the transcendental "I," the world as a whole, and God. It is by virtue of the encounter between the forms of man's sensory intuition (space and time) and his perceptions that phenomena are formed. The forms arise from the subject himself; the perceptions, however—or the data of experience—have reference, ultimately, to things-in-themselves, which nevertheless remain unknowable, inasmuch as, in order to be known at all, it is necessary for things to appear clothed, as it were, in the forms of man's intuition and, thenceforth, to present themselves as phenomena and not as noumena. The thing-in-itself, accordingly, indicates the limit and not the object of knowledge.

Early criticism

These theses of Kant provoked criticism among the followers of Christian Wolff, the Leibnizian Rationalist, and doubts among the disciples of Kant, which, as they further developed into systems, marked the first period of Kantianism. Inasmuch as these disciples took the *Critique of Pure Reason* to be a *preface* to the study of the pure reason or of the transcendental system and not the system itself, they saw in this interpretation an explanation for the ambiguities to which the *Critique* (as they felt) was subject. Their doubts revolved around two points: first, Kant had erroneously distinguished three kinds of a priori knowledge, coordinate with the three aforementioned levels or faculties of the mind; and second, Kant had accepted the thing-in-itself as constitutive of knowledge. Regarding the first point, they claimed that Kant had accepted the three faculties and their respective transcendental characteristics without investigation, in which case this structure should be viewed, in accordance with the preliminary character of the *Critique,* as a triple manifestation of a single fundamental faculty. For this reason the distinction between the levels of intuition and understanding (or between the receptivity and spontaneity of the mind) had to be rejected—for the three transcendentals—space and time, the categories, and the ideas of reason—were not existents but were only functions of thought. Finally, these disciples argued that the existence of a single transcendental subject, the Ego, would render the thing-in-itself superfluous and even pernicious for the scientific treatment of epistemology.

This function of human thought (the transcendental subject), which serves as the absolute source of the a priori, was variously designated by different early Kantian thinkers: for the German Realist Karl L. Reinhold, it constituted the faculty of representation; for the Lithuanian Idealist Salomon Maimon, it was a mental capacity for constructing objects; for the Idealist Jakob S. Beck, a protégé of Kant's, it was the act of synthesis; for the empirical critic of Kantianism G.E. Schulze, it was experience in the sense intended by Hume—a volley of discrete sense impressions; for the theory of knowledge of the outstanding ethical Idealist Johann G. Fichte, it was the original positing of the Ego and the non-Ego—which meant, in turn, in the case of the aesthetic Idealist F.W.J. von Schelling, "the absolute self," and in the case of G.W.F. Hegel "the *Geist* or absolute Spirit," and finally, in the case of the pessimistic Romanticist Arthur Schopenhauer, "the absolute Will." In each case (excepting Schulze) the interpretation of the thing-in-itself in a realistic metaphysical sense was rejected in favour of various degrees of transcendental Idealism. Removed from the main current of Kantianism was the empirically oriented thinker Jakob Friedrich Fries (the one figure in this group who was not an Idealist in the true sense), who interpreted the a priori in terms of psychological faculties and elements.

Having earlier renounced these apostates on a large scale, Kant, at the end of his life, prepared a new exposition of the transcendental philosophy (the second part of his *Opus Postumum*), which showed that he was ready tacitly to accede to the criticisms of his adversaries.

NEO-KANTIANISM: SINCE 1860

General features of Neo-Kantianism

Nineteenth-century Neo-Kantianism. The rejection of all of philosophy by Positivism had the anomalous effect of, itself, evoking an awakening of Kantianism, for many thinkers wished to give to Positivism itself a philosophical foundation that, while respecting the phenomenological attitude, would yet be hostile to the metaphysics of Positivism, which was usually a tacit, but inconsequent, Materialism. It was justifiably held that Kant could provide such a foundation because of his opposition to metaphysics and his limitation of scientific knowledge to the sphere of phenomena. The complexity of the critical philosophy was such that the theoretical criticism could be approached in diverse ways and that, through the facts themselves, diverse interpretations of the *Critique of Pure Reason* could be obtained. In the order of their origin (though not of their worth or importance), there thus arose currents of Kantianism that were empiricist, logicist, realist, metaphysical, axiological, and psychological—of which the most important have survived into the 20th century.

The return to Kant was determined by the historical fresco of the incomparable historian of philosophy Kuno Fischer entitled *Kants Leben und die Grundlagen seiner Lehre* (1860; "Kant's Life and the Foundations of his Teaching"), which replaced the earlier work of the semi-Kantian Ernst Reinhold, son of the more notable Jena scholar (published 1828–30), and especially that of the outstanding historian of philosophy Johann Eduard Erdmann (published 1834–53). In 1865 the order: "Zurück nach Kant!" ("Back to Kant!") reverberated through the celebrated work of the young epistemologist Otto Liebmann, *Kant und die Epigonen* ("Kant and his Followers"), which was destined to extricate their spirits from the Positivistic morass and, at the same time, to divert the Germans from romantic Idealism.

Schools of Neo-Kantianism

Epistemological Neo-Kantianism. The empiricist, logistic, and realistic schools can be classed as epistemological.

Empiricist Neo-Kantianism was represented by the erudite pioneering physicist and physiologist Hermann L.F. von Helmholtz and, in part, by F.A. Lange, author of a famous study of Materialism. Helmholtz found support in Kant for his claim, first, that, although perception can *represent* an external thing, it usually does so in a way far removed from an actual description of its properties; second, that space and time comprise an empirical framework created for thought by the perceiving subject; and, third, that causality is an a priori law allowing the philosopher to infer a reality that is absolutely unknowable. Similarly, Lange reduced science to the phenomenal level and repudiated the thing-in-itself.

Logistic Neo-Kantianism, as represented in the most well-known and flourishing school of Kantianism, that at Marburg, originated with Hermann Cohen, successor of Lange at Marburg, who, in a book on Kant (1871), argued that the transcendental subject is not to be regarded as a psychic being but as a logical function of thought that constructs both the form *and* the content of knowledge. Nothing is *gegeben* ("given"), he urged; all is *aufgegeben* ("propounded," like a riddle) to thought—as when, in the infinitesimal calculus, the analyst generates motion by imagining thin slices of space and time and adding up their areas. Hence experience is a perfect construction of man's logical spirit. The example of Cohen inspired many other authors, among them Cohen's colleague at Marburg Paul Natorp, who, in his work on the logical foundations of the exact sciences, integrated even psychology into the Marburgian transcendentalism; and Ernst Cassirer, best known for stressing the symbolizing capacities of man, who, in his memorable work *Das Erkenntnisproblem in der Philosophie und Wissenschaft der neueren Zeit* (1906–20; *The Problem of Knowledge: Philosophy, Science, and History since Hegel,* 1966), transposed this same logisticism into a form that illumines the history of modern philosophy.

Realistic Neo-Kantianism, the third manifestation of epistemological Neo-Kantianism, was represented in the Realism of the scientific monist Alois Riehl and of his disciple Richard Hönigswald. In a work on the significance of the critical philosophy for the positive sciences (published 1876–87), Riehl held, in direct opposition to the Marburgian logisticism, that the thing-in-itself participates positively in the constitution of knowledge inasmuch as all perception includes a reference to things outside the subject.

Metaphysical Neo-Kantianism. Ten years after the appearance of the aforementioned ground-breaking book *Kant und die Epigonen,* its author, Otto Liebmann, introduced the new metaphysical approach in his book on the analysis of reality (1876), which came near to the Kantianism of Marburg. The Romanticist Johannes Volkelt, in turn, took up the theme of a critical metaphysics and expressed his persisting aspirations toward the Absolute in the claim that, beyond the certainties of man's own subjective consciousness, there exists a new kind of certainty in a transsubjective realm. Subjectivity is, thus, inevitably transcended, just as the sciences are surmounted when they presuppose a metaphysics. The influential spiritual moralist Friedrich Paulsen defended the claim that Kant had always behaved as a metaphysician, even in the *Critique of Pure Reason,* in spite of the epistemological restrictions that he imposed upon himself—a claim that made an impact that was felt throughout the following century.

Axiological Neo-Kantianism. Inasmuch as the two principal representatives of the axiological interpretation both taught at Heidelberg, this branch is also known as the Southwest German or Baden school. Its initiator was Wilhelm Windelband, esteemed for his "problems" approach to the history of philosophy. The scholar who systematized this position was his successor Heinrich Rickert, who had come from the tradition of Kuno Fischer. Drawing a parallel between the constraints that logic exerts upon thought and those that the sense of ought exerts upon ethical action, these thinkers argued that, while man's *action* must answer to an absolute value (the Good), his *thought* must answer to a regulative value (the True), which imposes upon him the duty of conforming to it. The *Critique of Pure Reason,* they held, elaborates this rule—which is not an entity but an imperative, or absolute, charge to act. Rickert regarded the critical endeavour as having been too narrow, since it was suited merely to physics. Actually, he charged, it should be the foundation for all of the sciences of the spirit. The distinctive characteristic of this school thus consisted in reintegrating German Idealism (as in Fichte and Hegel) into a rather personal Kantianism. Consequently, it succeeded in annexing more than one area of semi-Kantian thought: *e.g.,* "the philosophy of the spiritual sciences" of Wilhelm Dilthey, who held that intellectual life cannot be explained by means of naturalistic causality but only through historical understanding (*Verstehen*); "the life-philosophy" of the social philosopher Georg Simmel, who deviated from an earlier naturalistic relativism to the espousal of objective values; "the philosophy of value" of the experimental psychologist Hugo Münsterberg, author of one of the earliest systems of values; the "semi-Hegelianism" of Richard Kroner, a philosopher of culture and religion; and the general works of Bruno Bauch, Liebmann's successor at Jena. All of these philosophers were more or less related to axiological Neo-Kantianism.

Psychological Neo-Kantianism. An initial attempt to interpret Kantian transcendentalism in psychological terms was made by the Friesian Empiricist Jürgen Bona Meyer in his *Kants Psychologie* (1870). Later, a more important contribution in this field was made by the Göttingen philosopher of ethics and law Leonard Nelson and published in the *Abhandlungen der Fries'schen Schule* (1904 ff; "Acts of the Friesian School"). Even this title suggests an intimate agreement with the Kantianism of Fries's new critique of reason (1807); and Nelson, indeed, is regarded as the founder of the Neo-Friesian school. At a time when other Kantian schools were concerned with the transcendental analysis of objective or outer knowledge, Nelson held that, in the analysis of the subjective or inner self, the transcendental equipment of the mind—the a priori—is directly revealed. It thus fell to psychology to lay bare this equipment, which belongs in itself to the metaphysical order. It was upon this basis that the Marburg theologian Rudolf Otto, in his book *Das Heilige* (1917; *The Idea of the Holy,* 1958), ventured a type of religious phenomenology that has proved very successful.

Kantian philology

A discipline known as the Kant *Philologie,* concerned with the history, development, and works of Kant, has preempted a considerable portion of philosophical historiography since 1860. These studies began with the immense commentary on the *Critique of Pure Reason* produced in 1881–92 by Hans Vaihinger, known for his philosophy of the "As If" (which stresses man's reliance on pragmatic fictions), and with the founding of the new journal *Kantstudien* (1896) and the Kant-Gesellschaft ("Kantian Society," 1904)—both still extant. The most conspicuous result of this philological movement, however, was undeniably the monumental edition, in 23 volumes, of all of Kant's available works prepared (1900 ff) by the Academy of Sciences at Berlin under the editorship of the champion of humanistic studies, Wilhelm Dilthey. These volumes include: Sect. 1, Works; Sect. 2, Correspondence; Sect. 3, The "Nachlass." Since the transfer of this task to the University of Münster, Sect. 4, Kant's Lectures, has been undertaken. Those on logic and metaphysics (vols. 24–25) have been splendidly edited by Gerhard Lehmann.

Contemporary Neo-Kantianism. The recent development of Neo-Kantianism, except for innumerable historical studies, is very one-sided: no longer considered as exclusively epistemological, it merely prolongs the metaphysical school. Moreover, a large portion of the present Kant research is covered by the so-called *Problems of Kantianism* (see below). Important studies have been made on the development of Kant's philosophical thought, on Kant as a metaphysician, on his ontology and teachings on science, and on his transcendental deduction.

NON-GERMAN KANTIANISM

The Kantian awakening, in no wise limited to Germany, extended throughout Western philosophy. Its principal initiators were as follows: France was the first to open to its influence, beginning with the eclectic thinker Victor Cousin, who had studied German authors and made several trips to Germany. The relativistic personalist Charles Renouvier then defended a rather personal critical philosophy, which exerted an enduring influence through its impact upon the extreme Idealist Octave Hamelin of the Sorbonne; upon the metaphysician and cofounder of French neospiritualism Jules Lachelier; and upon his pupil, the philosopher of science Émile Boutroux.

English Kantianism

The English-speaking countries, on the other hand, have not seemed disposed to assimilate the critical philosophy as they did Hegelian Idealism. Except for the Scottish religious absolutist Edward Caird (*The Critical Philosophy of Immanuel Kant,* 1889), who was chiefly an Hegelian, there was in Britain at the close of the 19th century only another Scot, the critical Realist Robert Adamson, who was a Kantian. After him, however, can be cited the commentary, published in 1918, of Norman Kemp Smith, producer of the standard English translation of Kant's first *Critique,* and more recently, the remarkable exposition by the Oxford Kantian Herbert J. Paton, *Kant's Metaphysic of Experience* (2 volumes, 1936). Finally, Kantian methods can be discerned today in the later work of the prominent Oxford "ordinary language" philosopher, Peter F. Strawson, entitled *Individuals: An Essay in Descriptive Metaphysics* (1959). Kantianism became known in the United States toward 1840 primarily through the New England transcendentalist and poet Ralph Waldo Emerson—who was not, however, a Kantian himself. The physicist and logician Charles Sanders Peirce owes his Pragmatism largely to Kant's role as a counterweight against Hegelianism. The former southern California philosopher William H. Werkmeister represents a type of Neo-Kantianism inspired by the Marburg school (*The Basis and Structure of Knowledge,* 1948).

Italian scholars, on the other hand, have been vigorously engaged in Kantian studies since the initiative was taken by Alfonso Testa. The chief Neo-Kantian in Italy, however, was the Realist Carlo Cantoni, who took an anti-Positivist stance. Later, in the period from 1900 to 1918, Kantianism was represented by the extreme Realism of the theist Francesco Orestano. A school of Kantian philology has formed at Turin around the erudite Christian Idealist Augusto Guzzo and his journal *Filosofia.* More independent in spirit is the work of the critical ontologist Pantaleo Carabellese, Giovanni Gentile's successor at Rome.

ASSESSMENT OF KANTIANISM

At the present time there does not exist, either in Germany or elsewhere, a purely Kantian philosopher; but all acknowledge the obligation to come to grips with him. Within the four great currents of contemporary thought, however—*i.e.,* in Phenomenology, in the traditionalistic metaphysics, in Existentialism, and in the Positivistic Empiricism of the Vienna Circle and of Analytical philosophy—the predominant attitude toward Kant is negative.

Empiricist criticism

Problems of Kantianism. As far as epistemology is concerned, the critical philosophy constitutes a theory of science that agrees with current trends; for science must have a base that is empirical though also real. On the other hand, the transcendental or a priori is implicated; and severe complications ensue whenever the question is posed whether a type of apprehension can be acquired apart from experience that conveys, however, some new and genuine knowledge—whether, in short, synthetic a priori judgments can be made. Significantly, the founder of Phenomenology, the German philosopher Edmund Husserl, came back to the fold of Kantian transcendentalism after previously opposing it bitterly. As against the Kantian position, Empiricism entirely rejects the possibility (and even the meaning) of the synthetic a priori and, robbed thereby of everything traditionally regarded as the subject matter of philosophy, directs its philosophical inquiries principally to the study of language. The foremost recent analyst of language, however, the pioneering philosopher Ludwig Wittgenstein, imposed upon philosophy the obligation to limit reason (or the transcendental element in knowledge)—a semi-Kantian position, which he nonetheless later renounced. As for Existentialism, one of recent Germany's foremost philosophers, Martin Heidegger, has presented in his *Kant und das Problem der Metaphysik* (1929; Eng. trans., *Kant and the Problem of Metaphysics,* 1962) a highly personalized interpretation. A student of Cohen at Marburg, the metaphysician Nicolai Hartmann, became the harbinger of the Realistic approach, elaborating in his analysis of the metaphysics of knowledge (1921) an ontological relation that he discerned to obtain between two forms of being: between thought and reality. Accordingly, the principles of thought correspond, in his view, to those of reality—a position at odds with Kant (even when he is interpreted as a Realist). Moreover, Hartmann treated the problems of mathematics—so urgent in current philosophy—in a manner that is again completely opposed to Kant; in particular, he questioned the validity of Kant's a priori intuition (or positing) of the spatiotemporal framework in terms of which man thinks about the world, challenging Kant at this point not merely to accommodate the non-Euclidean geometries (with curved space) that afforded a Realist alternative to the a priori but above all to reflect the distinctly logistic position regarding the foundations of mathematics to which he adhered. Discussion of the status of the thing-in-itself in man's knowledge of the real remained on the philosophical agenda both during and after Hartmann's time, but invoked the same indecision as it always had. At a time when Hartmann was accepting the thing-in-itself almost naïvely, Empiricism (in all its forms) rejected it categorically and attempted to construe the real in terms merely of what Kant had called phenomena. In the realm of ethics, Phenomenologists and Existentialists were dissatisfied with the purely formal character of Kant's ethics—*i.e.,* with its lack of specificity—and substituted a "material" ethic, of concrete duties, which was no less absolute than that of Kant. For their part, Empiricists were only interested in the analysis of expressions of moral judgment, which they reduced to imperative statements that are emotive and aimed at winning adherents.

Objections to Kantianism. It must be acknowledged that Kant has furnished many of the most significant themes that are found in the currents of contemporary philosophy, even in the forms that they still assume today. Yet, as compared with the state of affairs that existed from 1860 to 1918, Kantianism has suffered an impressive decline—though a slight recovery seems to have occurred during the third quarter of the 20th century.

Scientism, formalism, subjectivism

What were the reasons for this decline? In general, since World War I the reduction of philosophy to the philosophy of science has no longer been accepted, though contemporary Positivistic Empiricism has offered hardly any objection to it. The philosophy of science comprises, in fact, only *one* problem area, not the entire assemblage of philosophical problems. From this a second objection arises: Kantianism in general is too formalistic to satisfy man's actual inquisitiveness, which inclines more and more toward concrete concerns. Kantianism restricts itself to examining the a priori forms of thought and cares little for its diverse contents. Were this objection pertinent only to the exact sciences, it would not be serious—for these sciences attend to their own applications; but the objection becomes very grave for the field of ethics. For this reason, the objection against Kant's formalism has been raised most passionately against his ethical treatise, the *Critique of Practical Reason*—as by Hartmann, by the Phenomenologist Max Scheler, and by others. This transcendental formalism immediately encounters the further objection of subjectivism—in spite of efforts (from the side of logic) to evade it—*i.e.,* it is blamed for obstructing man's apprehension of the real universality of his Ego, of the thinking subject, and for inexorably impelling the scholar to the view that man's knowledge is merely the product of subjective construction. This subjectivistic transcendentalism, by its intrinsic logic, denies man access to the external world. Not only does it debar him from the world of things-in-themselves but it also prevents him from granting objective reality to phenomena as such, inasmuch as the transcendental source is here viewed as playing a constructive role with respect to experience and the phenomenon.

These three major objections, which stand out in the midst of many criticisms of minor details, recur constantly in the Kantian literature of the past quarter of a century. The result of these objections, as far as the evaluation of the critical philosophy is concerned, is that it is repudiated in its entirety—without, however, being thereby considered barred by limitation. Kant thus remains, in spite of everything, an inexhaustible source of problems and ideas, comparable in this respect to Plato and Aristotle, with whom he forms the great triad of Western philosophical thought. (H.J. de V.)

MAJOR WORKS

PRE-CRITICAL WRITINGS: *Gedanken von der wahren Schätzung der lebendigen Kräfte und Beurteilung der Beweise derer sich Herr von Leibniz und anderer Mechaniker in dieser Streitsache bedient haben* (1746); *Allgemeine Naturgeschichte und Theorie des Himmels* (1755; *Kant's Cosmogony . . . ,* 1900 and 1968; *Universal Natural History and Theories of the Heavens,* 1969); *Principiorum Primorum Cognitionis Metaphysicae Nova Dilucidatio* (1755; Eng. trans. by F.E. England in *Kant's Conception of God,* 1929); *Metaphysicae cum geometria iunctae usus in philosophia naturali, cuius specimen I. continet Monadologiam physicam* (1756); *Versuch einiger Betrachtungen über den Optimismus* (1759); *Die falsche Spitzfindigkeit der vier syllogistischen Figurerewiesen* (1762; trans. in *Kant's Introduction to Logic and His Essay on the Mistaken Subtilty of the Four Figures,* 1963); *Der einzige mögliche Beweisgrund zu einer Demonstration des Daseyns Gottes* (1763; *Enquiry into the Proofs for the Existence of God,* 1836); *Versuch, den Begriff der negativen Grössen in die Weltweisheit einzuführen* (1763; *An Attempt to Introduce the Conception of Negative Quantities into Philosophy,* 1911); *Untersuchung über die Deutlichkeit der Grundsätze der natürlichen Theologie und der Moral* (1764); *Beobachtungen über das Gefühl des Schönen und Erhabenen* (1764, 1766, 1771; *Observations on the Feeling of the Beautiful and Sublime,* 1960); *Träume eines Geistersehers erläutert durch Träume der Metaphysik* (1766; *Dreams of a Spirit-Seer, Illustrated by Dreams of Metaphysics,* 1900; *Dreams of a Spirit Seer, and Other Related Writings,* 1969); *De Mundi Sensibilis atque Intelligibilis Forma et Principiis: Dissertatio* (1770; *Kant's Inaugural Dissertation and Early Writings on Space,* 1929); *Von den Verschiedenen Racen der Menschen* (1775).

CRITICAL AND POST-CRITICAL WRITINGS: *Critik der reinen Vernunft* (1781; rev. ed., *Kritik der reinen Vernunft,* 1787; *Critique of Pure Reason,* 1929, 1950); *Prolegomena zur einer jeden künftigen Metaphysik die als Wissenschaft wird auftreten können* (1783; *Prolegomena to Any Future Metaphysics,* 1951); *Grundlegung zur Metaphysik der Sitten* (1785; *The Fundamental Principles of the Metaphysic of Ethics,* 1938; *The Moral Law; or, Kant's Groundwork of the Metaphysic of Morals,* 1948; *Foundations of the Metaphysics of Morals,* 1969); *Metaphysische Anfangsgründe der Naturwissenschaft* (1786; *Metaphysical Foundations of Natural Science,* 1970); *Critik der practischen Vernunft* (1788; *Critique of Practical Reason,* 1949); *Critik der*

Urteilskraft (1790, 2nd ed. 1793; *Kant's Kritik of Judgment,* 1892, reprinted as *Kant's Critique . . .,* 1914; new version, *Critique . . .,* vol. 1, *Kant's Critique of Aesthetic Judgment* and vol. 2, *Critique of Teleological Judgment,* 1911–28, republished 1952); *Über eine Entdeckung, nach der alle neu Critik der reinen Vernunft durch eine ältere entbehrlich gemacht werden soll* (1790; 2nd ed., 1791); *Die Religion innerhalb der Grenzen der blossen Vernunft* (1793; 2nd ed., 4 pt., 1794; *Religion Within the Boundary of Pure Reason,* 1838; *Religion Within the Limits of Reason Alone,* 2nd ed., 1960); *Zum ewigen Frieden* (1795; 2nd ed., 1796; *Project for a Perpetual Peace,* 1796, many later editions called *Perpetual Peace;* 1915 ed. reprinted 1972); *Die Metaphysik der Sitten* (1797; 2nd ed., 1798–1803; *The Metaphysic of Morals,* 2 vol., 1799 and 1965; *The Metaphysic of Ethics,* 1836), comprising *Metaphysische Anfangsgründe der Rechtslehre* (*The Philosophy of Law,* 1887) and *Metaphysische Anfangsgründe der Tugendlehre* (*The Doctrine of Virtue,* 1964); *Der Streit der Facultäten* (1798); *Von der Macht des Gemüths durch den blossen Vorsatz seiner krankhaften Gefühle Meister zu seyn* (1798; *Kant on the Art of Preventing Diseases,* 1806); *Anthropologie in pragmatischer Hinsicht abgefasst* (1798; improved ed., 1800; *The Classification of Mental Disorders,* 1964); *Immanuel Kants Physische Geographie,* 3 vol. in 6 pt., 1801–04); *I. Kants Logik: Ein Handbuch zu Vorlesungen* (1800; *Logic,* 1819); *Immanuel Kant über Pädagogik* (1803; *Kant on Education,* 1899; *The Educational Theory of Immanuel Kant,* 1904; *Education,* 1960); *Welches sind die wirklichen Fortschritte, die Metaphysik seit Leibnizens und Wolf's Zeiten in Deutschland gemacht hat?* (1804).

BIBLIOGRAPHY

Kant. *Biography:* The main sources for Kant's life are three memoirs published in 1804: LUDWIG ERNEST VON BOROWSKI, *Darstellung des Lebens und Charakters Immanuel Kants* (reprinted 1968); REINHOLD B. JACHMANN, *Immanuel Kant geschildert in Briefen an einen Freund* (reprinted 1968); and CHRISTOPH WASIANSKI, *Immanuel Kant in seinen letzten Lebensjahren* (the basis of THOMAS DE QUINCEY's "The Last Days of Kant" included in his *Works*). See also JOHN H.W. STUCKENBERG, *The Life of Immanuel Kant* (1882); FRIEDRICH PAULSEN, *Immanuel Kant: His Life and Doctrine* (1902, reissued 1963; originally published in German, 1898); ERNST CASSIRER, *Kant's Life and Thought* (1981; trans. of 2nd German ed., 1921); and KARL VORLÄNDER, *Immanuel Kants Leben,* 3rd ed. (1974), and *Immanuel Kant: Der Mann und das Werk,* 2 vol. (1924, reissued 1977).

Editions: The standard edition of Kant's works is that of the Berlin Academy (later the DDR Academy), *Gesammelte Schriften* (1902–), 29 vol. by 1980, which contains Kant's lectures, correspondence, and literary remains as well as his published writings. There are also modern collected editions by ERNST CASSIRER, 11 vol. (1912–23); and by KARL VORLÄNDER, 10 vol. (1920–29). A convenient edition of the *Critique of Pure Reason* is that by RAYMOND SCHMIDT, 1926).

Aids to study: RUDOLF EISLER, *Kant-lexicon* (1930, reprinted 1971); HEINRICH RATKE, *Systematisches Handlexikon zu Kants Kritik der reinen Vernunft* (1929, reprinted 1965).

General works: The best introduction is STEPHAN KÖRNER, *Kant* (1955, reissued 1982). See also EDWARD CAIRD, *The Critical Philosophy of Immanuel Kant,* 2 vol. (1889, reprinted 1969); KUNO FISCHER, *Kants Leben und die Grundlagen seiner Lehre* (1860); ALOIS RIEHL, *Der philosophische Kriticismus . . . ,* 3rd ed., 3 vol. (1924–26); BRUNO BAUCH, *Immanuel Kant,* 4th ed. (1933), in German; HENRICH RICKERT, *Kant als Philosoph der modernen Kultur* (1924); MAX WUNDT, *Kant als Metaphysiker* (1924); MARTIN HEIDEGGER, *Kant and the Problem of Metaphysics* (1962, reissued 1972; originally published in German, 1929); HERMAN J. DE VLEESCHAUWER, *La Déduction transcendentale dans l'oeuvre de Kant,* 3 vol. (1934, reissued 1976), and *L'Évolution de la pensée kantienne* (1939); PANTALEO CARABELLESE, *Il problema della filosofia in Kant* (1938); GOTTFRIED MARTIN, *Kant's Metaphysics and Theory of Science* (1955, reissued 1974; originally published in German, 1951); HEINZ HEIMSOETH, *Studien zur Philosophie Immanuel Kants,* 2nd ed. (1971); RICHARD KRONER, *Kant's Weltanschauung* (1956; originally published in German, 1914); FRIEDRICH DELEKAT, *Immanuel Kant,* 3rd ed. (1969), in German.

Precritical writings: MARIANO CAMPO, *La genesi del criticismo kantiano* (1953); GIORGIO TONELLI, *Elementi metodologici e metafisici in Kant dal 1745 al 1768* (1969), in German.

The "Critique of Pure Reason": NORMAN KEMP SMITH, *A Commentary to Kant's "Critique of Pure Reason,"* 2nd ed. rev. (1923, reissued 1979); HERBERT J. PATON, *Kant's Metaphysic of Experience,* 2 vol. (1936); ALFRED C. EWING, *A Short Commentary on Kant's Critique of Pure Reason* (1938, reprinted 1974); THOMAS D. WELDON, *Kant's Critique of Pure Reason,* 2nd ed. (1958); HERMANN COHEN, *Kants Theorie der Erfahrung,* 4th ed. (1925), and *Kommentar zu Immanuel Kants Kritik der reinen Vernunft,* 4th ed. (1925); HANS VAIHINGER, *Kommentar zu Kants Kritik der reinen Vernunft,* 2nd ed., 2 vol. (1922, reissued 1970); HEINZ HEIMSOETH, *Transzendentale Dialektik,* 3rd vol. (1966–69). See also GRAHAM BIRD, *Kant's Theory of Knowledge* (1962, reissued 1973); PETER F. STRAWSON, *The Bounds of Sense* (1966, reissued 1975).

Ethical writings: JEFFIRIE G. MURPHY, *Kant: The Philosophy of Right* (1970); PAUL A. SCHILPP, *Kant's Pre-Critical Ethics,* 2nd ed. (1960, reprinted 1977); HERBERT J. PATON, *The Categorical Imperative* (1947, reissued 1971); ORNA NELL, *Acting on Principle: An Essay on Kantian Ethics* (1975); VIGGO ROSSVAER, *Kant's Moral Philosophy: An Interpretation of the Categorical Imperative* (1979); LEWIS W. BECK, *A Commentary on Kant's Critique of Practical Reason* (1960); A.E. TEALE, *Kantian Ethics* (1951, reprinted 1975); WILLIAM D. ROSS, *Kant's Ethical Theory* (1954, reprinted 1978); HERMANN COHEN, *Kants Begründung der Ethik,* 2nd ed. (1910); MAX SCHELER, *Der Formalismus in der Ethik und die materiale Wertethik,* 6th ed. (1980). See also PAUL MENZER's ed. of *Eine Vorlesung Kants über Ethik* (1924; Eng. trans., *Lectures on Ethics by Immanuel Kant,* (1930); MORRIS STOCKHAMMER, *Kants Zurechnungsidee und Freitheitsantinomie* (1961); HENRICH W. CASSIRER, *A Commentary on Kant's Critique of Judgement* (1938, reprinted 1970); ALFRED BAEUMLER, *Kants Kritik der Urteilskraft* (1923).

Particular problems: (*Science*): ERICH ADICKES, *Kant als Naturforscher,* 2 vol. (1924–25); JULES VUILLEMIN, *Physique et métaphysique kantiennes* (1955). (*Ontology*): CHRISTOPHER B. GARNETT, *The Kantian Philosophy of Space* (1939, reprinted 1965); MARTIN HEIDEGGER, *Kants These über das Sein* (1963), and *What Is a Thing?* (1968; originally published in German, 1962). (*Philosophy of history*): YIRMIAHU YOVEL, *Kant and the Philosophy of History* (1980); KLAUS WEYAND, *Kants Beschichtsphilosophie: Ihre Entwicklung und ihr Verhältnis zur Aufklärung* (1963). (*Political philosophy*): SUSAN M. SHELL, *The Rights of Reason: A Study of Kant's Philosophy and Politics* (1980); GEORGES VLACHOS, *La Pensée politique de Kant* (1962). (*Religion*): CLEMENT C.J. WEBB, *Kant's Philosophy of Religion* (1926, reprinted 1970); JOSEF BOHATEC, *Die Religionsphilosophie Kants . . .* (1938, reprinted 1966); ALLAN W. WOOD, *Kant's Rational Theology* (1978). (*Comparative studies*): JOHANNES B. LOTZ (ed.), *Kant und die Scholastik heute* (1955); KARL JASPERS, *Die grossen Philosophen,* 2 vol. (1957–81; abridged Eng. trans., *The Great Philosophers,* ed. by HANNAH ARENDT, 1966). (*Aesthetics*): DONALD W. CRAWFORD, *Kant's Aesthetic Theory* (1974); PAUL GUYER, *Kant and the Claims of Taste* (1979).

Opus postumum: ERICH ADICKES, *Kants Opus Postumum* (1920, reprinted 1978); GERHARD LEHMANN, *Kants Nachiasswerk und die Kritik der Urteilskraft* (1939).

Journals: Many important works of Kantian scholarship have been published in the periodical *Kant-studien* (quarterly).

Kantianism. Though the literature on Kant himself comprises innumerable titles, that on Kantianism is relatively scanty. One work that contains the complete history of Kantianism is *Freidrich Ueberwegs Grundriss der Geschichte der Philosophie,* 13th ed. (1953), vol. 3:606–620 and 4:1–128 for the first period and pp. 410–483 for the second. For the first period, there is an abundant literature. Of particular interest for the present purposes are JOHANN E. ERDMANN, *Versuch einer wissenschaftlichen Darstellung der Geschichte der neuern Zeit,* 2nd ed., vol. 3 (1923); and G. LEHMANN, "Kant im Spätidealismus und die Anfänge der neukantischen Bewegung," in *Zeitschrift für philosophische Forschung,* 17:438–456 (1963). For the second period, MARIANO CAMPO has begun the history in his *Schizzo storico della esegesi e critica Kantiana* (1959); and summaries have been written by LEWIS W. BECK in the *Encyclopedia of Philosophy,* 5:468–473 (1967, reissued 1972); and by HERMANN NOACK in his *Die Philosophie Westeuropas,* pp. 143–196 (1962). See also the *Enciclopedia Filosofica,* new ed., vol. 3, col. 1225, and vol. 4, col. 953 (1967); WOLFGANG RITZEL, *Studien zum Wandel der Dantauffassung* (1952); HENRI DUSSORT, *L'École de Marbourg* (1963); and HEINRICH RICKERT, *Die Heidelberger Tradition und Kants Kritizismus* (1934).

Karāchi

Karāchi, the principal seaport and the largest city in Pakistan, is located on the coast of the Arabian Sea immediately northwest of the Indus River Delta. It is the capital of the province of Sind as well as the headquarters of the district of Karāchi. It is also a major commercial and industrial centre. The city proper covers an area of 228 square miles (591 square kilometres), while the metropolitan area of Greater Karāchi spreads out over an area of 560 square miles.

The city has been variously called Caranjee, Crochey, Krotchey, Currachee, and Kurrachee. All its names are believed to be derived from the Sindhi name of the original settlement that initially stood on the spot—Kalachi-Jo-goth (meaning the village of Kalachi—the headman of the tribe).

The impetus to Karāchi's development originally came from its role as the port serving the Indus River Valley and the Punjab region of British India. The development of air travel subsequently increased Karāchi's importance. It is also the port serving the landlocked country of Afghanistan.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. Karāchi Harbour, on the shores of which the city is situated, is a safe and beautiful natural harbour. It is protected from storms by Kiamāri Island, Manora Island, and Oyster Rocks, which together block the greater part of the harbour entrance in the west.

A low-lying coastal strip runs along the shore of the harbour. Away from the coast, the ground rises gently to the north and east to form a large plain, from five to 120 feet (1 1/2 to 37 metres) above sea level, on which the city of Karāchi is built. The Malīr River, a seasonal stream, passes through the eastern part of the city, and the Layāri River, also seasonal, runs through the most densely populated northern section. Some ridges and isolated hills occur in the north and east; Mango Pīr, the highest elevation, is 585 feet high.

The 560 square miles that comprised the Federal Capital Area of Pakistan in 1948 are considered, for all practical purposes, to form the Karāchi metropolitan area. Almost half of the area is occupied by the city and its suburbs, and the surrounding 332 square miles consist of agricultural land and government wasteland.

Climate. Karāchi has pleasant weather for the greater part of the year. May and June are the hottest months, when the mean maximum temperature is about 93° F (34° C). Spells of enervating weather occasionally prevail in May and October, during which the temperature shoots up to 105° F (41° C). The coolest months are January and February, during which the mean minimum temperature remains about 56° F (13° C). A biting north wind occasionally blows in these months, during which the temperature may drop to 40° F (4° C). The relative humidity varies from 58 percent in October, the driest month, to 82 percent in August, the wettest month. The average rainfall is eight inches (203 millimetres); most of the rain falls during a total of nine or 10 days in the months of June, July, and August.

The city faces some pollution problems. High humidity in the region does not permit evaporation of stagnant water in some places, while fumes from factories and automobiles contribute to air pollution, in spite of land and sea breezes.

Plant and animal life. The natural vegetation is scanty. Seaweed rises in tangles, and mangroves grow along some of the shores. Coarse grass, cactus, and castor plants occur on the plains and hills, and date and coconut palms grow in the river valleys.

The common wild animals are wolves, chinkaras (a type of gazelle), hog deer, jackals, wild cats, and hares. Domestic animals include sheep, goats, horses, and cows. Local birds include geese, ducks, snipe (game birds related to the woodcock), cranes, flamingos, and ibis (wading birds related to the heron). There are various types of snakes found in the region, particularly cobras, kraits, vipers, and python.

The city layout. The most striking aspect of Karāchi's layout is the west-to-east parallel alignment of the four arterial roads—Nishter Road (formerly called Lawrence Road), Mohammed Ali Jinnah Road (formerly Bandar Road), Shahrah-e-Liaquat (Frere Road), and I.I. Chundrigar Road (McCleod Road). Beginning at Mereweather Tower in the vicinity of the port, these roads run through the centre of the city. Several roads, such as Napier Road, Dr. Zia-ud-din Ahmed Road (Kutchery Road), and Garden Road, cut perpendicularly across these arteries from north to south.

The old town

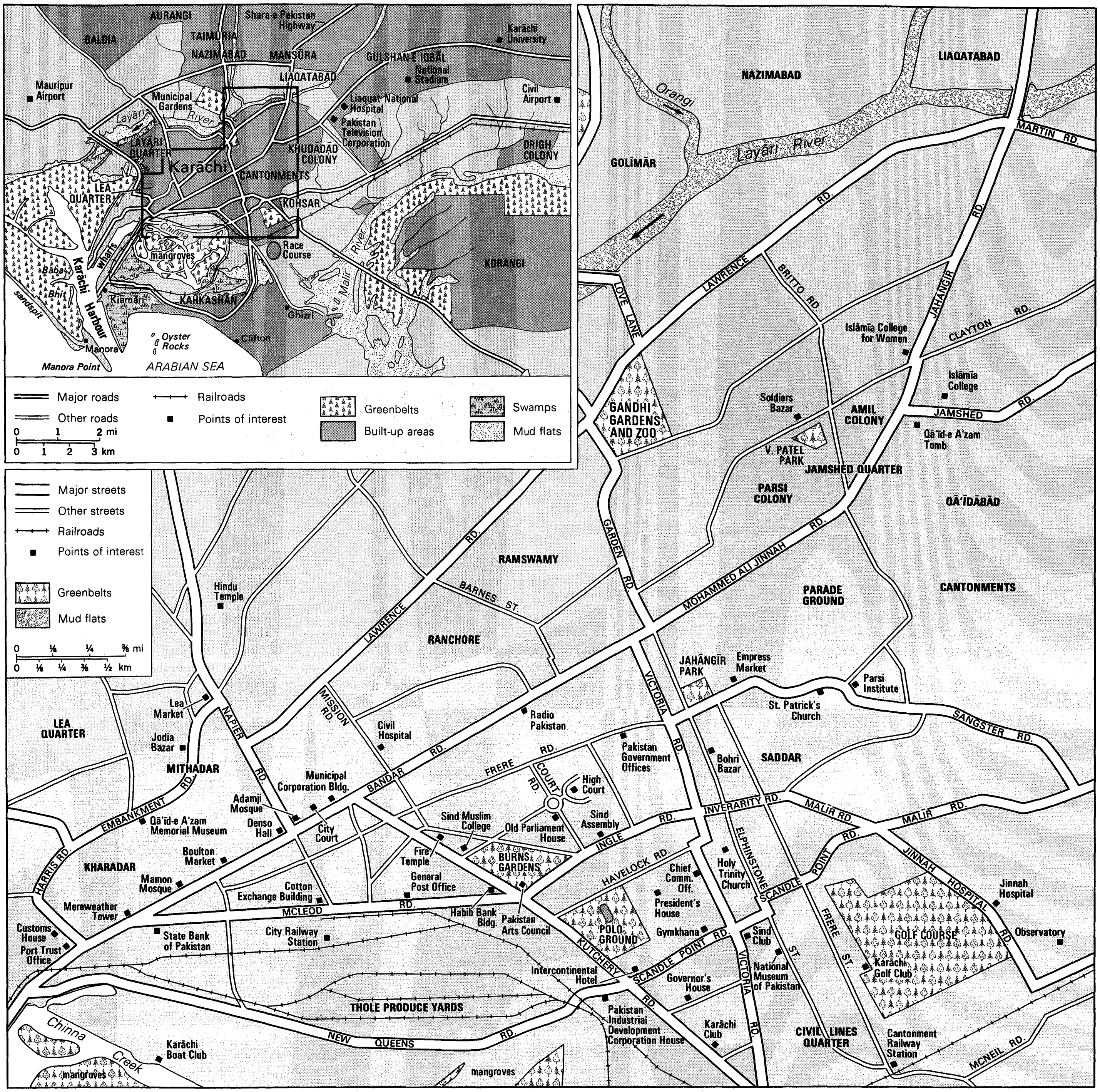
The old town lies near the port, to the north of M.A. Jinnah Road, and with extensions stretching along the material roads for over a mile; unplanned, it is reminiscent of medieval towns of the Near East or Europe. East of the old town are such districts as the Drigh Cantonment, the Civil Lines (residential areas for senior civil service officers), and the Saddar Bazar. This area is planned on a checkerboard pattern, and shows European characteristics. Beyond this stretch several radial roads, along which growth has taken the form of neighbourhood units; each unit is laid out with straight, broad roads connected by smaller streets.

Land use

The land-use pattern of the city is complex. In the central area, the preponderance of residential property tends to form a matrix within which all other functions are distributed. There is, however, a marked concentration of commercial buildings at the western ends of M.A. Jinnah Road and I.I. Chundrigar Road. Wholesale businesses are located in the old town, retail businesses along M.A. Jinnah Road and in Saddar Bazar, and the government offices on Shahrah-e-Liaquat, near Saddar. The outer areas are dominated by dormitory suburbs interspersed with a scattering of cantonments (military quarters), agricultural tracts, salt works, airports, railway stations, and marshalling yards.

The city proper has old and decayed buildings, occupied by members of the middle and lower income groups. Further from the city centre are modern bungalows occupied by richer persons; the outer zone is occupied by workers.

Karāchi has a variety of types of buildings. The central area contains apartment bungalows, barracks, and mul-



The city of Karāchi and (inset) its metropolitan area.

tistoried buildings; the outer areas are characterized by bungalows, blocks of flats, and quarters (streets of small houses). Buildings of the British period were constructed with stone in western styles of architecture; other stone buildings in the central city show a blending of Eastern and Western styles, and have towers, domes, pillars, arches, hanging balconies, and rectangular courtyards. Buildings in the outer areas are built of cement blocks, and with few exceptions they show no uniformity in design. Some follow contemporary North American design, while others incorporate features of traditional Muslim architecture.

THE PEOPLE

No ethnic group predominates in the city. Cultural and social activities essentially revolve around religion. The population is almost entirely Muslim, but there are also small Christian, Hindu, Parsi, Buddhist, and Jain minorities. Most of the Muslims derive from Indo-Pakistani stock, except for "Makranis" and "Shiddies," who are also descended from Negro elements, and who originated during the era of the slave trade in the days before British rule, when Karāchi was an important slave-trading centre. Some of the members of the Christian minority are of Indo-Pakistani origin, while others are descended from Portuguese or other European groups.

THE ECONOMY

Manufacturing

Industry. Textiles and footwear are the principal items manufactured, followed by such items as metal products, food and beverages, paper and printing, wood and furniture, machinery, chemicals and petroleum, leather and rubber, and electrical goods. Karāchi is also an important centre for handicrafts and cottage industries that produce handloomed cloth, lace, carpets, articles made of brass and

The Defence Society Mosque in Karāchi.
William MacQuitty—Camera Press

bell metal (an alloy of copper and tin), pottery, leather goods, and gold and silver embroidery. Karāchi handles the entire seaborne trade of Pakistan and of landlocked Afghanistan.

Finance. There are more than 25 banks in Karāchi that have branches throughout Pakistan; these include the State Bank of Pakistan, the Habib Bank Ltd., the National Bank of Pakistan, the United Bank Ltd., the Industrial Development Bank of Pakistan, and the Agricultural Development Bank of Pakistan. The city is also the centre of about two dozen insurance companies, which play an important role in the economic development of the country by investing large sums in power development, housing programs, jointstock companies, government loan securities, and savings certificates.

Karāchi has a stock exchange that handles an overwhelming proportion of transactions in government securities and in the shares of most of the important industrial and financial institutions.

Transportation. The Karāchi–Peshāwar highway links the city with the interior of Pakistan, while the Karāchi–Ormāra highway extends along the coast. The Karāchi to Zāhedān highway connects it with Iran and other Middle Eastern countries. Express roads radiate from the city centre, while feeder roads connect the express roads with local streets.

Karāchi is the terminus of Pakistan's railway system, which mainly serves to transport goods between Karāchi and the interior. There are also passenger trains, as well as a circular railway that skirts the city on the north and the east, for commuter traffic and the transport of goods between the port and the industrial areas.

Karāchi Airport provides international and domestic services. The port of Karāchi is one of the busiest east of Suez.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The city is administered by five institutions, the heads of which are appointed by the government. The Karāchi Municipal Corporation, constituted in 1852, performs a large number of civic functions affecting more than three-fourths of the population of Greater Karāchi. The Korangi-Lāndhi and Drigh-Malīr municipal committees were established in 1966 and 1970, respectively, to provide civic facilities to the suburban areas developed after 1947. The Karāchi Cantonment Board is the administrative body for the areas where the military are quartered. The Karāchi Port Trust administers the affairs of the port and is entrusted with the development and maintenance of the harbour.

Sources of water supply

Public utilities. The three main sources of the city's water supply are Lake Hāleji, 55 miles (90 kilometres) away, fed by the Indus River; wells that have been sunk in the dry bed of the Malīr River, 18 miles away; and Lake Kalri, 60 miles away, also fed by the Indus waters. Although the city's water mains stretch for many miles, some of the outer areas, such as Lāndhi, Malīr, New Karāchi, and Mauripur, still have an acute water shortage.

The Karāchi Electric Supply Corporation is responsible for electricity services. It has several power stations located in the city; these stations use natural gas, diesel oil, or both. A nuclear power station is operated at Paradise Point.

Karāchi Municipal Corporation maintains a fleet of vehicles for refuse collection, night soil removal, dog catching, and antimalarial and antifly operations. Sweepers are employed to clean the streets. Sewage is disposed of by two underground drainage systems, and there are two sewage treatment plants, one serving the city proper and the other the outlying areas.

Health and security. Karāchi proper has more than 20 general hospitals, as well as several hospitals specializing in tuberculosis, skin diseases, leprosy, and epidemic diseases. There are also child-welfare centres and dispensaries, in addition to general hospitals in the suburbs.

There are several well-equipped fire-fighting stations; separate fire brigade units are attached to the railway network. In addition, the Port Trust and Pakistan International Airlines (PIA) have services that can be used in emergencies.

The police are administered by the Sind provincial administration; the inspector general of police is assisted by a force of more than 1,100. The city is divided into 40 police districts.

Education. Karāchi has more than 900 schools, of which the majority are primary schools and the rest are secondary schools. More than half of all these are privately run, the rest being run by the government. Among schools established by different religious communities are Karāchi Grammar School, St. Joseph's Convent School, and St. Patrick's High School, all of which are Christian; a school for the Parsi community; and Sind Madrassa, a Muslim school.

The University of Karāchi

The University of Karāchi is the primary institution of higher education. It has more than 20 graduate departments in arts and sciences, as well as a graduate school of business administration. Courses in a variety of subjects, including commerce and law, are provided by about 75 colleges affiliated to the university. In addition, there is a medical college, as well as two engineering colleges, a polytechnic institute, a college of home economics, and two teacher-training colleges.

CULTURAL LIFE

The Arts Council of Pakistan is the primary cultural institution in the city; it organizes various cultural functions including art exhibitions, and offers training in music. The Ghanshyam Art Centre and the Bulbul Academy promote Pakistani dancing and other cultural activities.

Karāchi does not have well-established theatre, but amateur dramas and variety shows are frequently staged in Katrak Hall. Motion pictures are more popular, and there are more than 50 cinemas.

Karāchi has a small museum containing relics of the early Indus Valley civilization and examples of the Greco-Buddhist art of Gandhāra (a region of ancient India in what is now northwestern Pakistan); it also has some ethnological collections representing life in different regions of Pakistan.

The library of the University of Karāchi is the city's largest, but there are other libraries containing books of a popular nature. Material of a more scholarly nature is to be found in the British Council Library, the American Center Library, and the Liaquat Memorial Library. The departmental libraries of the State Bank of Pakistan, the Pakistan Institute of Development Economics, and the National Archives contain collections of books on economics and on national matters.

There is a general shortage of open spaces and parks in Karāchi. Gandhi Gardens and Fatima Jinnah (Burns) Gardens are popular parks. There are a number of fine swimming and fishing beaches, such as Paradise Point, Hawkes Bay, Sandspit, Manora, and Clifton. The Karāchi Zoo is located in the Gandhi Gardens, and contains a varied collection of animals, birds, and reptiles.

Sports and games facilities are mostly provided by such associations as the Karāchi Gymkhana, the Parsi Gymkhana, the Agha Khan Gymkhana, and the Young Men's Christian Association (YMCA). Various organizations and educational institutions have their own playgrounds. The largest sports area is the National Stadium, which contains playgrounds for cricket, hockey, football (soccer), and tennis. There are also boating, yachting, and flying clubs.

History

Origins of the port

Karāchi was a small fishing village when a group of traders moved there in the early 18th century from the decaying port of Kharak Bandar nearby. Besides the natural protection against monsoon storms, Manora Head furnished an excellent site for the defense of the harbour, and the Talpura *amīr*s who gained Karāchi from the *khān* of Kalāt in 1795 erected a permanent fort on it. The settlement expanded rapidly, and was already of significance when it was captured in 1839 by the British, who annexed it in 1842, together with the province of Sind. It then became an army headquarters for the British, and also began to develop from a fishing village into the principal port for the Indus River region.

In 1843 a river-steamer service was introduced between Karāchi and Multān, about 500 miles up the Indus. Port facilities were improved from 1854 onward. In 1861 a railway was built from Karāchi to Kotri, 90 miles upstream on the right bank of the Indus, opposite Hyderābād. In 1864 direct telegraph communications were established with London and with the interior. With the opening of the Suez Canal in 1869, the importance of Karāchi grew, and it became a full-fledged seaport. By 1873 it possessed an efficient and well-managed harbour.

Karāchi was connected directly with the hinterland when the railway line was extended from Kotri in 1878 to join the Delhi-Punjab railway system at Multān. In 1886, the Karāchi Port Trust was established as the port authority, and between 1888 and 1910 the East Wharf—186,000 feet in length—was constructed. When the Punjab emerged as the granary of India in the 1890s, Karāchi became the region's principal outlet. By 1914 it had become the largest grain exporting port of the British Empire.

After World War I, manufacturing and service industries were installed. By 1924 an aerodrome had been built, and Karāchi became the main airport of entry to India. The city became the provincial capital of Sind in 1936.

Growth after 1947

With the creation of Pakistan in 1947, Karāchi not only became the capital and premier port of the new country but also a centre for industry, business, and administration. Although Rāwalpindi became the interim capital in 1959, some governmental agencies, including the Public Service Commission, as well as "skeleton" staffs of various ministries, remained in Karāchi, which continued to play a role as a multifunctional city serving the entire country.

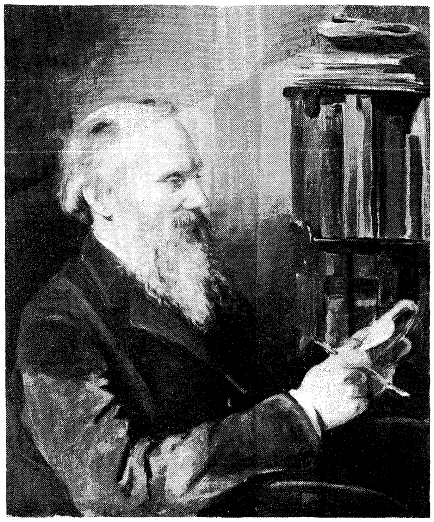
BIBLIOGRAPHY. NAZIR AHMAD, *Survey of Shelterless Persons in Karachi 1959* (1959), is a detailed report of the problems faced by the city due to the influx of refugees from India following the establishment of Pakistan. KARACHI DEVELOPMENT AUTHORITY, *The Greater Karachi Resettlement Housing Programme* (1961), analyzes the housing requirements, activities, and needs in metropolitan Karāchi. A.F. BAILLIE, *Kurrachee: Past, Present and Future* (1890), gives Karāchi's history and its growth in the early years of British rule. R.F. BURTON, *Scinde; or, the Unhappy Valley,* 2 vol. (1851), and *Sind Revisited,* 2 vol. (1877), are classic surveys of Sind under British rule. The *District Census Report of Karachi, 1972,* provides official demographic data. H. FELDMAN, *Karachi Through a Hundred Years* (1960), written to commemorate the centenary of the Karāchi Chamber of Commerce and Industry, gives details of the role played by the Chamber in the economic development of Karāchi from 1860 to 1960. H.T. LAMBRICK, *Sind: A General Introduction,* 2nd ed. (1975), includes a discussion of Karāchi. RICHARD F. NYROP *et al., Area Handbook for Pakistan* (1975), discusses many aspects of modern Karachi. Karāchi Development Authority MP Reports have been written for the preparation of the Master Plan of Karāchi and cover all aspects of city development. GUSTAV RANIS, *Industrial Efficiency and Economic Growth: A Case Study of Karachi* (1961), is an analysis of economic development; and IMTIAZUDDIN HUSAIN, MOHAMMED AFZAL, and AMJAD ALI BAHADUR RIZVI, *The Social Characteristics of the People of Karachi* (1965), a study of the socioeconomic characteristics of the people.

(Z.A.K.)

Kelvin

William Thomson, who was knighted in 1866 and was raised to the peerage in 1892 (as Baron Kelvin of Largs) in recognition of his work in engineering and physics, was foremost among the small group of British scientists who helped to lay the foundations of modern physics. His contributions to science included a major role in the development of the second law of thermodynamics; the absolute temperature scale (measured in kelvins); the dynamical theory of heat; the mathematical analysis of electricity and magnetism, including the basic ideas for the electromagnetic theory of light; the geophysical determination of the age of the Earth; and fundamental work in hydrodynamics. His theoretical work on submarine telegraphy and his inventions for use on submarine cables aided Britain in capturing a preeminent place in world communication during the 19th century.

The style and character of Thomson's scientific and engineering work reflected his active personality. While a student at the University of Cambridge, he was awarded



Kelvin, oil painting by Elizabeth King, 1886–87. In the National Portrait Gallery, London.

By courtesy of the National Portrait Gallery, London

silver sculls for winning the university championship in racing single-seater rowing shells. He was an inveterate traveller all of his life, spending much time on the Continent and making several trips to the United States. In later life he commuted between homes in London and Glasgow. Thomson risked his life several times during the laying of the first transatlantic cable.

Thomson's worldview

Thomson's worldview was based in part on the belief that all phenomena that caused force—such as electricity, magnetism, and heat—were the result of invisible material in motion. This belief placed him in the forefront of those scientists who opposed the view that forces were produced by imponderable fluids. By the end of the century, however, Thomson, having persisted in his belief, found himself in opposition to the positivistic outlook that proved to be a prelude to 20th-century quantum mechanics and relativity. Consistency of worldview eventually placed him counter to the mainstream of science.

But Thomson's consistency enabled him to apply a few basic ideas to a number of areas of study. He brought together disparate areas of physics—heat, thermodynamics, mechanics, hydrodynamics, magnetism, and electricity—and thus played a principal role in the great and final synthesis of 19th-century science, which viewed all physical change as energy-related phenomena. Thomson was also the first to suggest that there were mathematical analogies between kinds of energy. His success as a synthesizer of theories about energy places him in the same position in 19th-century physics as Sir Isaac Newton has in 17th-century physics or Albert Einstein in 20th-century physics. All of these great synthesizers prepared the ground for the next grand leap forward in science.

Early life. William Thomson was born on June 26, 1824, in Belfast, Ireland, the fourth child in a family of seven. His mother died when he was six years old. His father, James Thomson, who was a textbook writer, taught mathematics, first in Belfast and later as a professor at the University of Glasgow; he taught his sons the most recent mathematics, much of which had not yet become a part of the British university curriculum. An unusually close relationship between a dominant father and a submissive son served to develop William's extraordinary mind.

Influences on his career

William, age 10, and his brother James, age 11, matriculated at the University of Glasgow in 1834. There William was introduced to the advanced and controversial thinking of Jean-Baptiste-Joseph Fourier when one of Thomson's professors loaned him Fourier's pathbreaking book *The Analytical Theory of Heat,* which applied abstract mathematical techniques to the study of heat flow through any solid object. Thomson's first two published articles, which appeared when he was 16 and 17 years old, were a defense of Fourier's work, which was then under attack by British scientists. Thomson was the first to promote the idea that Fourier's mathematics, although applied solely to the flow of heat, could be used in the study of other forms of energy—whether fluids in motion or electricity flowing through a wire.

Influence of Fourier

Thomson won many university awards at Glasgow, and at the age of 15 he won a gold medal for "An Essay on the Figure of the Earth," in which he exhibited exceptional mathematical ability. That essay, highly original in its analysis, served as a source of scientific ideas for Thomson throughout his life. He last consulted the essay just a few months before he died at the age of 83.

Thomson entered Cambridge in 1841 and took his B.A. degree four years later with high honours. In 1845 he was given a copy of George Green's *An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism.* That work and Fourier's book were the components from which Thomson shaped his worldview and which helped him create his pioneering synthesis of the mathematical relationship between electricity and heat. After finishing at Cambridge, Thomson went to Paris, where he worked in the laboratory of the physicist and chemist Henri-Victor Regnault to gain practical, experimental competence to supplement his theoretical education.

The chair of natural philosophy (later called physics) at the University of Glasgow fell vacant in 1846. Thomson's father then mounted a carefully planned and energetic campaign to have his son named to the position, and at the age of 22 William was unanimously elected to it. Despite blandishments from Cambridge, Thomson remained at Glasgow for the rest of his career. He resigned his university chair in 1899, at the age of 75, after 53 years of a fruitful and happy association with the institution. He was making room, he said, for younger men.

Thomson's scientific work was guided by the conviction that the various theories dealing with matter and energy were converging toward one great, unified theory. He pursued the goal of a unified theory even though he doubted that it was attainable in his lifetime or ever. The basis for Thomson's conviction was the cumulative impression obtained from experiments showing the interrelation of forms of energy. By the middle of the 19th century it had been shown that magnetism and electricity, electromagnetism, and light were related, and Thomson had shown by mathematical analogy that there was a relationship between hydrodynamic phenomena and an electric current flowing through wires. James Prescott Joule also claimed that there was a relationship between mechanical motion and heat, and his idea became the basis for the science of thermodynamics.

Joule's theory of interconvertability

In 1847 Thomson first heard Joule's theory about the interconvertability of heat and motion at a meeting of the British Association for the Advancement of Science. Joule's theory went counter to the accepted knowledge of the time, which was that heat was an imponderable substance (caloric) and could not be, as Joule claimed, a form of motion. Thomson was open-minded enough to discuss with Joule the implications of the new theory. At the time, though he could not accept Joule's idea, Thomson was willing to reserve judgment, especially since the relation between heat and mechanical motion fit into his own view of the causes of force. By 1851 Thomson was able to give public recognition to Joule's theory, along with a cautious endorsement in a major mathematical treatise, "On the Dynamical Theory of Heat." Thomson's essay contained his version of the second law of thermodynamics, which was a major step toward the unification of scientific theories.

Thomson's work on electricity and magnetism also began during his student days at Cambridge. When, much later, James Clerk Maxwell decided to undertake research in magnetism and electricity, he read all of Thomson's papers on the subject and adopted Thomson as his mentor. Maxwell—in his attempt to synthesize all that was known about the interrelationship of electricity, magnetism, and light—developed his monumental electromagnetic theory of light, probably the most significant achievement of 19th-century science. This theory had its genesis in Thomson's work, and Maxwell readily acknowledged his debt.

Thomson's contributions to 19th-century science were numerous. He advanced other scientists' ideas, including those of Michael Faraday, Fourier, and Joule. Thomson subjected experimenters' results to mathematical analysis and showed the generality of the experimental results. He formulated the concept that was to be generalized into the dynamical theory of energy. He also collaborated with a number of leading scientists of the time, among them Sir George Gabriel Stokes, Hermann von Helmholtz, Peter Guthrie Tait, and Joule. With these partners he advanced the frontiers of science in several areas, particularly hydrodynamics. Furthermore, Thomson originated the mathematical analogy between the flow of heat in solid bodies and the flow of electricity in conductors.

The transatlantic cable

Although a mild man by nature, Thomson became involved in several controversies, and one, over the feasibility of laying a transatlantic cable, changed the course of his professional work. His involvement with the project began with an inquiry from Stokes, a lifelong correspondent on scientific matters, who asked in 1854 for a theoretical explanation of the apparent delay in an electric current passing through a long cable. In his reply Thomson referred to his early paper "On the Uniform Motion of Heat in Homogeneous Solid Bodies, and its Connexion with the Mathematical Theory of Electricity" (1842). Thomson's idea about the mathematical analogy between heat flow and electric current worked well in his analysis of the problem of sending telegraph messages through the planned 3,000-mile (4,800-kilometre) cable. His mathematical equations describing the flow of heat through a solid wire, as Thomson showed, were applicable to questions arising over the velocity of a current in a cable.

The publication of Thomson's reply to Stokes prompted a rebuttal by E.O.W. Whitehouse, the Atlantic Telegraph Company's chief electrician. Whitehouse claimed that practical experience refuted Thomson's theoretical findings, and for a time Whitehouse's view prevailed with the directors of the company. Despite their disagreement, Thomson participated, as chief consultant, in the hazardous early cable-laying expeditions. In 1858 Thomson patented his telegraph receiver, called a mirror galvanometer, for use on the Atlantic cable. (The device, along with his later modification called the siphon recorder, came to be used on most of the worldwide network of submarine cables.) Eventually the directors of the Atlantic Telegraph Company fired Whitehouse, adopted Thomson's suggestions for the design of the cable, and decided in favour of the mirror galvanometer. In doing so the company saved years of effort and a large amount of money, and in 1866 Thomson was knighted by Queen Victoria for his work.

Later life. After the successful laying of the transatlantic cable, Thomson became a partner in two engineering consulting firms. These companies played a major role in the planning and construction of submarine cables during the frenzied era of expansion that resulted in a worldwide network of telegraph communication. In the process, Thomson became a wealthy man who could afford a 126-ton yacht and a baronial estate.

Thomson's broad interests in science not only encompassed electricity, magnetism, thermodynamics, and hydrodynamics but also included geophysical questions concerning tides, the shape of the Earth, atmospheric electricity, thermal studies of the ground, rotation of the Earth on its axis, and geomagnetism. Believing that all science must be subjected to the same analytical rigour, he did not hesitate to enter the controversy over Charles Darwin's theory of evolution. Thomson opposed Darwin, remaining "on the side of the angels."

Investigations of the age of the Earth

Thomson challenged the views on geological and biological change of the early uniformitarians, including Darwin, who claimed that the Earth and its life had evolved over an incalculable number of years, during which the forces of nature always operated as at present. On the basis of thermodynamic theory and Fourier's studies, Thomson estimated in 1862 that more than 1,000,000 years earlier the Sun's heat and the temperature of the Earth must have been considerably greater and that these conditions had produced violent storms and floods and an entirely different type of vegetation. His views, published in 1868, particularly angered Darwin's supporters. Thomas Henry Huxley used the occasion of the Anniversary Address of the President of the Geological Society of London in 1869 to reply to Thomson. Although Thomson's speculations as to the age of the Earth and the Sun were inaccurate, he did succeed in pressing his contention that biological and geological theory had to conform to the well-established theories of physics.

In 1884 Thomson delivered a series of lectures at Johns Hopkins University on the state of scientific knowledge and wondered aloud about the failures of the wave theory of light to explain certain phenomena. Even when relaxing, Thomson questioned. The time spent on his yacht, the "Lalla Rookh," roused his interest in the sea, and from that interest came a number of patents: a compass that was adopted by the British Admiralty; a form of analogue computer for measuring tides in a harbour and for calculating tide tables for any hour, past or future; and sounding equipment. He established a company to manufacture these items and a number of electrical measuring devices. Like his father, he published a textbook, *Treatise on Natural Philosophy* (1867), a work on physics that he coauthored with Tait, which helped shape the thinking of the the next generation of physicists.

Thomson was said to be entitled to more letters after his name than any other man in the Commonwealth. He received honorary degrees from major universities throughout the world and was lauded by many engineering societies and scientific organizations. He was elected a fellow of the Royal Society in 1851 and served as its president from 1890 to 1895. He published more than 600 papers and was granted dozens of patents. He died on December 17, 1907, at Netherhall, his estate near Largs, North Ayrshire, Scotland, and he was buried in Westminster Abbey, London.

BIBLIOGRAPHY. HAROLD ISSADORE SHARLIN and TIBY SHARLIN, *Lord Kelvin: The Dynamic Victorian* (1979), provides an interpretation of the sources of Thomson's originality. SILVANUS P. THOMPSON, *The Life of Lord Kelvin*, 2nd ed., 2 vol. (1976), written by a man who knew and admired Thomson, includes a complete bibliography of his published works and complete lists of his patents and of his honours and awards. ANDREW GRAY, *Lord Kelvin: An Account of His Scientific Life and Work* (1908, reprinted 1973), is an admiring but not thorough account. ELIZABETH THOMSON KING, *Lord Kelvin's Early Home* (1909), consists of reminiscences of Thomson's early family life by his oldest sister. DAVID B. WILSON (comp.), *Catalogue of the Manuscript Collections of Sir George Gabriel Stokes and Sir William Thomson, Baron Kelvin of Largs in Cambridge University Library* (1976), is a guide to the large collection of Kelvin's papers at Cambridge and to the papers of Stokes, Kelvin's lifelong scientific correspondent. JOE D. BURCHFIELD, *Lord Kelvin and the Age of the Earth* (1975), gives an account of Kelvin's role in the argument between Darwin and British geologists.

(H.I.S.)

Kepler

The Renaissance astronomer and astrologer Johannes Kepler is best known for his discovery of the three principles of planetary motion, by which he clarified the spatial organization of the solar system. Moreover, he founded modern optics by presenting the earliest correct explanation of how human beings see. He was the first to set forth accurately what happens to light after it enters a telescope, and he designed a particular form of that instrument. His ideas provided a transition from the ancient geometrical description of the heavens to modern dynamical astronomy, into which he introduced the concept of physical force.

Early life. On December 27, 1571, in the German town of Weil der Stadt, then a "free city" within the Holy Roman Empire, Johannes was born prematurely, the offspring of an unhappy marriage. His father was a ne'er-do-well mercenary soldier, his mother the quarrelsome daughter of an innkeeper. Small in stature, Johannes never enjoyed robust health, but his superior intelligence was recognized even when he was a young child. Coming from a poor family, he would have received no education had not the dukes of Württemberg adopted the enlightened policy of providing generous scholarships for the bright sons of their impoverished subjects.

With such help Kepler in 1587 was able to attend the University of Tübingen, where he had the good fortune to study astronomy under Michael Mästlin, a professor who may have been unique in his day, for he was convinced that the astronomical system propounded by Nicolaus Copernicus was basically true: the Earth is a planet that rotates daily around its own axis and revolves annually around the Sun. Kepler's youthful acceptance of Copernican astronomy profoundly affected the subsequent course of his life.

Major achievements. After obtaining the B.A. in 1588 and the M.A. in 1591, Kepler planned to become a Lutheran minister. But in 1594, during his last year of training in theology at Tübingen, the teacher of mathematics in the Lutheran high school of Graz, in Austria, having died, Kepler was strongly recommended by the Tübingen faculty to fill the vacancy. Kepler did not finish the theology course at Tübingen but went to Graz the same year. On a summer day in 1595, while he was teaching a class, a spectacular idea flashed through his mind. Ancient Greek geometry had proved that there were five regular solids, or "Platonic bodies": tetrahedron (pyramid), cube, octahedron (formed by eight equilateral triangles), dodecahedron (12 pentagons), and icosahedron (20 equilateral triangles). The ancients knew that these five solids could be enclosed in a sphere, and that there can be no additional regular solids. Sustained by a vision of mathematical harmonies in the skies, a vision he derived from the philosophy of Plato and the mathematics of the Pythagoreans, Kepler tried to relate planetary orbits with geometrical figures.

Kepler, engraving by an unknown artist, *c.* 1730, after a contemporary painting.
Archiv fur Kunst und Geschichte

The notion of cosmic harmony

According to Copernican astronomy there were six planets, whose orbits were regulated by the turning of invisible spheres. But why were there only six planets and not nine or 100? Was the cosmos so constructed that one of the five regular solids intervened between each pair of the unseen spheres, which carried the six Copernican planets? This nest of alternating planets and regular solids constituted the main theme in Kepler's *Prodromus Dissertationum Mathematicarum Continens Mysterium Cosmographicum* ("Cosmographic Mystery"), which he published in 1596 under the auspices of the Tübingen faculty. The Platonic and Pythagorean components in Kepler's conception of celestial harmony, however mystical in origin, helped to lead him to the three principles of planetary motion now known by his name.

Kepler sent copies of his first major work to a number of scientists, including Tycho Brahe, who was soon to become the imperial mathematician of the Holy Roman Empire. Although Brahe did not agree with the underlying Copernican foundation of Kepler's *Mysterium Cosmographicum,* he was so impressed by the author's knowledge of astronomy and skill in mathematics that in 1600 he invited him to join his research staff in the observatory at Benatek (now Benátky nad Jazerou), outside Prague. When Brahe died the next year, Kepler was promptly appointed his successor as imperial mathematician. His first publication at Prague, *De Fundamentis Astrologiae Certioribus* (1601; "The More Reliable Bases of Astrology"), rejected the superstitious view that the stars guide the lives of human beings. Nonetheless, his deep feeling for the harmony of the universe included a belief in the harmony between the universe and the individual, and his skill in astrological prediction was much in demand.

While Kepler was watching a rare conjunction of Mars, Jupiter, and Saturn in October 1604, a supernova appeared that remained visible for 17 months. This event was evidence that the realm of the fixed stars, considered since ancient times as pure and changeless, could indeed experience change. He published the results of his observations in 1606 as *De Stella Nova in Pede Serpentarii* ("The New Star in the Foot of the Serpent Bearer").

Contributions to optics

Kepler now had access to Brahe's incomparable collection of astronomical observations, the result of decades of unremitting and painstaking toil by the greatest naked-eye observer of the heavens and the leader of a highly qualified team of astronomers. As a member of the team, Kepler had been assigned to investigate the planet Mars. But, before he could use the raw observations, Kepler felt that he had to solve the problem of atmospheric refraction: how is a ray of light, coming from a distant heavenly body located in the less dense regions of outer space, deflected when it enters the denser atmosphere surrounding the Earth?

Kepler incorporated his results in a book that he modestly entitled *Ad Vitellionem Paralipomena, Quibus Astronomiae Pars Optica Traditur,* (1604; "Supplement to Witelo, Expounding the Optical Part of Astronomy"); Witelo (Latin Vitellio) had written the most important medieval treatise on optics. But Kepler did much more than add to his work. He made an analysis of the process of vision that provided the foundation for all of the advances in the understanding of the structure and function of the human eye. Kepler wrote that every point on a luminous body in the field of vision emits rays of light in all directions, but that only those rays can enter the eye that impinge on the pupil, which functions as a diaphragm. He stated that the rays emanating from a single luminous point form a cone, the circular base of which is in the pupil. All of the rays are then refracted within the normal eye to meet again at a single point on the retina, identified by Kepler as the sensitive receptor of the eye. If the eye is not normal, the second short interior cone comes to a point not on the retina but in front of it or behind it, causing blurred vision. For more than three centuries eyeglasses had helped older persons to see better. But nobody before Kepler was able to explain how these little pieces of curved glass had worked.

After the invention of the telescope had been reported to Galileo, who promptly proceeded to make his astounding discoveries, Kepler applied the same ideas concerning optics to the explanation of how the telescope works. Although Galileo's findings were received in general with skepticism and ridicule, Kepler acknowledged the Italian's accomplishments in his *Dissertatio cum Nuncio Sidereo Nuper ad Mortales Misso a Galilaeo Galilaeo* in 1610.

Kepler's three principles of planetary motion

Galileo did not return the compliment. He chose to ignore the epoch-making results Kepler had published the preceding year. In his *Astronomia Nova* ("New Astronomy") of 1609, Kepler had demonstrated that the orbit of the planet Mars is an ellipse. Although it had been believed since antiquity that the planets, being heavenly bodies, were perfect and therefore could move only in perfect circles or combinations of circles, Copernicus had correctly classified the Earth as one of the planets; and it was fully accepted that the Earth was far removed from perfection. Kepler extended Copernicus' reasoning to the other planets and was the first to declare that the other planets resemble the Earth in being material bodies. That a material body, being imperfect, need not travel in a perfectly circular orbit was a conclusion made by Kepler after he tried unsuccessfully to fit the orbit of Mars to Brahe's observations in every possible combination of circles his ingenuity could devise. Because none of them worked, he tried noncircular paths until he found the true solution: Mars revolves in an elliptical orbit with the Sun occupying one of its two focuses.

The pre-Keplerian dogma that permitted only circular paths entailed the concept of uniform motion—*i.e.,* the moving body or point must traverse equal arcs in equal intervals of time. Such a conception of uniform motion as measured along an arc was, of course, incompatible with an elliptical orbit. But Kepler found an alternative form of uniformity. This new uniformity equated equal areas with equal times. With the Sun remaining stationary in one focus of the ellipse, the planet, while revolving along the periphery of its elliptical orbit, would sweep out, in equal intervals of time, equal areas of the ellipse, not equal arcs along the periphery of the ellipse.

In 1619, 10 years after Kepler published these first two principles of planetary motion (the elliptical orbit and equality of areas), he published the *Harmonice Mundi (Harmonics of the World),* in which he expounded his third principle, which related a planet's mean distance from the Sun to the time it takes to complete its elliptical orbit around the Sun. The cube of the distance proved to be in a constant ratio to the square of the time required for all the planets to complete such an orbit. The enunciation of this rule (which is sometimes called the $^{3}/_{2}$ ratio) completed Kepler's contribution to the understanding of planetary motion and helped to prepare the way for Sir Isaac Newton's exposition of universal gravitation, which affects all of the material bodies in the physical universe.

Later life. Meanwhile, Kepler's patron, the Holy Roman Emperor, had been compelled by his brother to abdicate, and Kepler himself had found it desirable to leave Prague, then the capital of the empire. Although he was reappointed imperial mathematician by the new emperor, Kepler moved to Linz, in Austria. His first wife had died in Prague; Kepler remarried in 1613. Once, when buying supplies for his new home, Kepler became unhappy about the rough-and-ready methods used by the merchants to estimate the liquid contents of a wine barrel. Because the curved containers they used were of various shapes, Kepler sought a mathematical method for determining their volumes. Following the model established by Archimedes, the most talented mathematician of antiquity, Kepler, in his volumetric researches, investigated the properties of nearly 100 solids of revolution—made by rotating a two-dimensional surface on one of its axes—that had not been considered by Archimedes. Starting with an ordinary wine barrel, Kepler enormously extended the range of Archimedes' results. He did so by refusing to confine himself, as Archimedes had done, to cases in which a surface is generated by a conic section—a curve formed by the intersection of a plane and a cone—rotating about its principal axis. Kepler's additional solids are generated by rotation about lines in the plane of the conic section other than its principal axis.

While he was in Linz, Kepler published his *Epitome Astronomiae Copernicanae* (1618–21; *Epitome of Copernican Astronomy*). He modelled this title after the highly successful introduction to astronomy that had been published by his former Tübingen professor in a number of editions. But, whereas Mästlin had deemed it prudent pedagogical practice to keep Copernicanism out of an elementary textbook, which he therefore entitled simply "Epitome of Astronomy," Kepler emphasized his open espousal of the new cosmology by inserting the provocative label "Copernican."

Witchcraft trial of his mother

In Linz in 1620, Kepler heard that his mother had been indicted on the charge of being a witch. Such a defendant was often subjected to torture and, if convicted, was usually burned at the stake. If his mother had suffered this fate, Kepler's own status as imperial mathematician of the Holy Roman Empire and mathematician of Upper Austria might have been irreparably impaired. He rushed to her defense, therefore, not only out of filial devotion but also out of prudent self-interest. Only his skillful intervention saved her from torture and a fiery death.

Kepler had planned to publish his *Tabulae Rudolphinae* (*Rudolphine Tables*), named in honour of his first imperial patron, Rudolph II, in Linz. But this work, the final outcome of long years of unceasing reflection and tireless calculation, could not be printed there because of a rebellion by the peasants, who were infuriated by a combination of being forced to return to Catholicism and to pay heavy taxes. Kepler had to find another home and a new patron.

Albrecht von Wallenstein, duke of Friedland and Żagań, a successful soldier of fortune who had put his private army at the disposal of the empire in the Thirty Years' War, accepted the responsibility of satisfying Kepler's financial needs. The astronomer moved to Żagań in Silesia, where he was able to establish his own printing press. The *Rudolphine Tables* were printed at Ulm, Germany, in 1627, before Kepler went to Żagań in 1628. But Wallenstein turned out to be someone on whom Kepler could not rely.

Leaving his family behind in Żagań, Kepler went west to collect the interest due on two promissory notes he held in exchange for money he had deposited in Austria. On his way he stopped at Regensburg, where the Imperial Diet was in session. He fell acutely ill and died on November 15, 1630. The tremendous upheavals suffered by Germany in the Thirty Years' War later obliterated his grave.

BIBLIOGRAPHY. MAX CASPAR (comp.), *Bibliographia Kepleriana*, 2nd ed. by MARTHA LIST (1968), a complete enumeration of the publications by and about Kepler; and ARTHUR BEER and PETER BEER (eds.), *Kepler: Four Hundred Years: Proceedings of Conferences Held in Honor of Johannes Kepler* (1975), covering the years 1967–75 and including Martha List's supplement to Max Caspar's 2nd ed. of the *Bibliographia*. The best biography is MAX CASPAR, *Kepler* (1959, reissued 1962; originally published in German, trans. from the 3rd ed. of 1958). For Kepler's works, see JOHANNES KEPLER, *Gesammelte Werke*, ed. by WALTHER VON DYCK and MAX CASPAR, 19 vol. (1937–81), a magnificent edition; JOHANNES KEPLER, *Opera Omnia*, ed. by CHRISTIAN FRISCH, 8 vol. (1858–71, reprinted 1971), still a useful collection; *Epitome of Copernican Astronomy*, bk. 4–5; and *Harmonies of the World*, bk. 5, trans. by CHARLES GLENN WALLIS in "Great Books of the Western World," vol. 16 (1952). Other

English-language selections include CAROLA BAUMGARDT, *Life and Letters*, with an introduction by ALBERT EINSTEIN (1951); JOHN LEAR (ed.), *Kepler's Dream: With the Full Text and Notes of Somnium, sive Astronomia Lunaris, Joannis Kepleri*, trans. by PATRICIA F. KIRKWOOD (1965); *The Six-Cornered Snowflake*, ed. and trans. by COLIN HARDIE, with essays by L.L. WHYTE and B.F.J. MASON (1966), including parallel Latin and English texts and also containing numerous annotations and photographs; and *The Secret of the Universe*, trans. by A.M. DUNCAN, with an introduction by E.G. AITON (1981), and including parallel Latin and English texts. On Kepler's science, see ROBERT SMALL, *An Account of the Astronomical Discoveries of Kepler* (1804, reprinted 1963); EDWARD ROSEN (ed. and trans.), *Kepler's Conversation with Galileo's Sidereal Messenger* (1965), on Kepler's enthusiastic reaction to Galileo's spectacular discoveries with the telescope, and *Kepler's Somnium* (1967), on Kepler's imaginary space flight to the Moon. Specialized studies on Kepler by EDWARD ROSEN are "In Defense of Kepler," in ARCHIBALD R. LEWIS (ed.), *Aspects of the Renaissance: A Symposium*, pp. 141–158 (1967); "Galileo and Kepler: Their First Two Contacts," *Isis*, 57:262–264 (1966); "Kepler's *Harmonics* and His Concept of Inertia," *Am. J. Phys.*, 34:610–613 (1966); and "Kepler and Witchcraft Trials," *Historian*, 28:447–450 (1966). Other studies include ARTHUR KOESTLER, *The Watershed: A Biography of Johannes Kepler* (1960); ANGUS ARMITAGE, *John Kepler* (1966), a biography contrasting Kepler's thought with medieval superstitions; and ALEXANDRE KOYRÉ, *The Astronomical Revolution: Copernicus-Kepler-Borelli* (1973; originally published in French, 1961).

(Ed.R./Ed.)

Kiev

The capital of the Ukrainian Soviet Socialist Republic, a port on the Dnepr River, and a large railroad junction, Kiev is the third largest city of the Soviet Union and first among cities of the Ukraine. Kiev has an ancient and proud history. As the centre of Kievan Rus, the first Russian state, 1,000 years ago, it acquired the title "Mother of Russian Cities." It was severely damaged during World War II, but by the mid-1950s it was fully restored and by the 1970s had become a thriving, modern capital, with a well-developed economic and cultural life.

The article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

Site. The city stands on the Dnepr River just below its confluence with the Desna and 591 miles (951 kilometres) from its mouth in the Black Sea. The original location was on the high and steep right bank, which rises above the river in an imposing line of bluffs, culminating in Batyyeva Hill, 330 feet (100 metres) above mean river level. This precipitous and wooded bank, topped by the golden domes and spires of churches and bell towers and by modern high-rise apartment buildings, makes the city an attractive and impressive sight from across the Dnepr. Since World War II, Kiev has extended on to the wide, low, and flat floodplain on the left bank.

Climate. Kiev has a moderately continental climate. The average January temperature is 21.6° F (−5.8° C), and winter days with temperatures above freezing are not uncommon; in cold spells with a northerly or northeasterly airstream, temperatures may drop sharply, and an absolute minimum of −27.4° F (−33° C) has been recorded. Snow cover lies usually from mid-November to the end of March; on average, the frost-free period lasts 180 days but in some years surpasses 200 days. Summers are warm, with a July average of 67.1° F (19.5° C) and a recorded maximum of 102.2° F (39° C). The mean annual rainfall is 25 inches (625 millimetres), with maximum precipitation in June and July.

Layout. The city limits enclose an area of 300 square miles (780 square kilometres) on both banks of the Dnepr. It is divided into 12 administrative wards: Darnitsky, Dneprovsky, Leningradsky, Leninsky, Minsky, Moskovsky, Pechersky, Podolsky, Radyansky, Shevchenkovsky, Zaliznichny, and Zhovtnevy.

The Upper Town

The focus of Kiev is the area of the ancient Upper Town, crowning the high bluff of the Dnepr. Although very largely of postwar construction, this central area retains its old street pattern, and most of the surviving historical and architectural monuments are located there. First among these is the Cathedral of St. Sophia, now a museum. It was founded in the 11th century and remains, despite certain Baroque modifications in the 18th century, one of the finest and most beautiful examples of early Russo-Byzantine ecclesiastical architecture. It has a nave and four aisles, and it is crowned by five domes. The interior is magnificently decorated with frescoes and mosaics; it contains the tomb of Yaroslav, during whose reign the cathedral was built.

Close by is the Baroque Church of St. Andrew (Andreyevskaya), designed by Bartolomeo Rastrelli and built in the mid-18th century; its site on the crest of the steep slope to the river makes it a striking landmark. Other historical relics in the central area include the ruins of the Golden Gate, also built in the 11th century in the reign of Yaroslav; the Zaborovsky Gate, built in 1746–48; and the remains of the Desyatinnaya church, or Church of the Tithes, built in 989–996 by St. Vladimir.

Within and immediately adjacent to the area of the former Old Town are many of the city's museums, theatres, and public buildings as well as the principal shops, including the central department store and the covered market. The axis of the centre is the street known as Kreshchatik, which runs along the bottom of a small valley the sides of which have in part been landscaped with terraced gardens interspersed with tall, modern office and apartment buildings. The greenery of the gardens, the trees lining the street, the squares that it intersects—all combine with the variegated colours of brick, red and gray granites, and decorative ceramic tiles to give Kreshchatik an attractive and colourful aspect, much admired by Kiev's inhabitants. Among important buildings on the street is that of the city council, where the 800 elected deputies hold their meetings.

Intersecting Kreshchatik at right angles is the wide, poplar-lined Boulevard of Taras Shevchenko, on which stands the university with its eye-catching red-washed walls. There too is the Cathedral of St. Vladimir (still in use as a church), built in 1850–96 in Byzantine style and containing impressive paintings by Viktor Vasnetsov and other Russian artists. Notable among the many statues in central Kiev are those that commemorate the Cossack leader Bohdan Khmelnytsky and the Ukrainian poet Taras Shevchenko.

North of the old centre is the former trading and Jewish quarter, Podol, with a rectangular pattern of streets and the old merchants' trading exchange, the House of Contracts, built in 1817. There too is the river port. South of the centre is the Pechersky district, along the top of the river bank. This district contains many of the principal buildings of the Ukrainian republican government, including the glass-domed palace of the Supreme Soviet of the Ukrainian S.S.R., built in 1936–39, and the 10-story block that houses the Council of Ministers. Nearby is the attractive Mariinsky Palace, built in 1752–55 for the tsaritsa Elizabeth and reconstructed in 1870; it is now used for receptions by the Ukrainian government.

Pechersky district

At the southern end of this district is the Kievo-Pecherskaya Lavra (Monastery of the Caves), one of the most famous and important monasteries in Russian history, founded in the early 11th century. It was at the *lavra* that the monk Nestor wrote the earliest surviving Russian chronicle. Although the Cathedral of the Assumption (inside the walls of the monastery) was blown up in 1941, Trinity Church, of the same period, survives. Also within the walls are the 17th-century Church of All Saints and an impressive 18th-century bell tower, rising 315 feet. A major feature of the monastery is the system of catacombs beneath it in which are the mummified bodies of early monks and saints, including that of Nestor. Although a museum open to the public, the Pecherskaya Lavra is still in use as a monastery.

South from the *lavra* is yet another monastery, the Vydubetsky, dating from the 11th century; it too was severely damaged in World War II.

All along the steep river bank, fronting the Upper Town and Pechersky district, an attractively landscaped park has been laid out overlooking the Dnepr. With the views it affords, the park forms one of the most striking features of the city. It contains an open-air theatre, sports stadium, and restaurant, and a funicular railway climbs the 300-foot slope. Within the park are also many memorials. Dominating the northern end is the statue of Prince Vladimir, who brought Christianity to Russia, the statue that marks the place where in 988 the people of Kiev were baptized en masse. The southern end, called the Park of Glory, has an 85-foot granite obelisk rising above the grave of the Unknown Soldier and a memorial garden. Also located in the park are the grave of Gen. Nikolay Vatutin, commander of the Soviet forces that liberated Kiev in 1943, and a rotunda marking the supposed grave of the early Varangian chief Askold.

Suburbs

Around these central districts of Kiev stretch extensive suburbs of factories and residential neighbourhoods. The low priority given to housing during the Stalin period means that the greater part of these suburbs has been built since 1956. The neighbourhood units, known as microregions, consist of groupings of apartment buildings housing 2,500 to 5,000 people, together with basic services, local shops, a health centre, cinema, and primary school. Since the late 1960s the apartment buildings have usually been of 12 to 20 stories and of prefabricated construction. Most apartments have only two or three rooms, and population densities are therefore high, in the new residential developments as much as in the older central areas. The growing ownership of private cars poses problems in the provision of garage space in these new districts. A feature of development since World War II has been the rapid spread of the city on the low left bank of the Dnepr, previously almost devoid of settlement. The left bank is linked to the main part of Kiev by a railway bridge and by the imposing Ye.O. Paton road bridge, which is 4,920 feet long and named after its designer.

Between the neighbourhood units are substantial areas of park and green space. These include the very large botanical gardens of the Ukrainian Academy of Sciences, the smaller university botanical gardens (established in the mid-19th century), and in the southwestern suburbs the extensive permanent exhibition of the Ukrainian economy. On the city outskirts are several areas of forest, which are much used for recreation. In the south is the Goloseyevsky Forest Park, dominated by deciduous trees, and to the north are nearly 10,000 acres (4,450 hectares) of the Pushcha-Voditsa Forest Park, mainly covered by coniferous species.

Soviet cities on the whole tend to a certain monotony of appearance. A number of factors combine to make Kiev an exception to this rule and one of the most attractive urban places in the Soviet Union—the site, with its sharply contrasted relief and wide views across the Dnepr, the abundance of greenery in and around the city, and the many buildings of historic interest and beauty.

Anatoly Poddubny—Tass/Sovfoto

The Bohdan Khmelnytsky Square in Kiev. In the foreground is the Cathedral of St. Sophia, now a museum.

THE ECONOMY

Industry. Kiev, as capital of the Ukraine, has major administrative functions, with considerable employment in the offices of ministries responsible for the republic's economy. The city is also an important industrial centre, possessing a wide range of manufactures. Factories are found in all quarters of the city, with major concentrations to the west of the city centre and on the Dnepr left bank.

Engineering industries, based on metal from the iron and steel plants of the Dnepr bend region and the Donbass coalfield, take pride of place and include the production of complex machinery and precision tools and instruments. The Bolshevik plant makes equipment for chemical works, such as conveyor lines for vulcanized rubber, linoleum, and fertilizer factories; the Gorky works producesmetal-cutting machines. Other engineering products are aircraft, hydraulic elevators, electrical instruments, armatures, river- and seagoing craft, motorcycles, and cinematography apparatus.

Chemical industry

Another important sector is the chemical industry, making resin products, fertilizers, plastics and chemical fibres, the last at the Darnitsa viscose rayon plant on the left bank. Timber working and the making of bricks and reinforced concrete items are also well developed. Consumer goods manufactured include cameras, thermos flasks, knitwear, footwear, a range of foodstuffs, and watches. Kiev is also a large publishing centre.

Power for the many enterprises is supplied by natural gas, piped from Dashava in the western Ukraine, and by electricity from the Kiev hydroelectric station on the Dnepr. This station, completed in 1968, is at Vyshgorod, just upstream of the city. Twenty-five miles southeast of Kiev is the still more powerful Tripolye thermal electric station.

Transportation. Transportation for the industries and for the city as a whole is provided by a good communications network. Trunk railways and all-weather roads link Kiev to Moscow, to Kharkov and the Donets Basin (Donbass), to the southern Ukraine and the port of Odessa, and to the western Ukraine and Poland. The navigability of the Dnepr has been improved by a series of barrages and reservoirs. Borispol airport operates direct flights to most major Soviet cities and to many Ukrainian towns, as well as some international connections to Romania and Bulgaria. Within Kiev itself there is efficient subway and rail, bus, streetcar, and trolleybus service.

SOCIAL AND CULTURAL LIFE

Kiev's ancient tradition as a cultural centre is still vigorously alive. The Kiev State University heads an array of 20 institutions of higher education, notable among which are the Polytechnic (founded in 1898), the Agricultural Academy, and the medical, art, and architectural institutes.

There is a large number of general secondary schools, evening schools for adults, and specialist technical schools. A range of research establishments is headed by the Academy of Sciences of the Ukrainian S.S.R., established in 1919, which also maintains the largest of the city's many libraries. Kiev is particularly noted in the Soviet Union for medical and cybernetic research. The emphasis on applied research is illustrated by the academy's Ye.O. Paton Institute of Electrical Welding.

There are several theatres, notably the Shevchenko Theatre of Opera and Ballet. Plays are presented at the Lesya Ukrainka and Ivan Franko theatres, which specialize in Russian and Ukrainian drama, respectively; drama is also frequently staged in the 4,000-seat auditorium of the Palace of Culture and in the Palace of Sport, which can seat 12,000 people. In addition there are youth, open-air, and musical comedy theatres. Kiev has a circus and more than 130 cinemas; films are made in a studio in the city. Concerts are regularly given at the Tchaikovsky Conservatory. The most important of the city's many museums are the Kiev State Historical Museum, the Kiev State Museum of Russian Art, and the Kiev State Museum of Ukrainian Art.

Sports

Kiev has good facilities for sports; the largest of its 15 stadiums, the Central Stadium, can accommodate 100,000 people. Aquatic sports take place on the reservoir of the Kiev dam at Vyshgorod and also on Trukhanov Island in the Dnepr opposite the city centre, where there is a fine beach and water sports centre. The city is well provided with health facilities, including general and specialized hospitals and local polyclinics, the latter serving residential neighbourhoods. Since the majority of women are employed, a number of nursery schools and crèches care for children below school age. Around the outskirts of Kiev are health resorts, sanatoriums, and children's holiday camps.

History

THE EARLY PERIOD

Origins and foundation. Kiev has a long, rich, and often stormy history. Its beginnings are lost in antiquity. Archaeological findings of stone and bone implements, the remains of primitive dwellings built of wood and skins, and large accumulations of mammoths' bones indicate that the first settlements in the vicinity date from the Late Paleolithic Period (some 15,000 to 40,000 years ago). As early as 3000 BC in the Neolithic Period and subsequently at the time of the Cucuteni-Tripolye culture at the end of the Neolithic, tribes engaged in agriculture and animal husbandry lived on the site of modern Kiev. Excavations continue to uncover many artifacts from settlements dating from the Copper, Bronze, and Iron ages. The tribes of the area traded with the nomadic peoples of the steppe to the south, Scythians, Sarmatians, and later Khazars, and also with the ancient Greek colonies that were located on the Black Sea coast.

According to the 12th-century chronicle *Povest vremennykh let* ("Tales of Bygone Years," also known as *The Russian Primary Chronicle*), Kiev was founded by three brothers, Kiy, Shchek, and Khoriv, leaders of the Polyane tribe of the East Slavs. Each established his own settlement on a hill, and these became the town of Kiev, named for the eldest brother, Kiy; a small stream nearby was named for their sister Lybed. Although the chronicle account is legendary, there are contemporary references to Kiev in the writings of Byzantine, German, and Arab historians and geographers. Archaeological evidence suggests that Kiev was founded in the 6th or 7th century AD.

The first Russian capital. Less legendary is the chronicle account of the Varangians, who seized Kiev in the mid-9th century. As in Novgorod to the north, a Russo-Varangian ruling elite developed. Kiev, with its good defensive site on the high river bluff and as the centre of a rich agricultural area and a group of early Russian towns, began to gain importance. In about 882 Oleg (Ukrainian Oleh), the ruler of Novgorod, captured Kiev and made it his capital, the centre of the first Russian state, Kievan (or Kiev) Rus. The town flourished, chiefly through trade along the Dnepr, going south to Byzantium and north over portages to the rivers flowing to the Baltic, the so-called "road from the Varangians to the Greeks," or "water road." Trade also went to the Caspian Sea and Central Asia.

Christianization

In 988 the introduction of Orthodox Christianity to Kiev enhanced its significance as the spiritual centre of Rus. By the 12th century, according to the chronicles, the city's wealth and religious importance was attested to by its more than 400 churches. The Cathedral of St. Sophia, parts of the Kievo-Pecherskaya Lavra, and the ruins of the Golden Gate remain today as witnesses to Kiev at the height of its splendour. The town was famed for its art, the mosaics and frescoes of its churches, its craftsmanship in silver, and the quality of many of its manufactures. One of Europe's major cities, Kiev established diplomatic relations with Byzantium, England, France, Sweden, and other countries. Travellers wrote of its population as numbering tens of thousands.

Throughout the period of Kievan Rus, however, the city was engaged in a succession of wars against the nomadic warrior peoples who inhabited the steppes to the south, in turn the Khazars, Pechenegs, and Polovtsy (Kipchaks). These conflicts weakened the city, but even greater harm was done by the endless, complex internecine struggles of the Russian princedoms into which Rus was divided. In 1169 Prince Andrew Bogolyubsky of Rostov-Suzdal captured and sacked Kiev. Thus by the late 12th century the power of the city had declined, and in the following century it was unable to resist the rising and formidable power of the Mongols. In 1238 a Mongol army under Batu, grandson of Genghis Khan, invaded Rus and, having sacked the towns of central Russia, in 1240 besieged and stormed Kiev. Much of the city was destroyed and most of its population killed. The Italian traveller Giovanni da Pian del Carpini six years later reported only 200 houses surviving in Kiev.

Kiev under Lithuania and Poland. In the 14th century, what was left of Kiev and its surrounding area came under the control of the powerful and expanding Grand Duchy of Lithuania, which captured it in 1362. For a long time thereafter Kiev had little function except as a fortress and minor market on the vaguely defined frontier between Lithuania and the steppe Tatars, based in the Crimea. It frequently came under attack from the Tatars; in 1482 the Crimean khan, Mengli Giray, took and sacked the town. Almost the only survival of Kiev's former greatness was its role as the seat of an Orthodox metropolitan. A step forward came in 1516, when the grand duke Sigismund I granted Kiev a charter of autonomy, thereby much stimulating trade.

In 1569 the Union of Lublin between Lithuania and Poland gave Kiev and the Ukrainian lands to Poland. Kiev became one of the centres of Orthodox opposition to the expansion of Polish Roman Catholic influence, spearheaded by vigorous proselytization by the Jesuits. In the 17th century a religious Ukrainian brotherhood was established in Kiev, as in other Ukrainian towns, to further this opposition and encourage Ukrainian nationalism. Peter Mogila (Mohyla), a major theologian and metropolitan of Kiev from 1633 to 1646, founded there the Collegium (later the Academy of Kiev), which became a major focus of the struggle with Roman Catholicism.

The Cossack state

In the 17th century there was also increasing unrest among the Zaporozhian Cossacks of the Dnepr downstream of Kiev and an ever-growing struggle between them and the Polish crown. This culminated in the revolt of Bohdan Khmelnytsky, who, assisted by the Crimean Tatars, entered Kiev with his insurgent Cossacks in 1648. Under heavy pressure from the Polish forces, in 1654 Khmelnytsky and the Cossacks offered their allegiance to Moscow (the Pereyaslav Agreement); this was followed by a prolonged and confused period of strife and destruction, leading in 1667 to the Treaty of Andrusovo, by which Kiev and the Dnepr left bank part of the Ukraine became an autonomous Cossack state under the suzerainty and protection of Moscow. Thereafter further struggle ensued against the Turks, with the Cossacks constantly changing sides and engaging in internecine disputes. In 1686 Kiev was finally yielded to Russia by Poland and stood as the sole Russian outpost on the right bank of the Dnepr.

EVOLUTION OF THE MODERN CITY

Kiev under the tsars. The Second Partition of Poland in 1793, under Catherine the Great, brought the right bank Ukraine into Russia, and Kiev, assisted by the abolition in 1754 of the tariff barriers between Russia and the Ukrainian lands, began to grow in commercial importance. Catherine's reign was marked by the abolition of the old administrative system and of the post of Cossack hetman and the division of the Ukraine into new administrative provinces, for one of which Kiev became the centre. Subsequently it became the centre of a governor generalship, covering three provinces.

In the first half of the 19th century, Kiev developed as a major focus of Ukrainian nationalism, although severe persecution from the tsarist government forced the movement to shift the brunt of its activities to Lvov in the Austrian Ukraine. In Kiev, as in other Russian cities, there was clandestine revolutionary activity (beginning with the Dekabrists in the early 19th century) that culminated in a series of strikes and demonstrations leading to the revolution of 1905. An important role in this revolutionary movement was taken by students of the University of Kiev (now Kiev T.G. Shevchenko State University), which had been established in 1834.

In the 19th century the expanding economic importance of the Ukraine, and especially the growing export of grain, brought further commercial development to Kiev. Modern factory industry appeared; to the Arsenal, set up as early as the 18th century, were added timber working and the building of river craft. The town had significant industries processing agricultural products—leather, tobacco, distilling, brewing, and textiles. In the late 1860s Kiev was connected by rail to both Moscow and the Black Sea port of Odessa, further enhancing its role as a centre of industry, commerce, and administration. By the outbreak of World War I, the city had a population of some 350,000.

The revolutionary period. With the outbreak of the Revolution in 1917, a revolutionary soviet, the Central Rada (*rada,* "council"), was elected by the city workers, consisting primarily of Menshevik and Social Revolutionary members, with strong support from Ukrainian nationalist groups. In January 1918 the Rada proclaimed an independent Ukrainian state with Kiev as its capital. Minor uprisings by Bolshevik workers, who were mostly concentrated in the Arsenal works, were suppressed, but Red Army troops came to their aid and on February 8, 1918, entered Kiev.

Petlyura government

By the Treaty of Brest-Litovsk (March 3) between the Bolshevik government and the Germans, however, the new Soviet government recognized the independence of the Ukraine, which was promptly occupied by German troops. A puppet Ukrainian government was set up in Kiev by the Germans, but it collapsed with the German surrender to the Allies in November 1918 and the subsequent withdrawal of German troops. Once more an independent Ukraine was declared in Kiev, under the leadership of Simon Petlyura, but its brief and stormy history was a series of struggles between Ukrainian nationalist, White, and Red forces. In November 1919 Kiev was briefly taken by the White armies under Gen. A.I. Denikin before being finally occupied by the Red Army. Peace was still denied the city, with the outbreak of the Russo-Polish War. In May 1920 the Poles captured Kiev but were driven out in a counterattack.

The Soviet period. Kiev's role as the centre for Ukrainian nationalists caused the Soviet government to transfer the capital of the new Ukrainian Soviet Socialist Republic to Kharkov, and it was not until 1934 that Kiev resumed its capital status. Meanwhile, restoration of the city's shattered economy was undertaken. During the first Five-Year Plans, between 1928 and 1940, new machine tool, electrical, and chemical industries were established. By 1939 the population had reached 846,724. The German invasion in 1941 again brought severe suffering and destruction to the city. After a fierce 80-day battle, German forces entered it on September 19, 1941. More than 30,000 Jews, Soviet prisoners of war, and partisans who had remained in the city were massacred within days in a nearby ravine known as Baby Yar. Many of Kiev's other inhabitants were deported for forced labour and to concentration camps, including almost all the large prewar Jewish segment. In 1943 the advancing Soviet troops forded the Dnepr and, after bitter fighting, liberated Kiev on November 6. The city itself had suffered great destruction, including more than 40 percent of its buildings and some 800 of its industrial enterprises.

For its role in the war Kiev was later honoured by the Soviet government with the Order of Lenin, the title of Hero-City, and the Gold Star medal. In the first postwar Five-Year Plan, rapid reconstruction was undertaken. Since then Kiev has continued to grow and to strengthen its industrial base.

BIBLIOGRAPHY. О.К. КАСИМЕНКО (ed.), *Історія Києва,* 2 vol. (1959–60), a thorough historical study based on archaeological evidence as well as documentary sources; І.П. СТАРОВОЙТЕНКО, *Соціалістичний Київ* (1958), covers changes in economic, social, and cultural life during the Soviet period; also in Russian is the encyclopaedic handbook edited by А.В. КУДРИЦКИЙ, *Киев: Энциклопедический справочник* (1982); LEONID DAEN, PAVEL POZNYAK, and MARK CHERP, *Kiev: Travel Guide* (1970), a well-documented guide in English, and *Kiev Travel Guide,* published by the Novosti Press Agency Publishing House, is also in English.

(R.A.F./Ed.)

Kinshasa

Kinshasa, the capital of the Republic of Zaire, lies about 320 miles (515 kilometres) from the Atlantic Ocean on the south bank of the Congo (locally called the Zaire) River. One of the largest cities of sub-Saharan Africa, it is a special political unit equivalent to a Zairian region, with its own governor. The city's inhabitants are popularly known as Kinois.

Kinshasa is distinctive not only as the capital of Zaire but also as the centre of the dynamic and contradictory influences that have shaped the country's character in modern Africa. The only city not clearly identified with any particular region of the country, it is the seat of a long-lasting military government based, on the one hand, on the strength of the armed forces and, on the other, on a technique of political and social compromise that has gained the rather grudging collaboration of most of the citizens. Caught between spectacular wealth and massive poverty, most Kinois must spend a considerable amount of their time scrambling for necessities that are in erratic supply. Nevertheless, they have found the means to make Kinshasa a source of distinctive influence in intellectual and popular culture that has been felt throughout Africa.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site and climate. Kinshasa spreads out southward from the shoreline of the Congo River at Malebo Pool, a widening of the river. The plain on which the city lies varies mostly between 918 and 1,148 feet (280 and 350 metres) above sea level and is partly encircled by higher ground. The most heavily inhabited area of Kinshasa covers 58 square miles (about 150 square kilometres). The total area subject to city government, much of it sparsely populated, is 3,847 square miles.

The climate is hot year-round, with a dry season from May to September and a rainy season from October to May. The mean annual rainfall is slightly more than 60 inches (1,524 millimetres). Violent rainstorms occur frequently but seldom last more than a few hours. The hottest month is April, with mean daily maximum and minimum temperatures of 89° F (32° C) and 71° F (22° C), respectively. The corresponding figures for July, the coldest month, are 81° F (27° C) and 64° F (18° C). The higher suburbs are somewhat cooler than the central city. The surrounding countryside is heavily farmed savanna and gallery forest; the chief crops are cassava, sugarcane, oil palms, plantains, corn (maize), peanuts (groundnuts), and beans.

The city layout. The built-up area of Kinshasa is divided into industrial, residential, and commercial zones. Along the western edge of the central city an industrial zone (before 1966 called Léo-Ouest) flourishes near the site of the first depot established by Sir Henry Morton Stanley. To its east lies the riverside residential and administrative district of Gombe, which houses most of the European population and the Zairian elite; the central government buildings and the embassy district are located there. The eastern sector (known before 1966 as Léo-Est), of which the wide Boulevard du 30-Juin forms the main artery, is a major commercial area. The waterfront, along Kinshasa's northern edge, is lined with quays and large warehouses. Ndolo, east of Gombe, comprises a complex of port facilities and industrial plants. The poorer areas extend southward on the east and west of Kinshasa. Among Kinshasa's satellite cities, Ndjili, to the southeast, has become a residential area, while Kimpoko upstream has been developed as an outer port. During the 1970s, wealthy businessmen and politicians built mansions, often of spectacular opulence, in Binza, an area in the western hills overlooking the city.

Government and residential section

There are a variety of architectural styles in Kinshasa. High-rise apartment blocks, luxuriously appointed banks, stores, and the offices of large corporations and government agencies characterize the centre of town. Some date from shortly after World War II, but the most prominent were constructed during the economic boom of the early 1970s. They include the Parliament, the president's palace (built by the Chinese), the headquarters of the national broadcasting corporation (the Voice of Zaire), the International Trade Center, the headquarters of the mineral marketing agency, and the unfinished tower dedicated to Patrice Lumumba. Spacious villas surrounded by ornamental shrubs and flower gardens, and often also by high walls and iron bars, stand on tree-lined, paved boulevards that mark the elite residential districts. Dwellings in the less affluent communities often consist of tin-roofed, concrete-block houses and multi-unit dwellings on unpaved streets, often inaccessible to vehicles. These in turn give way to hastily assembled shelters and rough pathways in the extensive squatting zones of the city, where many of the most recent immigrants reside.

Architectural landmarks

THE PEOPLE

The population of Kinshasa grew slowly at first (from 5,000 people in 1889 to 23,000 in 1923) but increased rapidly after 1940; after 1950 it doubled about every five years and by the 1980s was estimated to be more than 3,000,000, of whom about one-third lived in the squatting zones. Much of the population growth has been the result of Zairian migration and government expansion, but widening of the city's boundaries has caused some of the increase. Kinshasa has a young population. More than half the people are under 22 years of age, and only about 3 percent of the population is over 50.

Migration of peoples from the rural areas intensified greatly after independence as colonial restrictions were relaxed. Political troubles and the economic decline of rural areas and their lack of amenities and opportunities, as well as the attractions of the city, have contributed to this rural exodus. In its early years the city received immigrants from West Africa and the various neighbouring countries of Central Africa; since independence, however, most new inhabitants have come from within Zaire, especially the nearby regions of Bandundu, to the west, and Bas-Zaïre (Lower Zaire), to the south and east.

Population growth through migration

THE ECONOMY

Industry and commerce. Kinshasa is the most important consumer centre of the republic and the core of its industrial and commercial activity. The city serves as the headquarters of major public corporations and of privately owned industrial and commercial companies. It dominates the financial and commercial life of the republic and

Dynamic centre of Zairian commerce and industry houses the head offices of the principal banks. Among Kinshasa's main industries are food processing, and those producing beer, textiles, footwear, woodwork, paper, packing cases, tires, metalwork, tobacco, and chemicals. Construction and various service industries also contribute to the city's economy.

The rapid expansion of the city's population has created serious problems of food supply; there is a constant threat of shortages, posing an implicit political problem as well. The busy central market is complemented by suburban markets lined with wooden stalls and by hawkers and street vendors selling in minute quantities to passers-by. The region of Bas-Zaïre supplies at least half of the food consumed in the capital. Other foodstuffs come from more distant regions of Zaire or are imported. For those who can afford it, South Africa has been an important source of meat and fruit and vegetables, which are flown in. For the poor, however, Kinshasa is in some ways like an overgrown village, whose people forage at a considerable distance for firewood and keep gardens where they can find good soil. The demands of this vast urban population have caused extensive erosion in the surrounding countryside, as the soil is exhausted from overcultivation and trees cut for charcoal have not been replanted.

Transportation. Kinshasa's transportation system is inadequate in many respects. Economic problems and shortage of foreign exchange have caused severe deterioration, and there has been a continual crisis for lack of spare parts and replacement vehicles. Kinshasa is well served by roads, but its dense and rapidly increasing population causes much congestion. The city is connected by a paved road to Matadi, Zaire's principal port at the head of navigation on the Congo estuary, and by another to Kikwit, to the east. The railway line from Matadi, bypassing the rapids on the river below Kinshasa, brings in most of the country's imports, some of which are then conveyed upriver. The Congo is navigable to Kisangani, 1,050 miles upstream, and a vast network of navigable stretches on its tributaries, connected by railways, brings almost all inland traffic carrying exports destined for Matadi down the Congo and through the port of Kinshasa. Ndjili International Airport, to the southeast, is one of Africa's largest air stations. A busy ferry connects the city to Brazzaville, the capital of the People's Republic of the Congo, across Malebo Pool. Within Kinshasa public transportation consists of grossly overcrowded buses, minibuses, taxis, and *fula-fula* (trucks adapted to carry passengers).

The port of Matadi

ADMINISTRATION AND SOCIAL CONDITIONS

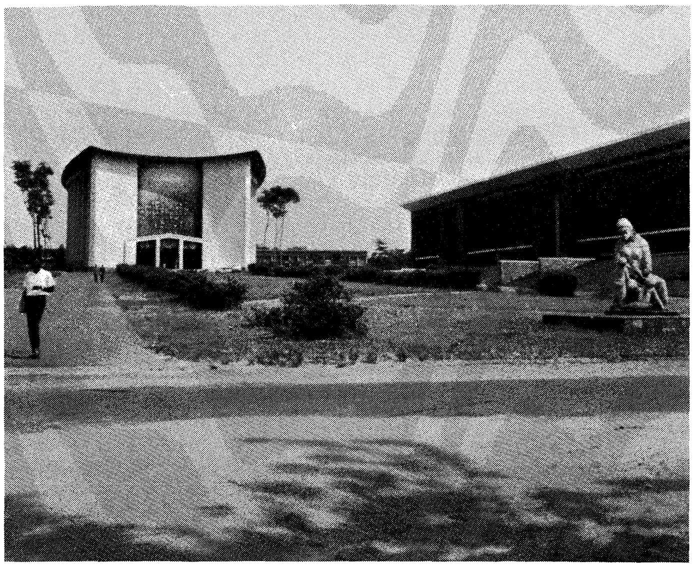
Government and services. Despite the government's declared policy of decentralization, the capital remains the place where all administrative decisions of importance are made; consequently, it is the centre of the nation's political life. The city houses the national government: the office of the president, the Political Bureau (the highest organ of the single party, the Mouvement Populaire de la Révolution [MPR]), the Central Committee of the party, and the executive and legislative councils. Since 1982 the urban administration has consisted of a governor and two vice-governors, appointed by the president. They head the city council, consisting of the 24 zone commissioners appointed, also by the president, from among the councillors elected in each zone. Despite a pervasive sense of tension, political disturbances have been few.

City government structure

The administration is unable to adequately provide such services as running water, electricity, and sanitation throughout the city; town-planning and building-control agencies have had difficulty coping with the rapid growth of the city, much of which consequently lacks basic urban facilities. Some areas suffer from eroded housing lots and roadways, clogged open drains, and accumulated trash. Although there has been some increase, the rate of violent crime is relatively low.

Health and education. Medical facilities, like other city services, are overwhelmed by the expansion of the population. The hospitals, clinics, and dispensaries of the public health system are insufficient in number and unevenly distributed, which, coupled with the problems of transportation, limits the health care they can provide for the public.

Health and education problems


Campus of the Université de Kinshasa.
Michel Huet—HOA-QUI

The primary and secondary education system is similarly overextended, lacking sufficient facilities and teachers to cope with population growth. Institutions of higher education include the Université de Kinshasa (formerly Lovanium University), the largest of the country's three universities; a teacher-training college; a national school of administration and law; a school of telecommunications; and an academy of fine arts, as well as institutes of social research, political party indoctrination, medical training, and commerce. Kinshasa's School of Catholic Theology is internationally distinguished.

CULTURAL LIFE

Kinshasa is the dynamic centre of the nation's popular culture, of which the language is Lingala, the urban lingua franca. Zaire's popular music is renowned throughout Africa; well-established bands from the country tour abroad to Europe and the Americas; popular music celebrities receive wide attention, and it is not unusual for their latest hit song to give their names to fashions in women's dress materials, a medium of intense social competition. Like the popular songs, paintings sold on sidewalks express the social themes of the day. Daily papers, which are closely censored, and several periodicals are available to the populace. Television is an important medium of official communication to the people; it broadcasts news, presidential speeches, a form of propaganda entertainment called "animation," popular bands, and occasional old European films. Radio and television broadcasting is in French, the official language, and local languages.

Radio and television

Modern Kinshasa has produced a considerable flowering of literature in novels, plays, and poetry by local writers. Painting and sculpture produced by artists of the Académie des Beaux-Arts are exhibited and sold at the academy. The collection of the Institut des Musées Nationaux is of great archaeological, ethnographic, and musicological as well as aesthetic interest, and it is of immense importance for scholars of traditional African art. Although traditional art of value may no longer be available in the city, workshops in the suburbs turn out imitations of masks and sculptures that represent all parts of Africa, as well as carved work in ivory and malachite. The city is known for some excellent restaurants and is the site of numerous nightclubs and motion-picture theatres.

History

The land on which Kinshasa grew was inhabited in ancient times, as were all the shores of Malebo Pool. The present city evolved from two villages, Nshasa and Ntamo (later known as Kintamo), dominated by the Bahumbu and frequented by Bateke fishermen and traders. The explorer

Sir Henry Morton Stanley, on his visit in 1877, formed an alliance with the ruler of Kintamo, a wealthy ivory trader, and, despite French efforts to forestall him, was able to acquire a trading post site on his return in 1881. He named this post Léopoldville, after his patron, Léopold II, king of the Belgians. Although Stanley succeeded in opening river traffic as far north as Stanleyville (Kisangani since 1966) by portaging prefabricated steamers around the cataracts of the lower river, Léopoldville remained unimportant until the railway line from downstream Matadi was completed in 1898. A pipeline from Matadi to carry crude oil to the upriver steamers at Léopoldville was completed in 1914, and an air service was inaugurated between Léopoldville and Stanleyville in 1920. As a result, the administrative headquarters of the then Belgian Congo was transferred there from Boma in 1923.

Rail and pipelines spur growth

As industries were established, residential zones grew up around them. In the 1930s the zones of Kinshasa, Barumbu, and Lingwala grew up near the port. After 1950 Lemba, Matete, and parts of Ndjili, to the southeast, were built to house the workers of the new industrial district of Limete, but the more centrally located communes (now zones) of Dendale (now Kasa-Vubu), Bandalungwa, and Ngiri-Ngiri became the social and political heart of the city. In 1960 Léopoldville became the capital of the new republic. Its name was changed to Kinshasa in 1966.

BIBLIOGRAPHY. A full description of Kinshasa is contained in MARC PAIN, *Kinshasa: Écologie et organisation urbaines,* 3 vol. (1978–79); the Republic of Zaire's *Atlas de Kinshasa,* issued by the INSTITUT GÉOGRAPHIQUE DU ZAÏRE, BUREAU D'ÉTUDES D'AMÉNAGEMENTS URBAINES (1975), contains an excellent descriptive text supporting the 44 thematic maps. Works of historical significance include HENRY M. STANLEY, *The Congo and the Founding of Its Free State,* 2 vol. (1885, reprinted 1970); and CRAWFORD YOUNG, *Politics in the Congo: Decolonization and Independence* (1965). JEAN S. LA FONTAINE, *City Politics: A Study of Leopoldville, 1962–63* (1970), describes the city's social organization; and MBUMBA NGIMBI, *Kinshasa, 1881–1981: 100 ans après Stanley: Problèmes et avenir d'une ville* (1982), describes its administrative history and problems. KANKUENDA M'BAYA, *Les Industries du pôle de Kinshasa* (1977), discusses the city's economic role.

(J.O.A./W.MacG./J.MacG.)

Korea

Extending southward from the Chinese historic area of Manchuria (now the provinces of Liaoning, Kirin, and Heilungkiang) and the U.S.S.R. on the northeastern Asian mainland, the peninsula of Korea is approximately 600 miles (970 kilometres) long and from 125 to 200 miles (200 to 320 kilometres) wide. It reaches southwest to within approximately 120 miles of Honshu, the principal island of Japan, and to within approximately the same distance of the Shantung Peninsula of China. Elongated and irregular in shape, the Korean peninsula separates the Yellow Sea and the Sea of Japan (called the East Sea in Korea). There are about 3,500 islands, mainly along the southern and western coastline.

The name by which Korea is best known to its own people is Chosŏn, which may be translated as "land of the morning calm." The Republic of Korea (south) uses Tae Han for its official name; the Democratic People's Republic of Korea (north) continues to use the name Chosŏn. The western name, Korea, was derived from the Koryŏ dynasty (AD 918 to 1392) and may be literally translated as "high and beautiful."

Because of its strategic location between China, Japan, and the Soviet Far East, Korea has long suffered from the inroads of aggressive neighbours. At the end of World War II, Korea was divided by an artificial barrier, the 38th parallel, for the purpose of accepting the surrender of Japanese troops. This division was perpetuated by the U.S. occupation of the south and the U.S.S.R. occupation of the north. In 1948 the Republic of Korea was formed in the south, and the Democratic People's Republic of Korea came into being in the north. Aggression from the north in June 1950 brought on the Korean War, with resultant devastation to many parts of the peninsula. Since the truce of 1953, reconstruction and development have taken place. The peninsula is still divided by a demilitarized zone that cuts across the land in a wavy line from north of the 38th parallel in the east to south of it in the west.

The article is divided into the following sections:

Physical and human geography

Korea is a mountainous land of diverse geology. The northern interior forms a broad continental base; it is similar geologically to the adjacent areas of Manchuria. The drainage divide of central Korea lies in the eastern part of the peninsula: a high, mountainous terrain composed chiefly of Archeozoic (Early Precambrian) rocks—granites, gneiss, mica schist, and other metamorphics. The southeastern part of the peninsula, of considerably less relief, is composed largely of Cretaceous and Tertiary sedimentary rocks interspersed by Late Tertiary granitic intrusions. The latest major mountain-building episode in Korea appears to have occurred during Jurassic time, all older rocks undergoing the deformation. The latest uplifts have determined the orientation of the peninsula.

Although volcanism is recorded in the dike rocks and lava flows associated with both the geosynclinal sedimentaries and later Mesozoic deposits, it has not been pronounced in more recent geologic time. Korea has no active volcanoes and very rare earthquake shocks, a striking contrast with nearby Japan.

Archaeological, linguistic, and legendary sources support the view that the Korean peninsula was settled by tribes, Tungusic in origin, that migrated in waves from Manchuria and Siberia. They settled along the coasts and moved up the river valleys. These people formed the dominant racial stock of the Korean people and were the originators of the Korean language. Physically, the Koreans are Mongoloid, like the Chinese and Japanese, and share many of their characteristics. Chinese cultural influence is obvious in much borrowed vocabulary as well as in the written language; Chinese was the classical language for many centuries.

(Ed.)

History

Cultural ancestry

The Korean people share a common racial origin with other peoples of North Asia, and the Korean language belongs to the Altaic language family of the region. There was a close relationship between Korean culture and that of neighbouring peoples in the Neolithic Period and the Bronze Age. For example, Korean combware pottery, widely used in the Neolithic Period, is commonly found in the region; Korean bronze daggers, belt hooks, and multi-knobbed mirrors also display the traits of bronze tools unearthed in the area. In this manner, the lineage and formation of early Korean culture can be roughly traced. One branch of the cultural ancestry went as far as the Shantung Peninsula and formed an element of ancient Chinese culture; another went southward to Japan to form the main current of early Japanese culture.

Because of this background, the early Korean people were in constant confrontation with the Chinese. China was antagonistic toward the northern peoples, with whom the Koreans were associated, and naturally assumed an aggressive attitude toward the Koreans. Thus, Korea's formation and development were nurtured in its struggles against China.

After the unification of the Korean peninsula by Silla, Korea maintained friendly relations with China out of the need to deter invasions by the northern nomadic peoples. Later, Manchuria and Mongolia came under the successive control of the northern nomadic peoples and served as a base for invasion of Korea and China. To cope with this threat, Korea and China felt the need for a military alliance.

An even more important reason for friendly relations between Korea and China was cultural homogeneity. After the Iron Age, both China and Korea developed agricultural economies and thus shared many cultural similarities. China's culture was more advanced than Korea's, however, and Korea absorbed much from China—*e.g.,* written Chinese characters, laws and decrees, Confucianism, and painting. Nevertheless, Korea had its own language and invented its own alphabet; it developed its own way of life and adapted and improved upon all its borrowed culture to fit its own needs.

By the late 19th century the influence of the West began to be felt in East Asia. This was chiefly because Japan, the less advanced nation in the region, took the lead after opening its doors to the West. Stimulated by Japan's Westernization program, Korea tried to carry out reforms, doing away with traditional systems. But Korea fell prey to Japanese imperialism, which took advantage of reform movements in Korea for its own benefit. On liberation from Japanese rule in 1945, Korea was thrown into the vortex of world politics.

THE DAWN OF HISTORY

The Stone Age. Stone artifacts of the Paleolithic Period were unearthed at Kulp'o-ri in North Hamgyŏng Province (Hamgyŏng-pukto) and at Sŏkch'ang-ni in South Ch'ungch'ŏng Province (Ch'ungch'ŏng-namdo). Of 13 stratified Paleolithic sites, each cultural stratum produced chipped-stone tools of different shapes. Radioactive-carbon dating indicates the Paleolithic provenance of Sŏkch'ang-ni. Inhabited sites with round fireplaces were discovered there along with carved pebbles.

Neolithic life

The Neolithic Period was well established by 3000 BC, and a major characteristic was the use of combware pottery, chiefly found at seashore and river-basin sites, where inhabited places and shell mounds were also discovered. Stone spears and flint arrowheads have also been found, as well as bone hooks and stone weights, used for fishing. Remains of the Late Neolithic Period include stone plows and sickles, which indicate the beginning of farming. People lived in dugouts, mostly shallow round or rectangular hollows with fireplaces in the centre and possibly covered with thatched roofs. These shelters were huddled together in groups. The size of such villages is yet to be determined, but legends indicate the family members might have lived together, forming clan communities.

The use of metals and the emergence of tribal states. Bronzeware was probably first used about the 8th century BC, though some scholars think that it predates the 10th century. As the Bronze Age started, the design of pottery changed to undecorated earthenware. The uncovering of such pottery indicates that Bronze Age Korean people lived on hillsides, in dugouts raised slightly above ground. Half-moon-shaped stone knives and grooved stone axes show that rice farming was practiced, and bronze daggers and bronze arrowheads indicate wars of conquest. Dolmens, used as tombs, which were discovered in south Manchuria and the Korean peninsula, show the boundary of ancient Korean culture. Because only important persons were buried in dolmens, their number and location indicate that many small Bronze Age tribal states were probably established by powerful men.

The most advanced state was ancient Chosŏn, established in the Taedong-gang (Taedong River) Basin. According to myth, the son of Heaven, Hwanung, descended to Earth and married a bear-turned-girl, who bore a son, Tangun, the founder of Chosŏn. Perhaps Tangun, who called himself a grandson of Heaven, ruled a tribal state in which rituals and politics were not separated.

Chosŏn developed into a league of tribes in the area of the Taedong and Liao rivers (*c.* 4th century BC). Around this time ironware came to be used. Iron plows and sickles indicate the use of animals in farming and more efficient harvesting methods. Wooden houses were built on the ground, and *ondol,* a floor-heating system, was developed. Iron weapons were manufactured. The appearance of horse equipment and coaches indicates that horses and chariots were employed in wars.

Wiman (Wei Man in Chinese), said to have defected from China, became ruler of Chosŏn about 194 BC. More likely, he was indigenous to Chosŏn. Wiman's Chosŏn was overthrown by the Han Empire of China and replaced by four Chinese colonies in 108 BC.

DEVELOPMENT OF ANCIENT STATES

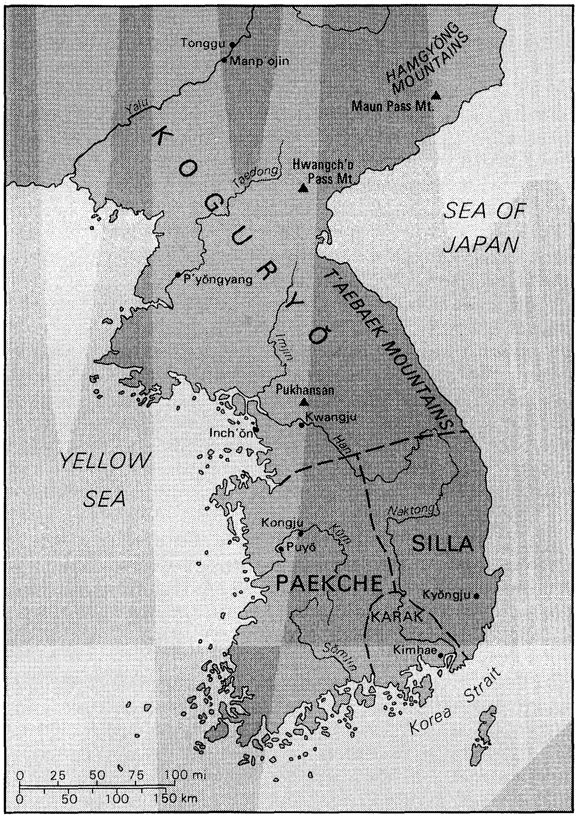
The Three Kingdoms. Apart from Chosŏn, the region of Korea developed into tribal states. To the north, Puyŏ developed in the Sungari River Basin in Manchuria. Chin, south of the Han-gang (Han River), was split into three—Mahan, Chinhan, and Pyŏnhan. These states were leagues, tribal federations centred on a leading state. The tribal league states stretched across a wide area from the Sungari Basin in Manchuria to the southern Korean peninsula. They evolved into three conflicting kingdoms—Koguryŏ, Paekche, and Silla. According to myths, Koguryŏ was founded by Chu-mong in 37 BC, Paekche by Onjo in 18 BC, and Silla by Pak Hyŏkkŏse in 57 BC. The actual development of a state, however, was begun for Koguryŏ by King T'aejo (reigned AD 53–146?), for Paekche by King Koi (reigned 234–286), and for Silla by King Naemul (reigned 356–402).

Common characteristics of the Three Kingdoms

The Three Kingdoms shared several common characteristics. They evolved into statehood through frequent wars of expansion. Centralized military systems were organized, and training institutions (*kyŏndang* in Koguryŏ, *hwarangdo* in Silla) were developed. The power of the king in each state was strengthened. Royal hereditary systems were established.

Another common trait was the appearance of central aristocrats, tribal chiefs who moved to the capital. These aristocrats were divided into several social classes and gained certain privileges as they advanced socially and politically. The typical class system was Silla's Kolp'um, or bone-rank system, in which the families of rulers customarily monopolized political power. Silla had a state conference, Hwabaek, composed of men of *chin'gol,* or "true bone" (royal or formerly royal origin), which made important state decisions.

From E. Reischauer and J. Fairbank, *East Asia, the Great Tradition,* Copyright © 1958 and 1960 by Edwin O. Reischauer and John K. Fairbank; Houghton Mifflin Company



Korea during the Three Kingdoms period (*c.* AD 400).

The states all experienced a centralization of power. The nations were divided into administrative units—the largest called *pu* in Koguryŏ, *pang* in Paekche, *chu* in Silla—that controlled many castles. The central government sent officials to these provincial units, where they saw to it that the people, as royal subjects, provided taxes and corvée labour.

The Three Kingdoms developed highly advanced cultures. Each compiled its history, apparently to consolidate the authority of the state. Also noteworthy was the introduction of Buddhism, characterized at the time as a nationalistic religion praying for the protection and welfare of the state.

Unified Silla. With the support of China, Silla conquered and subjugated Paekche in 660 and Koguryŏ in 668. Not until 676 did Silla drive out the Chinese and gain complete control of the Korean peninsula. But the surviving Koguryŏ people in northern Manchuria established Parhae, under the leadership of Dae Cho-yŏng, which soon came into direct confrontation with Silla. This period could well be called an age of opposing southern and northern states, but it is customary to place the primary focus on Silla because little is known about Parhae (though it built a highly civilized state that the Chinese referred to as Haedong-sŏngguk, the Prosperous Country of the East). After Parhae's collapse its territory was controlled by the northern nomadic peoples and was thus not part of Korean history.

Unified Silla saw the flowering of absolute monarchy, which reduced almost to nothing the influence of the Hwabaek. A central administrative body called Chipsabu was established to execute royal decrees. Aristocrats were now granted salaries and given land, but the latter was to revert to the state after tenure, thus reducing direct control of land and of the governed by the aristocracy. Monarchs built extravagant palaces and annexes and royal tombs. The nation was divided into administrative units: state, county, and ward. Five provincial capitals prospered as cultural centres.

Among the aristocracy, Avataṃsaka Buddhism provided the ideological backing for autocratic monarchy. The underprivileged general public was attracted most to the *Sukhāvatī-vyūha-sūtra,* promising bliss in the next world. The flowering of Buddhism produced many beautiful temples and great works of art, the most remarkable of which were Pulguk-sa, Sŏkkuram (a grotto shrine), and the bell at Pongdŏk-sa.

Confucianism prospered among low-echelon aristocrats, who used it as a foothold for political advancement. The National School, Kukhak, was established, and a sort of civil service examination system, called *toksŏ samp'um kwa,* was installed.

Disputes over succession in the 8th century

Emergence of provincial magnates. Frequent conflicts and rebellions over succession took place among the Silla aristocracy in the late 8th century. Aristocrats eventually restored the authority of the Hwabaek, overthrowing royal despotism. Even lower-ranking aristocrats demanded the abolition of restrictions imposed by the strict status system. New powerful families appeared in many provinces and grew because of the weakening of central control. Provincial military fortresses were established to repel Chinese pirates. The most active was the Ch'ŏnghae fortress led by Chang Po-go, who almost monopolized trade with China and Japan and had a private army of 10,000. Silla settlements in Chinese coastal cities in the Shantung Peninsula were also engaged in trade. Also powerful were the village rulers, who became castle lords by reinforcing control over military, administrative, and economic affairs. Many farmers, taxed by both the central government and castle lords, chose to become drifters or robbers, often staging rebellions.

Largely as a result of these trends, two provincial leaders, Kyŏnhwŏn and Kungye, established the Later Paekche (892) and Later Koguryŏ, also called Majin or T'aebong (901), kingdoms, which, along with Silla, are commonly referred to as the Later Three Kingdoms. In this period Sŏn (Zen) Buddhism was most popular, emphasizing the importance of realizing, through contemplation, the inborn Buddha nature of the individual.

KORYŎ

Social structure and culture. Koryŏ was founded in 918 at Songak (Kaesŏng) by Wang Kŏn, who in 935 established a unified kingdom in the Korean peninsula. Wang Kŏn went to great lengths to absorb the people of the overthrown states, even accepting the survivors of Parhae, which had been destroyed by the Khitan (Liao). Proudly declaring itself the successor of Koguryŏ, Koryŏ launched active campaigns to recover lost territory, clashing frequently with the Khitan in the north. Koryŏ eventually expanded its territory to the Yalu River (Korean Amnok-kang).

The Koryŏ ruling class consisted largely of provincial castle lords and former Silla aristocrats. The rulers regarded their family lineage highly. Marriage into a powerful family, especially a family of royal blood, was an important means for maintaining and elevating social and political status. Sons of a family above the fifth of nine official grades received official posts without going through civil service examinations.

The central government was run by the two supreme organs—Samsŏng, the highest administrative body, and Chungch'uwŏn, the secretariat to the king. These two formed the Supreme Council of State. Koryŏ thus practiced aristocratic council-centred politics. Central aristocrats were granted land that was later to be returned to the state. Officials above the fifth grade were given land for permanent possession. Even the land supposed to be returned was actually handed down for generations because the grantees' sons usually became officials. Aristocrats expanded their land by reclaiming, purchasing, or seizing by force, and land became the primary source of wealth.

Aristocrats believed in Buddhism as a religion, for spiritual fulfillment and personal happiness; and in Confucianism, for its political precepts and ethical principles. The same was true of the government, which built grand Buddhist temples, such as Hŭngwang-sa, to observe rituals and pray for the prosperity of the nation, but also set up a school named Kukchagam to teach Confucianism.

Military rule. Civilian officials constituted the core of the ruling class; the military was generally discriminated against. Indeed, the supreme commander for military affairs was a civilian. Military officials were not eligible for the second grade of the official hierarchy and were excluded from the Supreme Council. Even in the same official grade, military men received less land than did their civilian counterparts. This discrimination eventually led to a military coup d'etat in 1170. The revolutionaries massacred a large number of civilian officials and gained complete control of government. Struggles for hegemony soon occurred among the leaders and were ended by Gen. Ch'oe Ch'ung-hŏn, who established a military regime of his own that lasted about 60 years.

The military coup of 1170

The monarch remained as a figurehead, deprived of political power, which was in the hands of the Ch'oe families. The Ch'oes set up a private army for personal security and a new public military organization for national security, but the latter also served, in effect, as their private army. They also established a body of civilian officials to take care of the state personnel administration, thus controlling both military and civilian leaders.

Buddhism was suppressed and retreated to remote mountain areas, where it formed a new Sŏn sect called Chogye-jong, which became the main current of Korean Buddhism. The underprivileged farmers, stimulated by a general political atmosphere in which subordinates rose against superiors, staged rebellions across the country over a period of 30 years. The upheavals were at first natural and spontaneous protests against unfair oppression, but they developed into organized campaigns for liberation and for the seizure of power. The rebellions, eventually brought under control through appeasement and the use of naked force, were nevertheless instrumental in improving the lot of the underprivileged.

In 1231 the Mongols invaded Koryŏ, and the Ch'oe regime resisted for about 30 years. Even farmers and servants stood up bravely, and the Mongols, who had conquered most of Eurasia, could not take Koryŏ by force. As the exploitation of farmers by the Ch'oe grew more severe, however, the people became estranged, and the regime was finally overthrown by civilian leaders, who concluded a peace treaty with the invaders.

Social change in later Koryŏ. After the peace treaty Koryŏ was subject to some political interference from the Mongols but retained its political and cultural independence. Koryŏ went to some lengths to show its national and cultural superiority over the invaders by turning out highly advanced poems and national history books.

This period saw the remarkable development of huge farms, run by powerful aristocrats, which were scattered across the country. The landowners lived in the capital and sent private vassals or servants to collect taxes from the tenants, the common people who farmed the land. These tenants were often required to pay taxes to more than one owner because it became popular for landholders to share ownership. Tenants were also subject to forced labour and military duty for the state. Many farmers chose to become servants, thus getting protection from the aristocrats and avoiding the state duties. Some aristocrats also caught drifters and illegally made them servants. Thus, the number of servants increased, but these servants were actually on a level with tenants. They were not slaves in the Western sense but were more like serfs. The increase in the number of landholders and servants reduced the source of state tax revenue and the number of people to be mobilized in war.

Through civil-service examinations, the central government recruited a new bureaucratic force consisting of *sadaebu* (scholar officials), who generally had small farms in their native towns under their own management. These men were usually not satisfied with Buddhism and mere interpretations of the classics, and they adopted Neo-Confucianism, which brought a metaphysical approach to the understanding of the universe. Because of the financial straits of the government, the new officials could not receive land commensurate to their rank; thus, their demand for land reform was strong. Eventually, with the support of Gen. Yi Sŏng-gye, they seized power and established a new land-distribution system, under which land was granted according to the rank of office. These reforms ended the Koryŏ dynasty in 1392 and established the Yi dynasty. (K.-b.L.)

End of the Koryŏ dynasty

THE YI DYNASTY

The establishment of a Confucian state. The Yi dynasty was established and the region under its control named Chosŏn with permission of the emperor of China. The Yi dynasty, with 26 monarchs, ruled until the Japanese annexation of Korea in 1910. Hanyang (now Seoul) was made the capital. The Confucian ethical system was adopted officially and replaced Buddhism, which had become corrupt. Many Confucian institutions of learning were set up, and Neo-Confucian scholars gained government posts through civil-service examinations.

The early Yi dynasty flourished intellectually and culturally, especially in the reign of Sejong the Great, the fourth monarch. With the technique of movable-type printing, developed in Korea in 1234, many publications in such fields as medicine, astronomy, geography, history, and agriculture were produced. In 1420 a royal academy called Chiphyŏnjŏn was established, and many promising young scholars engaged in study and research there. In 1443 the Korean phonetic alphabet, Han'gŭl, was completed under Sejong's direction.

In the reign of Sejo, the seventh monarch, a powerful centralized and civilian-oriented government structure emerged. Laws were codified. The highest administrative body was the Supreme State Council. The country was divided into eight administrative provinces, and all officials were appointed by the central government.

Late in the 15th century, Korean scholars made original contributions to the theoretical refinement of Confucianism. In the mid-16th century many of these scholars were recruited to government positions. Idealistic in orientation, they criticized the bureaucratic establishment and recommended drastic measures toward the realization of Confucian ideals. But relentless counterattacks and pressures forced most of the scholars to quit their posts. They set up private academies of Confucian learning, called *sŏwŏn.* These academies produced many eminent scholars, including Yi Hwang (T'oegye) and Yi Yi (Yulgok), whose distinct theories of the universe evolved into mutually antagonistic schools.

Foreign invasions. In 1592 Toyotomi Hideyoshi, the military ruler of Japan, sent a large expeditionary force to Korea in an alleged attempt to invade China. Korean land forces suffered a series of defeats, but Korean naval forces, led by Adm. Yi Sun-shin, secured full control of the sea. Yi won the greatest naval victory in Korean history over the Japanese troops off the southern coast. People of almost all ranks, even Buddhist priests, volunteered to fight the Japanese. Contingents came from Ming China to help Korea. After about a year the Japanese were forced to withdraw. In 1597 Toyotomi launched another invasion, but, after his sudden death in 1598, the Japanese again withdrew. The war left most of Korea in waste. Palaces, government buildings, and private houses were burned; cultural treasures were lost or destroyed. Some scholars and artisans were taken to Japan, where they were required to teach the methods of Korea's more advanced technology.

Korea's greatest naval victory

In the early 16th century, nomadic Manchu violated the borders of both Ming China and Korea. Ming and Korean punitive attacks on Manchu strongholds in 1619 were beaten back, and in 1627 Manchu nomads overran the northern regions of Korea. Only after Korea had agreed to recognize their claim to "brotherhood" did the Manchu pull out of the occupied territory. In 1637 the Manchu captured Seoul and wrested an unconditional surrender from the king. The Manchu then overthrew the Ming and established the Ch'ing dynasty. Thus, the tribute that Korea had paid to the Ming now had to be switched to the Ch'ing.

Silhak and popular culture. A series of significant changes in Korea began in the mid-17th century and made a great impact on virtually every sector of Korean society in the 18th century. In agriculture, rice transplantation became popular. Irrigation systems were improved. Advances in farming brought dramatic boosts in agricultural produce and raised the standard of living for farmers. With the cultivation of such special crops as tobacco and ginseng, commerce and trade developed apace. The government started minting coins and collecting farm rents in cash. Markets were held in many places across the country. Particularly active were merchants from Kaesŏng, who had a nationwide network that put every fair in the country within their sphere of influence.

In the realm of scholarship, attention shifted from traditional theorizing to matters of relevance—to the needs of the society and nation. Such scholars are often referred to as the Silhak, or Practical Learning school. Silhak scholars fell into four major groups. One group advocated comprehensive administrative reform, calling upon the government to rationalize the systems of civil-service examination, education, taxation, and land administration. Another group stressed the need to foster commerce, industry, and technology. A third conducted critical examinations of the Confucian Classics. The fourth group focussed on the study of the history, geography, and language of Korea. Their research and publications provided a basis for study for the generations that followed.

Comparable new trends appeared in arts and letters. Mass-oriented literary and artistic works came into fashion, a great change from the tradition of catering exclusively to the upper classes. The new works were not only written in the easy-to-read Han'gŭl but also gave frank expression to popular discontent. Singing dramas or traditional Korean operas, most of them adapted from popular novels, were also popular with the masses. Many artists specialized in pictures of blacksmiths at work, farmers in the field, traditional wrestling matches, and in landscapes of the countryside. Pottery with a simple blue and white glaze was produced in large quantities for mass consumption.

Arrival of Europeans

The introduction of Catholicism. Europeans began to arrive in East Asia in the mid-17th century. In 1656 a Dutch merchant ship went aground off the southern shore of Cheju-do (Cheju Island). The 36 survivors were taken to Seoul for detention. About 13 years later Hendrik Hamel and seven others escaped and returned home. Hamel wrote an account of his experiences—the first book on Korea published in the West.

Along with the European merchants in East Asia came Catholic priests. Korea's first contact with Christianity was through missionaries in China. Korean envoys to China in the 16th century brought back with them a world atlas and scientific instruments made by the priests, as well as literature on science and Christianity. Some Silhak scholars had converted to Catholicism by the late 18th century, even before missionaries reached Korea. Most of the early converts were aristocratic scholars. A number of commoners were later attracted to the Catholic Church, finding a hope of salvation in the Christian doctrine of equality of all men before God and a new source of joy in the Christian belief in life after death. Catholicism spread from Seoul to the provinces slowly but steadily.

The incompatibility of Catholicism with Confucianism posed a serious problem. The two sides could not compromise on the great importance Confucianism attached to ancestor worship, which Catholicism rejected as pure idolatry. The government began to suppress the Catholic Church in the belief that it defied the existing sacrosanct mores of Confucianism. During persecutions in 1801, 1839, and 1866, scholar converts were either put to death or forced to turn renegade; foreign missionaries were ferreted out and beheaded. But rank-and-file Catholics rallied around the church and kept it alive. In 1831 the Holy See set up a Korean parish. French priests smuggled themselves into the country and engaged in clandestine activities.

The advent of Silhak, mass-oriented arts, and Catholicism in the 17th and 18th centuries is an indication of a modern Korea in the making. But in the 19th century young princes came to the throne, and power-hungry maternal relatives seized power and plunged the government into a state of collapse. One peasant uprising followed another in the provinces, and the whole nation was seething with popular discontent and resentment.

Many farmers sought refuge in religion. A new religion founded (*c.* 1860) by Ch'oe Che-u, a fallen aristocrat scholar, advocated sweeping social reform; it had much in common with traditional animism, and it appealed to the farmers. This religion was called Tonghak, or Eastern Learning, as a counterpoise to Sŏhak, or Western Learning—*i.e.,* Catholicism.

CONTACT WITH WORLD POWERS

The opening of the door. King Kojong was too young to rule when he ascended the throne in 1864, and his father, Taewŏn-gun, was appointed regent. Taewŏn-gun set out to restore all the powers of the monarchy and pursued a policy of national isolation. He put into force bold political reform measures, such as liberal civil service recruitments and the abolition of a great number of private Confucian academies.

During his regency Western men-of-war and merchant vessels came in search of trade and friendship, but the regent refused them. Korean soldiers and civilians burned and sank the United States merchant ship "General Sherman" at P'yŏngyang in revenge for acts of plunder committed by the crew. Korean forces repulsed two attacks by French warships in 1866. In 1871 a United States flotilla came to retaliate for the "General Sherman" but was beaten back. Such incidents strengthened Taewŏn-gun's resolve to keep the country's doors closed.

Japan repeatedly made futile attempts to establish diplomatic relations with Korea. The Japanese military thereupon raised an outcry for a war of conquest on Korea. Meanwhile, Taewŏn-gun came under widespread criticism for the enormous financial burden he had imposed on the people. Popular resentment forced him to step down as regent in 1873. Relatives of Queen Min took over the helm of state and initiated policies opposed to Taewŏn-gun's. Japan, which had been watching every development in Korea, now dispatched a fleet and pressured Korea to sign a treaty of trade and friendship. The ports of Pusan, Wŏnsan, and Inch'ŏn were opened to the Japanese in 1876.

Treaty with Japan

The growing Japanese presence in Korea was disturbing to the Ch'ing rulers of China. When conservative soldiers tried to restore Taewŏn-gun, the Ch'ing used it as an excuse for stationing forces in Korea. Thus began a period of open Ch'ing interference in Korean affairs. The Ch'ing forced Korea to sign a treaty of trade that heavily favoured Chinese merchants. Korea signed a treaty of trade and friendship with the United States through the offices of China. Similar treaties with Great Britain, Germany, Russia, and France followed, and resident foreign missions were established in Seoul.

Once the doors were opened, a modernization movement was begun. Students and officials were sent to Japan and China. Western-style schools and newspapers were founded. The government, however, could not push ahead with a consistent policy of modernization, for the King was feebleminded and the ruling classes were hopelessly divided into radicals and moderates.

In a coup of 1884 the Radicals seized power and drew up a bold blueprint for reform. But a Ch'ing contingent moved in and overthrew their three-day-old regime. This led to the signing of the Li-Itō Convention, designed to guarantee a Sino-Japanese balance of power in the Korean peninsula.

The Tonghak Revolt and government reform. Government expenditures greatly increased, largely because of appropriations for machinery imports and government reorganization, and the financial picture was aggravated by obligations to pay reparations to foreign governments. Heavier tax levies were imposed on farmers, who provided the bulk of government revenue. The import of such necessities as cotton textiles upset the traditional self-sufficiency of the farming community. Usurious loans by Japanese rice dealers helped reduce the peasantry to abject poverty. Angry farmers turned increasingly to Tonghak.

Despite ruthless government persecution, Tonghak took deep root in the peasantry. Its followers staged large-scale demonstrations calling for an end to injustice. Negative official response precipitated the Tonghak Revolt (1894), in which the Tonghak followers and the peasantry put up a united front for popular liberation. Government troops armed with Western weapons suffered ignominious defeats in southern Korea, weakening the government's military grip on the country. Foreign intervention seemed the last resort open to the rulers. Ch'ing troops soon moved in at the request of the government. Then Japan dispatched its large units uninvited, and the two alien powers were in sharp and sudden confrontation.

Japanese hegemony

The rebels laid down their arms voluntarily to defuse the threat. But a war broke out in 1894; Japan emerged victorious, and the two powers signed the Treaty of Shimonoseki, recognizing Japanese hegemony in Korea.

The Japanese dictated to the Korean government a wide range of reforms. Korea set up a council to plan and initiate reforms and issued pertinent decrees. Western-style institutions and a cabinet were formed. Civil-service examinations were discontinued. Such social practices as class discrimination were abolished. But public reaction to the reform policy was unfavourable. The government realized that old practices and institutions die hard and that reform takes more than mere decrees and imitation of things Western.

International power struggle and Korea's resistance. Japan's supremacy in Korea and its subsequent seizure of the Liaotung Peninsula in Manchuria was more than Russia, with its long-cherished dream of southward expansion, could tolerate. With German and French support, Russia pressured Japan to return the peninsula to China. At the same time, encouraged by Russia, the Korean government showed signs of drifting toward an anti-Japanese course. The Japanese government, however, promptly engineered the assassination of Queen Min (October 1895), the suspected mastermind behind the anti-Japanese attitude. Fearing for his own life, the King took refuge in the Russian Legation, granting such concessions as mining and lumbering rights to Russia and other powers.

Popular movements for the restoration of Korean sovereignty arose under the leadership of such figures as Sŏ Chae-p'il (Philip Jaisohn). After many years of exile Sŏ organized, in 1896, a political group called the Tongnip Hyŏphoe (Independence Association); he also published a daily newspaper named *Tongnip Sinmun* (The Independent) as a medium for awakening the populace to the importance of sovereignty and civil rights. On the urging of the Tongnip Hyŏphoe, the King returned to his palace and declared himself emperor and the Korean empire equal to other nations.

The Boxer Rebellion in China led to a Russian invasion of Manchuria and to the Russo-Japanese War (1904–05). The Korean government at first declared neutrality. But under Japanese pressure, it signed an agreement allowing Japan to use much of the country for operations against the Russians.

Japan won the war, and the resulting Treaty of Portsmouth (September 1905), signed through the mediation of the United States, recognized Japan's undisputed supremacy in Korea. Its hand thus strengthened, Japan forced the Korean emperor into signing a treaty that made Korea a Japanese protectorate (December 1905).

Although the Korean emperor sent a secret emissary to the international peace conference held at The Hague in 1906 to urge the big powers to intercede with Japan on behalf of Korea, the mission failed, serving only to infuriate Japan. Under Japanese coercion, the Emperor then abdicated in favour of his son, Emperor Sunjong. The Korean Imperial Army was disbanded, and in 1910 Japan annexed Korea.

Japanese annexation of Korea

A Korean Army, led by deposed officials and Confucian scholars, had arisen against the Japanese in the southern provinces following the 1905 treaty. For five years the militiamen effectively harassed the Japanese occupation forces, especially in 1908 and 1909. With the annexation, however, they were driven into Manchuria. Large numbers of Koreans emigrated to Manchuria, Shanghai, and Hawaii around this time.

KOREA UNDER JAPANESE RULE

Military control. Japan set up a government in Seoul, with the governor generalship filled by generals or admirals appointed by the Japanese emperor. Koreans were deprived of freedom of assembly, association, the press, and speech. Many private schools were closed because they did not meet certain arbitrary standards. The colonial authorities used their own school system as a tool for assimilating Korea to Japan, placing heavy stress on the Japanese language and excluding such subjects as the language and history of Korea. The Japanese built a nationwide transportation and communications network and established a new monetary and financial system. They also encouraged Japanese commerce in Korea while barring Koreans from similar activities.

The government promulgated a land-survey ordinance that forced landowners to report the size and area of their land. By failing to report, many farmers were deprived of their land. Land and forestry owned jointly by a village or a clan were also appropriated by the Japanese because no single individual could lay claim to them. Much of the land acquired by the government was sold cheaply to Japanese. Many of the dispossessed took to the woods and became brand tillers; others emigrated to Manchuria and Japan in search of jobs (the majority of Korean residents in Japan today are their descendants).

The March 1st Movement. A turning point in Korea's resistance movement came on March 1, 1919, when nationwide anti-colonial rallies were staged. The former emperor Kojong, the supreme symbol of independence, had died a few days before, bringing mourners from all parts of the country into Seoul. A declaration of independence was read at a rally in Seoul on March 1. Waves of students and citizens took to the streets, calling for independence. This movement took the form of peaceful demonstrations appealing to the conscience of the Japanese. An estimated 2,000,000 persons took part. The authorities responded with brutal repression, unleashing their gendarmerie and army and navy units and thus ending the demonstrations. They arrested some 47,000 Koreans, of whom about 10,500 were indicted, and killed and wounded nearly 23,000.

In April, independence leaders, including Syngman Rhee, An Ch'ang-ho, and Kim Ku, formed a Korean provisional government in Shanghai. It brought together all Korean exiles and established an efficient liaison with leaders inside Korea. Japan realized that its iron rule had to be replaced with more sophisticated methods. The gendarmerie gave way to an ordinary constabulary force, and some freedom of the press was granted. But the oppressive and exploitative Japanese colonial policy remained as ruthless as ever though with less conspicuous methods.

Korean provisional government in exile

Taking advantage of a wartime business boom, Japan took leaps forward as a capitalist country. Korea became not only a market for Japanese goods but also a fertile and untapped market for capital investment. Meanwhile, industrial development in Japan was achieved at the sacrifice of agricultural production, creating a chronic shortage of rice. The government enforced projects for the improvement of rice production throughout Korea. Many farmers were ordered to turn their dry fields into paddies. The program was temporarily discontinued during the worldwide depression in the early 1930s. But it was soon resumed to meet the increased needs of the Japanese military in its war against China, which began in 1931. Most Koreans were forced to subsist on low-quality cereals imported from Manchuria instead of their own rice.

The end of Japanese rule. Of the several dailies and magazines founded shortly after the March 1st Movement, the newspapers *Dong-a Ilbo* and *Chosun Ilbo* spoke the loudest for the Korean people and inspired them with the ideals of patriotism and democracy. In the academic community, scholars conducted studies of Korean culture and tradition. Many novels and poems were written in colloquial Korean.

A major anti-Japanese mass rally was held in Seoul in 1926, occasioned by the funeral of Sunjong. A nationwide student uprising originated in Kwangju on November 11, 1929, calling an end to Japanese discrimination. These and other resistance movements were led by the New Cadre, composed of a whole spectrum of Korean intellectuals.

In 1931 the Japanese imposed military rule once again. After the outbreak of the Sino-Japanese War (1937) and of World War II (1941), Japan tried to obliterate Korea as a nation, forcing Koreans to worship at Shintō temples and even to adopt Japanese names and banning academic societies devoted to Korean studies as well as newspapers and magazines published in Korean. The Japanese needed manpower to replenish the dwindling ranks of their military and labour forces. They drafted hundreds of thousands of able-bodied Koreans to fight for Japan and to work in mines, factories, and military bases.

When Shanghai fell to the Japanese, the Korean provisional government moved to Chungking. It soon formed a liberation army of Korean independence fighters scattered all over China and declared war on Japan in 1942. The small liberation army fought with the Allied forces in China until the Japanese surrender in 1945, which ended 36 years of Japanese rule over Korea. (K.-r.L.)

DIVISION OF KOREA

The Cairo Declaration, issued on December 1, 1943, by the United States, Great Britain, and China, pledged independence for Korea "in due course." The vague phrase aroused the Korean provisional government in Chungking (southwest China) to request interpretation from the United States. Their request, however, received no answer.

At the Yalta Conference held in February 1945, Pres. Franklin D. Roosevelt proposed to Joseph Stalin of the U.S.S.R. a four-power trusteeship for Korea between the U.S., Great Britain, the U.S.S.R., and the Republic of China. Stalin generally agreed to Roosevelt's offer, but they did not reach any formal agreement on the future state of Korea, and after the Yalta meeting there was a growing uneasiness between the Anglo-American allies and the U.S.S.R.

Initial concord on Korea

Throughout the Potsdam Conference in July 1945, U.S. military leaders insisted on encouraging Soviet entry into the war against Japan. The Soviet leaders asked the U.S. about invading Korea, and the U.S. declined to do so on the grounds that such an expedition would not be practicable until after a successful landing had taken place on the Japanese mainland. The ensuing Potsdam Declaration included the statement that "the terms of the Cairo Declaration," which promised Korea its independence, "shall be carried out. . . ." In the terms of its entry into the war against Japan on August 8, the U.S.S.R. subscribed to support the independence of Korea. On the following day Soviet troops went into action in Manchuria and landed on the northern tip of Korea.

The General Order No. 1, drafted on August 11 by the United States for Japanese surrender terms in Korea, provided that Japanese forces north of the 38th parallel of latitude were to surrender to the Soviet commanders while those south of that line were to surrender to the U.S. commander. Stalin did not raise objections to the contents of the order, and on September 8 American troops arrived in southern Korea almost a month after the first Soviet entry. On the following day the United States received the Japanese surrender in Seoul. There were now two zones, for the Soviets had already begun to seal off the 38th parallel.

The historic decision to divide the peninsula has aroused speculation on several counts. Some believe that the division of Korea was made simply for the sake of military expedience in receiving the Japanese surrender. Others tend to think that the division was politically motivated to prevent the Soviet forces from occupying the whole of Korea. Considering the fact that American policy toward Korea during World War II had aimed to prevent any single power's domination of Korea, it may be reasonably concluded that the reason for the division was to stop the Soviet advance south of the 38th parallel.

After the simple discussion on trusteeship between Stalin and Roosevelt at the Yalta meeting, it was not until May 1945, at a meeting with the Chinese foreign minister in Moscow, that Stalin confirmed his agreement to set up a four-power trusteeship. The ensuing Moscow Conference, held in late December, created a four-power trusteeship and established a Joint U.S.–U.S.S.R. Commission of the rival U.S. and Soviet military commands in Korea for the settlement of the question of a unified Korea. When the Joint Commission convened in Seoul from March to May 1946, the Soviet delegates demanded that those Korean political groups that had opposed trusteeship be excluded from consultation. The United States refused, and on this rock foundered all attempts by the commission to prepare for the unification of Korea. The commission met again from May to August 1947, but it achieved nothing for the creation of a unified Korea.

The United States presented the entire matter of Korean unification to the United Nations in September 1946. The United Nations General Assembly adopted a resolution, proposed by the United States, rejecting the Soviet position. The resolution called for general elections in Korea under the observation of a UN Temporary Commission on Korea, those elected to make up a National Assembly, establish a government, and arrange with the occupying powers for the withdrawal of their troops from Korea. But the U.S.S.R. barred the Temporary Commission from entering North Korea. The South, however, held elections under the supervision of the Temporary Commission on May 10, 1948. The National Assembly convened on May 31 and elected Rhee as its speaker. Shortly afterward a constitution was adopted, and Rhee was elected president on July 20. Finally, on August 15, the Republic of Korea was inaugurated, and the military government came to an end. On December 12 the UN General Assembly declared that the republic was the only lawful government in Korea.

Construction of the People's Republic

Meanwhile, on November 18, 1947, the Supreme People's Assembly of North Korea set up a committee to draft a North Korean constitution. The committee adopted a new constitution in April 1948, and on August 25 elections for members of the Supreme People's Assembly were held with a single list of candidates. On September 3 the constitution of the Democratic People's Republic of Korea was ratified by the Supreme People's Assembly, which was holding its first meeting in P'yŏngyang. Kim Il-sung was appointed premier of the People's Republic, and on September 9 the Democratic People's Republic of Korea was proclaimed. The U.S.S.R. recognized the People's Republic as the only lawful government in Korea on October 12. (B.-h.H.)

South Korea

In 1948, by a United Nations resolution, the Republic of Korea (South Korea) was formed in the south, and in the same year the Democratic People's Republic of Korea (North Korea) was formed in the north. As a result of the Korean War, which began in June 1950, following an invasion of South Korea by troops from North Korea, a demilitarized zone was established in 1953. It runs roughly from the mouth of the Han-gang (Han River) on the west to a little south of the town of Kosŏng on the east coast and is about 150 miles (240 kilometres) in length. The present effective administrative area of South Korea is about 45 percent of undivided Korea: 38,221 square miles (98,993 square kilometres). South Korea is divided administratively into nine provinces (*do* or *to*) and two special cities (*t'ŭkpyŏlsi*), Seoul (Sŏul) and Pusan. It maintains diplomatic and trade relations with more than 100 countries. South Korea has not been admitted to the United Nations but has joined almost all of the United Nations agencies. The country has made considerable economic progress since the 1950s, with the government pursuing ambitious industrialization programs with special emphasis on mobilization of domestic capital and the promotion of export industries.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Korea is largely mountainous, with small valleys and narrow coastal plains. The Taebaek-sanmaek (Taebaek Mountains), forming the backbone of the peninsula, run in a north–south direction along the eastern coastline. From them extend several mountain ranges oriented in a northeast–southwest direction. Principal rivers (*gang*), such as the Han, Kum, and Naktong, all have their sources in the Taebaek-sanmaek, and they flow between the ranges. These mountains are not very high, none exceeding 5,700 feet (1,737 metres) above sea level. The highest peak in South Korea, Halla-san on Cheju-do (Cheju Island) is 6,398 feet (1,950 metres) above sea level. There are comparatively extensive lowlands along the lower parts of the Han, Kum, and Naktong rivers. The eastern coastline is relatively straight, whereas the west has an extremely complicated ria (*i.e.,* creek-indented) coastline with many islands. The Yellow Sea and the complex coastline causes one of the highest tidal ranges in the world—about 30 feet (nine metres) maximum at Inch'ŏn, the entry port for Seoul.

Geological foundation

Geologically, the Korean peninsula consists in large part of Precambrian rocks (more than 570,000,000 years old), such as granite and gneiss. There are two volcanic islands, Cheju and Ullŭng, and a small-scale lava plateau in Kangwŏn Province (Kangwŏn-do). Most soils derive from granite and gneiss. Sandy and brown-coloured soils are common, and they are generally well leached and have little humus content. Podzolic soils are found in the highlands resulting from the cold of the long winter season.

Climate. Because of continental influences the climate of Korea is characterized by a cold winter and a hot summer. The annual range of temperature is greater in the north and in interior regions than in the south and along the coast. The average monthly temperature in January drops below freezing except along the southern coast, and the July average monthly temperature rises to about 78° F (25° C). The average monthly temperature in January at Seoul is about 23° F (−5° C), and about 78° F (25° C) in August; at Pusan, on the southeast coast, the average January temperature is 35° F (2° C), and the August average is about 78° F (25° C), as in Seoul. The annual rainfall varies from about 40 to 55 inches (1,016 to 1,397 millimetres). Taegu, the driest area, has 38 inches, and Pusan, one of the wettest areas, receives 55 inches of annual rainfall. About 70 percent of the annual rainfall is received during the summer monsoons and shifting polar fronts. Occasionally, late summer typhoons cause heavy showers and storms along the southern coast. The frost-free season varies from 170 to 226 days.

Plant and animal life. The long, hot, humid summer is favourable for the development of extensive forests, which cover about two-thirds of the total land area. Because of fuel needs in the long cold winter, however, and the high population pressure, the original forest has almost disappeared. Except for subtropical broadleaf forests in a narrow belt along the southern coast, most areas contain broadleaf and coniferous trees. Wild-animal life is similar to that of northern China and the Manchurian region. Tigers, leopards, lynxes, and bears, formerly abundant, have almost disappeared, even in remote areas.

Settlement patterns. The pace of urbanization since 1960 has caused decreases in rural population, not only proportionally but also absolutely. Agriculture is the most important occupation in rural areas, and rural settlements are thus found close to arable lands—mainly in river valleys and coastal lowlands. Agglomerated villages are common, ranging from a few houses to several hundred. Villages are frequently located along the foothills facing toward the south, backed by hills that give protection from the severe northwestern winter monsoon winds. Fields are divided into tiny plots and are cultivated by manual labour and animal power. Two types are found: rice fields and upland fields. Rice fields are usually irrigated, while upland fields are unirrigated dry fields, in which barley, wheat, soybeans, and millet are grown. Along the coastline small clustered fishing villages are found. Although the fishing population is not a large portion of the rural population, fishing is an important industry for export and for obtaining protein foods. Logging, mainly of coniferous trees, is limited to the mountain areas of Kangwŏn and Kyŏngsang provinces. Settlements in mountain areas are usually scattered, in contrast to the lowlands. Logging is usually practiced during the off-season for agriculture.

Urban expansion

Rapid expansion of urban areas in the past decade, especially the expansion of Seoul and Pusan, has resulted in considerable changes in urban landscapes. Before 1960 there were very few high-rise buildings; even in Seoul most structures were lower than 10 stories. By the early 1970s, however, buildings of more than 20 stories had become common in the city. Because of this rapid growth, city services, such as water, transportation, and sewage systems, have not met the increasing needs. The difficulties in improving these services is exacerbated by the fact that old and new buildings are still located side by side; modern high-rise buildings stand next to traditional one-story Korean houses, even in the business district.

The people. *Ethnic distribution.* The Korean people may originally have had links with the people of Central Asia, the Baikal region, Mongolia, and the coastal areas of the Yellow Sea. Tools of Paleolithic type and other artifacts found in Sokch'ang-ni, near Kongju, are quite similar to those of the Baikal and Mongolian areas. The physical characteristics of Koreans show Mongolian racial traits, such as dark straight hair, straight noses, high cheek bones, and the Mongolian eyelid fold. The population is quite homogeneous, with only a small percentage of foreigners, most of them urban Chinese, in South Korea.

Since 1960 birth and mortality rates have decreased rapidly. The decrease in the birth rate has been caused chiefly by a national campaign for family planning conducted since 1965. It also reflects the increasing number of educated people in South Korea. The country, however, continues to struggle to reduce its population growth.

Migration patterns

Before World War II, Koreans migrated to two major regions: Manchuria and Japan. People migrating into Manchuria were mainly from northern Korea, while those who went to Japan were mostly from southern Korea. It is estimated that in 1945 about 2,000,000 Koreans lived in Manchuria and Siberia and about the same number in Japan. About one-half of the Koreans in Japan returned to South Korea just after 1945. The most important migration, however, was the north-to-south movement of people after World War II, especially the movement that occurred during and after the Korean War. About 2,000,000 people migrated to South Korea from the North during that period, settling largely in the major cities.

More than one-half of the South Korean population lives in 30 cities and the country's two metropolises, Seoul and Pusan. Seoul, the capital, contains about one-fifth of the total population.

Religion. There is little uniformity of religious belief in Korea, which is confusing to outsiders. In a typical Korean family, the women may adhere to the Buddhist religion, while the men may be followers of the Confucian ethical system. The teachings of Confucius, including ancestor worship, are probably the most important basic belief and are especially strong in the rural areas. Deep-rooted Buddhism is more popular with women, and old Buddhist temples are part of the typical rural and urban scene. Christianity is relatively new in Korea, although it claims a large number of devoted adherents and has had a profound effect on the modernization of Korean society. An eclectic religion, Ch'ŏndogyo—a combination of Buddhism, Confucianism, Christianity, and even Taoism—spread widely in the latter part of the 19th century. Shamanism still remains strong in the minds of rural people, especially among women. Ch'angga Hakhoe (Value Creation Learning Society), introduced from Japan in 1963, has become popular in the low-income areas of major cities; it is a militant society of followers of the 14th-century Japanese Buddhist Nichiren, organized in Tokyo as the Sōka-gakkai in the early 1930s.

MAP INDEX

Political subdivisions

Cheju-do	33·20n	126·30e
Chŏlla-namdo	34·45n	127·00e
Chŏlla-pukto	35·45n	127·15e
Ch'ungch'ŏng-namdo	36·30n	127·00e
Ch'ungch'ŏng-pukto	36·45n	128·00e
Kangwŏn-do	37·45n	128·15e
Kyŏnggi-do	37·30n	127·15e
Kyŏngsang-namdo	35·15n	128·30e
Kyŏngsang-pukto	36·15n	128·45e
Pusan (Pusan-si)	35·10n	129·05e
Seoul (Sŏul-t'ŭkpyolsi)	37·34n	127·00e

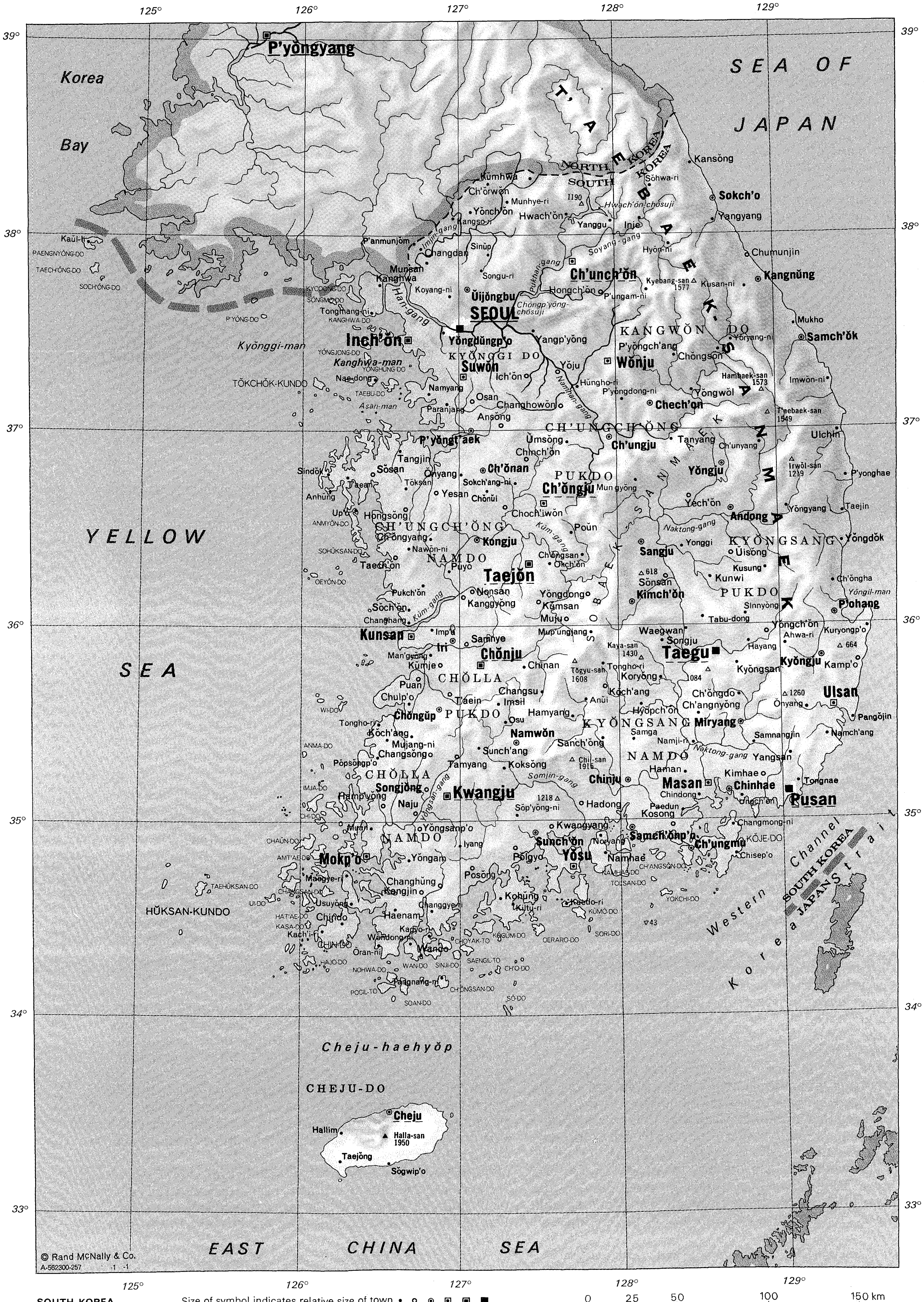
The name of a political subdivision if not shown on the map is the same as that of its capital city.

Cities and towns

Ahwa-ri	35·54n	129·02e
Andong	36·35n	128·44e
Anhŭng	36·41n	126·10e
Ansŏng	37·02n	127·16e
Anŭi	35·38n	127·48e
Changdan	37·56n	126·45e
Changgye-ri	34·33n	126·49e
Changhang	36·01n	126·40e
Changhowŏn	37·08n	127·39e
Changhŭng	34·41n	126·52e
Changmong-ni	34·58n	128·41e
Ch'angnyŏng	35·33n	128·29e
Changsŏng	35·20n	126·49e
Changsu	35·40n	127·32e
Cheju	33·31n	126·32e
Chinan	35·48n	127·25e
Chinch'ŏn	36·52n	127·26e
Chindo	34·28n	126·15e
Chindong	35·07n	128·29e
Chinhae	35·09n	128·40e
Chinju	35·11n	128·05e
Chisep'o	34·50n	128·42e
Choch'iwŏn	36·37n	127·18e
Ch'ŏnan	36·48n	127·09e
Ch'unch'ŏn	37·52n	127·43e
Ch'ŏngdo	35·38n	128·43e
Ch'ŏngha	36·13n	129·20e
Ch'ŏngju	36·39n	127·31e
Ch'ŏngsan	36·22n	127·46e
Chŏngsŏn	37·22n	128·38e
Chŏngŭp	35·36n	126·51e
Ch'ŏngyang	36·27n	126·48e
Chŏnju	35·49n	127·08e
Chŏnŭi	36·42n	127·11e
Chŏrwŏn	38·16n	127·12e
Chulp'o	35·37n	126·40e
Chumunjin	37·54n	128·49e
Ch'ungju	36·58n	127·58e
Ch'ungmu	34·51n	128·25e
Ch'unyang	36·56n	128·54e
Hadong	35·05n	127·44e
Haenam	34·34n	126·35e
Hallim	33·24n	126·16e
Haman	35·15n	128·24e
Hamp'yŏng	35·05n	126·30e
Hamyang	35·32n	127·42e
Hayang	35·55n	128·47e
Hongch'ŏn	37·42n	127·52e
Hongsŏng	36·36n	126·39e
Hŭngho-ri	37·14n	127·44e
Hwach'ŏn	38·06n	127·41e
Hyŏn-ni	37·57n	128·20e
Hyŏpch'ŏn	35·35n	128·08e
Ich'ŏn	37·17n	127·27e
Imp'a	35·59n	126·49e
Imsil	35·37n	127·15e
Imwŏn-ni	37·15n	129·20e
Inch'ŏn	37·28n	126·38e
Inje	38·05n	128·09e
Iri	35·56n	126·57e
Iyang	34·53n	127·00e
Kach'i-ri	34·27n	126·08e
Kaedo-ri	34·35n	127·39e
Kagyo-ri	34·25n	126·48e
Kamp'o	35·48n	129·29e
Kanggyŏng	36·10n	127·00e
Kanghwa	37·45n	126·30e
Kangjin	34·39n	126·45e
Kangnŭng	37·45n	128·54e
Kangsŏ-ri	38·06n	126·58e
Kansŏng	38·22n	128·29e
Kaŭl-li	37·58n	124·37e
Kiltu-ri	34·35n	127·20e
Kimch'ŏn	36·07n	128·05e
Kimhae	35·14n	128·52e
Kŏch'ang	35·26n	126·42e
Kŏch'ang	35·41n	127·55e
Kohŭng	34·37n	127·16e
Koksŏng	35·17n	127·17e
Kongju	36·27n	127·07e
Koryŏng	35·44n	128·15e
Kosŏng	34·58n	128·18e
Koyang-ni	37·42n	126·56e
Kŭmje	35·48n	126·52e
Kŭmsan	36·07n	127·30e
Kunsan	35·58n	126·41e
Kunwi	36·15n	128·34e
Kuryongp'o	35·59n	129·32e
Kusan-ni	37·43n	128·49e
Kusung	36·18n	128·55e
Kwangju	35·09n	126·54e
Kwangyang	34·59n	127·34e
Kyŏngju	35·51n	129·14e
Kyŏngsan	35·48n	128·43e
Maegye-ri	34·45n	126·18e
Man'gyŏng	35·52n	126·48e
Masan	35·11n	128·32e
Miryang	35·31n	128·44e
Mokp'o	34·48n	126·22e
Muan	34·58n	126·26e
Mujang-ni	35·26n	126·32e
Muju	36·02n	127·40e
Mukho	37·33n	129·06e
Munhye-ri	38·10n	127·19e
Mun'gyŏng	36·44n	128·07e
Munsan	37·51n	126·48e
Mup'ungjang	35·58n	127·49e
Nae-dong	37·16n	126·27e
Naju	35·03n	126·43e
Namch'ang	35·26n	129·16e
Namhae	34·50n	127·54e
Namji-ri	35·23n	128·29e
Namwŏn	35·25n	127·21e
Namyang	37·14n	126·44e
Nawŏn-ni	36·25n	126·40e
Nonsan	36·12n	127·05e
Noryang	34·56n	127·52e
Okch'ŏn	36·20n	127·34e
Ŏnyang	35·34n	129·07e
Ŏnyang	36·47n	127·00e
Ŏran-ni	34·22n	126·29e
Osan	37·11n	127·04e
Osu	35·31n	127·18e
Paedun	35·03n	128·21e
Pangŏjin	35·29n	129·26e
Paranjang	37·08n	126·55e
P'ohang	36·03n	129·20e
Pŏlgyo	34·52n	127·21e
Pŏpsŏngp'o	35·20n	126·27e
Posŏng	34·47n	127·04e
Poŭn	36·29n	127·43e
Puan	35·45n	126·44e
Pukch'ŏn	36·13n	126·45e
P'ungam-ni	37·43n	128·11e
Pusan	35·06n	129·03e
Puyŏ	36·18n	126·54e
P'yŏngch'ang	37·23n	128·22e
P'yŏngdong-ni	37·10n	128·02e
P'yonghae	36·46n	129·28e
P'yŏngt'aek	37·00n	127·05e
Samch'ŏk	37·27n	129·10e
Samch'ŏnp'o	34·57n	128·03e
Samga	35·25n	128·05e
Samnangjin	35·23n	128·50e
Samnye	35·55n	127·05e
Sanch'ŏng	35·26n	127·54e
Sangju	36·26n	128·09e
Seoul (Sŏul)	37·33n	126·58e
Sindŏk	36·47n	126·10e
Sinnyŏng	36·04n	128·46e
Sinŭp	37·54n	127·12e
Sŏch'ŏn	36·05n	126·41e
Sŏgwip'o	33·13n	126·34e
Sŏhwa-ri	38·15n	128·13e
Sokch'ang-ni	36·44n	127·17e
Sŏkch'o	38·12n	128·36e
Songjŏng	35·10n	126·46e
Sŏngju	35·55n	128·16e
Songu-ri	37·49n	127·09e
Sŏnsan	36·16n	128·17e
Sŏp'yŏng-ni	35·01n	127·24e
Sŏsan	36·47n	126·26e
Sŏul, see Seoul		
Sunch'ang	35·23n	127·07e
Sunch'ŏn	34·57n	127·28e
Suwŏn	37·17n	127·01e
Tabu-dong	36·03n	128·31e
T'aean	36·46n	126·16e
Taech'ŏn	36·22n	126·34e
Taegu	35·52n	128·35e
T'aein	35·40n	126·55e
Taejin	36·34n	129·24e
Taejŏn	36·20n	127·26e
Taejŏng	33·15n	126·17e
Tamyang	35·21n	126·58e
Tangjin	36·54n	126·37e
Tangnang-ni	34·12n	126·52e
Tanyang	36·57n	128·21e
Tŏksan	36·43n	126·39e
Tongho-ri	35·31n	126·29e
Tongho-ri	35·49n	127·54e
Tongmang-ni	37·37n	126·26e
Tongnae	35·12n	129·05e
Ŭijŏngbu	37·44n	127·03e
Ŭisŏng	36·22n	128·41e
Ulchin	36·59n	129·23e
Ulsan	35·34n	129·19e
Ŭmsŏng	36·56n	127·41e
Ungch'ŏn	35·07n	128·44e
Up'o	36·36n	126·22e
Usuyŏng	34·35n	126·18e
Waegwan	35·58n	128·24e
Wando	34·18n	126·47e
Wŏndong-ni	34·23n	126·40e
Wŏnju	37·22n	127·58e
Yanggu	38·06n	127·59e
Yangp'yŏng	37·30n	127·29e
Yangsan	35·21n	129·03e
Yangyang	38·04n	128·36e
Yech'ŏn	36·40n	128·26e
Yesan	36·41n	126·50e
Yŏju	37·18n	127·37e
Yŏngam	34·48n	126·40e
Yŏngch'ŏn	35·58n	128·56e
Yŏngch'ŏn	38·07n	127·05e
Yŏngdŏk	36·26n	129·23e
Yŏngdong	36·10n	127·48e
Yŏngdŭngp'o	37·31n	126·54e
Yonggi	36·24n	128·24e
Yŏngju	36·50n	128·37e
Yŏngsanp'o	34·58n	126·44e
Yŏngwŏl	37·12n	128·28e
Yŏngyang	36·40n	129·07e
Yŏryang-ni	37·30n	128·43e
Yŏsu	34·46n	127·44e

Physical features and points of interest

Amt'ae-do, *island*	34·50n	126·07e
Anma-do, *island*	35·21n	126·02e
Anmyŏn-do, *island*	36·30n	126·22e
Asan-man, *bay*	36·56n	126·51e
Changsan-do, *island*	34·38n	126·10e
Ch'angsŏn-do, *island*	34·52n	128·01e
Chaŭn-do, *island*	34·53n	126·03e
Cheju-do, *island*	33·20n	126·30e
Cheju-haehyŏp, *strait*	33·40n	126·28e
Chi-do, *island*	35·04n	126·13e
Chii-san, *mountain*	35·20n	127·44e
Chin-do, *island*	34·25n	126·15e
Ch'o-do, *island*	34·14n	127·15e
Chŏngp'yŏng-chŏsuji, *lake*	37·40n	127·30e
Ch'ŏngsan-do, *island*	34·11n	126·54e
Choyak-to, *island*	34·22n	126·54e
East China Sea	32·45n	126·30e
Hajo-do, *island*	34·17n	126·03e
Halla-san, *mountain*	33·22n	126·32e
Hambaek-san, *mountain*	37·09n	128·55e
Han-gang, *river*	37·45n	126·11e
Hat'ae-do, *island*	34·32n	126·03e
Hŭksan-kundo, *islands*	34·30n	125·20e
Hwach'ŏn-chŏsuji, *lake*	38·07n	127·52e
Imja-do, *island*	35·05n	126·05e
Imjin-gang, *river*	37·47n	126·40e
Irwŏl-san, *mountain*	36·50n	129·06e
Japan, Sea of	38·30n	129·30e
Kanghwa-do, *island*	37·40n	126·27e
Kanghwa-man, *bay*	37·20n	126·35e
Kasa-do, *island*	34·28n	126·03e
Kŏgŭm-do, *island*	34·27n	127·11e
Kŏje-do, *island*	34·52n	128·37e
Korea Strait	34·00n	129·00e
Kŭm-gang, *river*	36·00n	126·40e
Kŭmŏ-do, *island*	34·31n	127·45e
Kyebang-san, *mountain*	37·43n	128·29e
Kyodong-do, *island*	37·45n	126·16e
Kyŏnggi-man, *bay*	37·25n	126·00e
Naktong-gang, *river*	35·07n	128·57e
Namhae-do, *island*	34·48n	127·57e
Namhan-gang, *river*	37·31n	127·18e
Nohwa-do, *island*	34·12n	126·35e
Oeraro-do, *island*	34·27n	127·30e
Oeyŏn-do, *island*	36·14n	126·05e
Paengnyŏng-do, *island*	37·57n	124·40e
Pogil-to, *island*	34·09n	126·33e
Pukhan-gang, *river*	37·31n	127·18e
P'yŏng-do, *island*	37·35n	125·43e
Saengil-to, *island*	34·19n	126·59e
Sinji-do, *island*	34·20n	126·50e
Soan-do, *island*	34·09n	126·39e
Sobaek-sanmaek, *mountains*	36·00n	128·00e
Soch'ŏng-do, *island*	37·46n	124·45e
Sŏ-do, *island*	34·02n	127·18e
Sohŭksan-do, *island*	34·04n	125·07e
Somjin-gang, *river*	34·58n	127·46e
Songmo-do, *island*	37·42n	126·18e
Sori-do, *island*	34·26n	127·48e
Soyang-gang, *river*	37·52n	127·40e
T'aebaek-san, *mountain*	37·06n	128·55e
T'aebaek-sanmaek, *mountains*	37·20n	128·50e
Taebu-do, *island*	37·15n	126·35e
Taech'ŏng-do, *island*	37·49n	124·43e
Taehŭksan-do, *island*	34·40n	125·25e
Tŏgyu-san, *mountain*	36·48n	127·23e
Tŏkchŏk-kundo, *islands*	37·14n	126·07e
Tolsan-do, *island*	34·38n	127·45e
Ui-do, *island*	34·37n	125·51e
Wan-do, *island*	34·21n	126·42e
Western Channel	34·40n	129·00e
Wi-do, *island*	35·36n	126·17e
Yellow Sea	36·00n	125·00e
Yokchi-do, *island*	34·38n	128·15e
Yŏngil-man, *bay*	36·02n	129·26e
Yŏngjong-do, *island*	37·30n	126·31e
Yŏngsan-gang, *river*	34·54n	126·32e



SOUTH KOREA

Size of symbol indicates relative size of town • ○ ◎ ▣ ▣ ■

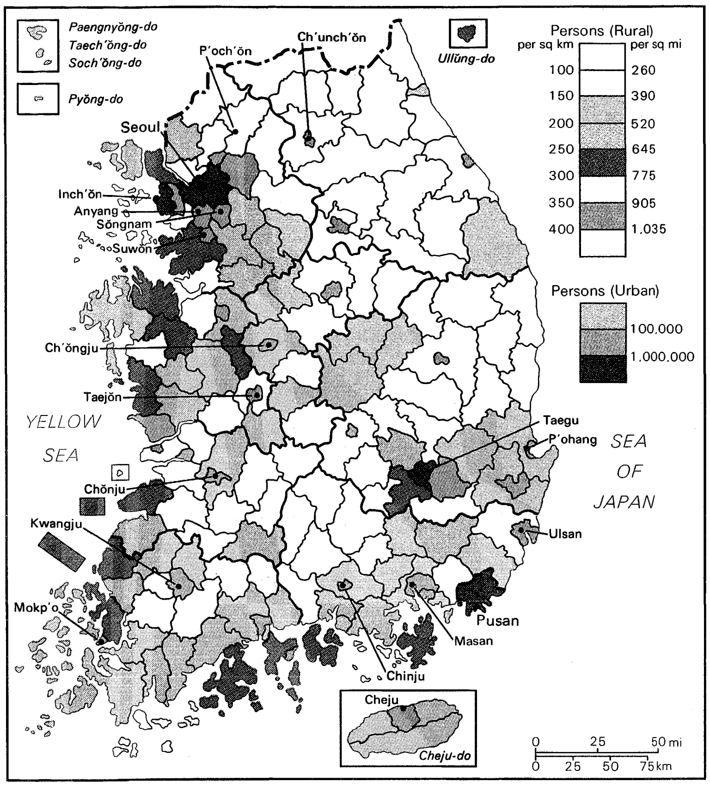
Elevations in metres

0 25 50 100 150 km

0 25 50 100 mi

The economy. The South Korean economy has progressed rapidly in the manufacturing and mining sectors under its economic development plans. Export-oriented manufacturing, which has received strong government support, has been particularly successful, and the country's balance of payments has steadily improved, although a payments deficit continued to exist into the early 1980s.

Resources. Natural resources in South Korea are meagre. The leading resources are coal, iron ore, graphite, gold and silver, tungsten, lead, and zinc, comprising almost two-thirds of the total value of mineral resources. Iron ore is exported mainly to Japan, and tungsten is exported to the United States. Both are produced in Kangwŏn Province. Gold and silver are produced mostly in Ch'ungch'ŏng and Kyŏngsang provinces.



Urban population distribution and rural population density of South Korea.

Energy resources consist mainly of coal, petroleum, and hydroelectric potential. Anthracite coal is the major exploited energy resource, production having increased rapidly after 1960. Almost three-fourths of the hydroelectric power stations are located along the Han River, not far from Seoul. Thermal electric-power stations are located in the urban centres, with the exception of one near the coal-producing area. Since the first oil refinery started to produce petroleum products in 1964, power stations have changed gradually from coal to oil. In contrast to North Korea, thermal electric power is far more important than hydroelectric power.

Agriculture and forestry. Both the farm population and the proportion of national income from agriculture are gradually decreasing, with less than one-third of the population engaged in agriculture. More than one-seventh of the gross national product comes from agriculture, forestry, and fishing; and less than one-fourth of the republic's area is cultivated for rice, barley, wheat, soybeans, and other crops. Rice is the most important crop, constituting about two-fifths of all farm products in value. Double cropping of rice and barley is common in Kyŏngsang and Chŏlla provinces.

Development of industry

Industry. The economic development plans stress improvement of mining and manufacturing. Heavy industry, including chemical, metal, and machinery industries, continues to be developed and constitutes about one-third of all industrial products in value. The textile industry has been the most important single industry in terms of value and employment, although its share of the market has decreased. South Korea's principal export markets are the United States, Japan, and Southeast Asian countries.

The government exercised strong controls on industrial development after 1962, giving most support to such large industrial establishments as fertilizer plants and oil refineries. As a result, small and middle industries that were privately managed became increasingly difficult to finance, and consumer goods and consumer spending were discouraged. By the beginning of the Fifth Republic in 1980, however, these trends began to be reversed, especially in credit policies, as the government increasingly divested itself of direct involvement in industry.

Transportation. South Korea's transportation system has expanded and improved to a considerable extent, especially with the introduction of modern highway and air services. It has not met the needs of the country, however, as indicated by congested urban transportation facilities. Bus transportation networks are well developed and serve most of the rural centres. In the rural areas agricultural products are still hauled by oxcart and by human labour. Internal air transportation began in the early 1970s under the aegis of Korean Air Lines. Major cities, such as Pusan, Taegu, Kwangju, Cheju, and Kangnŭng, have scheduled air services. The international airport of Kimp'o, near Seoul, serves most international airlines. In the 1970s the government expanded port facilities at Pusan, Inch'ŏn, and Cheju, which increased their freight-handling capabilities more than fourfold.

The bulk of Korean railroads are government owned. Until 1960 railway travel was the major means of inland transportation for both freight and passengers, but road transport has since become more important. Railroads are almost all of standard gauge, and the Seoul–Pusan line through Taejŏn and the Seoul–Inch'ŏn line are double-tracked. A four-lane highway from Seoul to Inch'ŏn was opened in 1968, and a modern superhighway between Seoul and Pusan was opened in 1970. Road transport has come to account for more than 90 percent of passenger travel and 60 percent of freight transport. (C.Le.)

Administrative and social conditions. *Government.* The government instituted after a constitutional referendum in 1980 is known as the Fifth Republic; that name was retained after another constitutional referendum in 1987. The republic is patterned mainly after the presidential system of the United States and is based on separation of powers among the legislature, the executive, and the judiciary. The government system, highly centralized during the Third and Fourth republics, is less so under the Fifth Republic. The president, since 1987 chosen by direct popular election, is the chief of state, head of the executive branch, and commander of the armed forces. The State Council, the highest deliberative body, is composed of the president, the chairman, the prime minister, the heads of executive ministries, and ministers without portfolio. The prime minister is appointed by the president and approved by the elected National Assembly. Provincial governors are appointed by the central government.

Provisions for presidential government

Extensive constitutional revisions were adopted by referendum in October 1987, modifying considerably revisions that had been adopted in 1980. Under these reforms the president's term of office was limited to one five-year term with no reelection. The powers of the National Assembly, which had been reinstated in 1980 after a period of curtailment, were strengthened, with its 276 members chosen, as previously, by a combination of direct and indirect election to four-year terms. All restrictions on political parties were ended.

South Korea had a two-party system until 1972, when the power of the pro-government party increased substantially and the activity of the opposition was restricted. During the 1980s the opposition was allowed to resume political participation, but it also tended to fragment; thus, the political system became more multiparty in character. The Democratic Justice Party (until 1981 called the Democratic Republican Party) has been the ruling party since its founding in 1963. The New Korea Democratic Party (NKDP), founded in 1985 as a coalition of opposition groups, and the Reunification Democratic Party, an offshoot of the NKDP founded in 1987, have become the major opposition parties. (C.Le./Ed.)

Justice. The judicial branch comprises the Supreme Court, three appellate courts, 12 district courts, and a family court. The Supreme Court is empowered to interpret the constitution and all other state laws and to review the legality of government regulations and activities. The chief justice is appointed by the president with the consent of the National Assembly, upon recommendation of the Judge Recommendation Council.

Education. Six years of primary school education is compulsory, and by 1980 almost all children of school age were enrolled. Most primary school graduates go on to three years of middle school, and many middle school graduates go on to high school or to technical schools. About half of the high-school graduates go to higher educational institutions. Before World War II there were fewer than 20 major college-level institutions in South Korea, but in the decades following the war the number has increased more than fourfold. Admission to the colleges and universities is in most cases granted through competitive entrance examinations.

Health and welfare. The availability of medical services has increased, but medical facilities and the number of personnel are inadequate to meet the country's needs. Improvement has been hampered partly by a continuous migration of medical personnel to foreign countries.

Government welfare activities

Government welfare activities are new and limited in range. The programs include care of disabled war veterans, homes for the aged and for homeless and disabled war widows and orphans, vocational training of women, and care of juvenile delinquents. Since the devastation of the Korean War, United Nations agencies, civilian and military agencies of the United States, and private volunteer agencies have played a significant role in improving living conditions in the south. The shortage of housing remains a problem, but it has been partially solved by central and local government-housing programs in metropolitan areas. The national police force is strongly developed in order to counteract Communist infiltration.

Living conditions in South Korea, relatively poor compared to conditions in developed nations at the end of the Korean War, have improved steadily. From 1968 to 1979 per capita disposable income increased more than sevenfold. Public health and sanitation have been greatly improved, thus reducing epidemics. Average life-expectancy rates rose from about 53 years in the late 1950s to about 70 years by the mid-1980s. Despite improvement of social conditions there is little sign of a narrowing of the gap between rural and urban and low- and high-income groups.

Cultural life. Shamanism, Buddhism, and the philosophy of Confucius constitute the most important background of modern Korean culture. Since World War II, especially after the Korean War in 1950, the modern trends have rapidly progressed. Traditional thought, however, still plays an important role under the surface. Korea belongs to the Chinese cultural realm, although Koreans have maintained a distinct cultural identity throughout their history. The National Museum maintains artifacts of Korean culture, including many national treasures, chiefly in the central museum in Seoul. Its four branches are located at Kyŏngju, Puyŏ, Kongju, and Chŏnju.

Architecture. After the Three Kingdoms period, Korean culture was strongly influenced by the Chinese, although this influence was given a distinctive Korean stamp. Korean architecture shows Chinese influence, but it is adapted to local needs and environment, utilizing wood and granite, the most abundant building materials. Beautiful examples are found in old palaces, Buddhist temples, stone tombs, and Buddhist pagodas.

Painting and ceramics. One of the earliest examples of Korean painting is found in the mural paintings in the kings' tombs of Koguryŏ. The best known mural paintings are those in the Sangyong Tomb, at Yonggang, located in North Korea. Ceramic arts became highly developed, flourishing during the Koryŏ period and diffusing to Japan, and every province continues to produce its distinctive ceramic wares.

Dance and music. Folk dances survive, and folk music, accompanied by native musical instruments, is performed occasionally at ceremonies and festive occasions. The government has made an effort to preserve the traditional arts. The National Classical Music Institute (formerly the Prince Yi Conservatory), for example, plays an important role in the preservation of folk music. It has had its own training centre for national music since 1954. The Korean National Symphony Orchestra and the Seoul Symphony Orchestra give concerts in Seoul and Pusan. (C.Le.)

For statistical data on the land and people of South Korea, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL.

HISTORY

The end of Japanese rule after World War II caused political confusion among Koreans in both zones. In South Korea various political parties sprang up. Although they were roughly divided into rightists, leftists, and middle-of-the-roaders, they had a common goal: the immediate attainment of self-government. As early as August 16, 1945, some Koreans organized a Committee for the Preparation of Korean Independence. On September 6 the People's Republic of Korea was proclaimed by delegates attending a national assembly called by the committee. The republic was headed by Woon-hyung Lyuh, who was closely associated with the leftists. But the U.S. military government, under Lt. Gen. John R. Hodge, the commanding general of the United States armed forces in Korea, refused to recognize the republic, asserting that the military government was the "only government" in Korea, as stipulated in General Order No. 1. The U.S. policy in Korea was to establish a trusteeship that would supersede both the U.S. and the Soviet occupation forces in Korea. The exiled Korean provisional government, on returning, declared itself a political party, not a government.

In late December the Council of Foreign Ministers (representing the United States, the Soviet Union, and Great Britain) met in Moscow and decided to create a four-power trusteeship of up to five years. On receiving the news, Koreans reacted violently. On February 14, 1946, to soothe the discontent, the military government created the Representative Democratic Council as an advisory body to the military government. This body was composed of Koreans and had as its chairman Syngman Rhee, former president of the Korean government in exile.

In October the military government created an Interim Legislative Assembly, half of whose members were elected by the people and half appointed by the military government. The assembly was empowered to enact ordinances on domestic affairs but was subject to the veto of the military government. The anti-trusteeship feeling came to a climax a few months later, when the assembly condemned trusteeship in Korea.

Elections were held in South Korea on May 10, 1948, under the supervision of the United Nations Temporary Commission on Korea. The National Assembly met for the first time on May 31 and elected Syngman Rhee as its speaker. In July a constitution was adopted, and Rhee was elected president. The Republic of Korea was inaugurated on August 15, and the military government ended. On December 12 the UN recognized the new government as the only lawful government in Korea.

The Korean War. South Korea began to organize a police constabulary reserve in 1946. On December 14, 1948, the Department of National Defense was established. By June 1950, when the war broke out, South Korea had a 98,000-man force equipped only with small arms, which was barely enough to deal with internal revolt and border attacks. The United States occupation forces completely withdrew from Korea by June 29, 1949, leaving behind them a force of about 500 men as a U.S. Military Advisory Group to the Republic of Korea to train the South Korean armed forces. On June 7, 1949, in a message to the United States Congress calculating the security measures necessary for the protection of the Pacific from Communist domination, Pres. Harry S. Truman declared that Korea had become a testing ground in the ideological conflict between Communism and democracy. On that ground, he asserted, U.S. aid should be granted with a long-range program. On October 6, 1949, the United States granted South Korea $10,200,000 for military aid and $110,000,000 for economic aid for the fiscal year 1950, the first year of a contemplated three-year program. In addition, the U.S. Congress approved $10,970,000 for military aid on March 15, 1950. The military equipment committed under the U.S. military-assistance program was still en route, however, when North Korean troops invaded the South. South Korea was unprepared to resist the total invasion from the North.

Early in 1946 the Soviet authorities in North Korea had organized a 20,000-man constabulary and army units, and in August 1946 the North Korean Army was established, its title being changed to the Korean People's Army in February 1948. The Soviet occupation forces left North Korea in December 1948, leaving behind 150 advisors for each army division for training purposes. On March 17, 1949, the U.S.S.R. concluded a reciprocal-aid agreement with North Korea in which it agreed to furnish heavy military equipment; and by June 1950 North Korean forces numbered 135,000, including a tank brigade. As early as 1946, the Soviets were sending thousands of Koreans to the U.S.S.R. for specialized training, and during 1949–50 the People's Republic of China transferred about 12,000 Korean troops from its army to the North Korean forces. The North Korean forces were thus far superior to the forces of South Korea in training and equipment.

Cost of the war

The North Korean troops launched a full-scale invasion of South Korea on June 25, 1950. The war concluded with an armistice on July 27, 1953, having lasted for three years and one month and having accounted for about 4,000,000 casualties, including civilians. South Korean casulties were some 1,313,000 (1,000,000 civilians); Communist casualties were estimated at about 2,500,000 (including 1,000,000 civilians). The United States lost 33,629 dead in action, South Korea 47,000, and the UN forces 3,194; but the estimated losses of the People's Republic of China in action were 900,000 men and of North Korea 520,000. During the war, 43 percent of Korea's industrial facilities was destroyed and 33 percent of its homes devastated.

UN intervention. On June 26 (June 25, New York time) the UN Security Council approved a resolution describing the invasion of South Korea as a "breach of the peace and an action of aggression" and called upon the members to render every assistance in restoring peace. The Soviet Union was unable to impose a veto because its delegate had been boycotting the meetings to protest the fact that the People's Republic of China had no seat in the United Nations. On June 27 President Truman issued the order for United States air and naval forces to resist Communist aggression in Korea, and that afternoon the Security Council of the United Nations ratified Truman's decision to send air and sea aid to Korea, calling upon the UN members to render such assistance to Korea as might be necessary to restore peace. But Seoul, the South Korean capital, fell on June 28, and most of the South Korean Army was destroyed. On June 30 Truman ordered United States ground forces in Japan into Korea; the first U.S. troops reached the battlefield on July 4. The UN approved the creation of a unified command in Korea, and Gen. Douglas MacArthur was appointed commander. Sixteen member nations sent armed contingents, but the United States furnished the great bulk of the air units, naval forces, supplies, and money.

The North Koreans continued to advance recklessly despite the presence of U.S. troops in the field. In early August the UN retreat came to an end in a defense perimeter along the Naktong-gang (Naktong River) line (forming a semicircle in southeast Korea). South Korea was now almost overrun by the North Koreans, except for the small beachhead around Pusan-hang (Pusan Harbor, in Korea's extreme southeast). On September 15 MacArthur counterattacked, catching the Communists on the flank by an amphibious attack on Inch'ŏn (on the coast west of Seoul). They were trapped and either surrendered or fled in panic. By October 1 the UN forces were back at the 38th parallel. On September 27 the U.S. Joint Chiefs of Staff ordered MacArthur to destroy the North Korean armed forces, and two days later Truman authorized him to advance into North Korea. On October 7 the UN General Assembly approved the resolution to permit entry into North Korea and created a UN Commission for the Unification and Rehabilitation of Korea. On October 20 the UN forces entered P'yŏngyang, the North Korean capital, and on October 26 reached the Manchurian border at the Yalu River.

Chinese intervention. The Chinese Communists, who had moved troops along the Yalu after the Inch'ŏn landing, in November entered Korea in overwhelming numbers. By the end of 1952, 1,200,000 Chinese were engaged in the war under the command of P'eng Te-huai. They forced the UN forces to retreat in disorder, and Seoul was re-evacuated on January 4, 1951. But around P'yŏngtaek (about 30 miles south of Seoul) the Chinese were halted, and in February the UN General Assembly formally condemned the People's Republic of China as an aggressor. The UN counteroffensive began in late January. By March 31 the UN forces had again reached the 38th parallel. MacArthur now publicly advocated an extension of the war to China because of the Chinese intervention, but this advocacy was regarded as a challenge to the United States president's conduct of foreign policy. Consequently, on April 11, Truman dismissed MacArthur from all of his commands, and Gen. Matthew B. Ridgway took his place. From then until the armistice, the UN forces fought a holding action along the 38th parallel; indeed, in many places the UN forces were slightly north of the line.

Armistice and aid. The Soviet delegate to the United Nations proposed a discussion of a cease-fire and an armistice on June 23, 1951, and on July 10 negotiations began between the United Nations and the Communist commanders at Kaesŏng, later resumed at P'anmunjŏm (both about 30 miles northwest of Seoul; the former in North Korea, the latter in South Korea). Many issues stood between the two negotiators, the first being the Chinese demand that all foreign troops be withdrawn from Korea in the face of the steadfast refusal of the United States to withdraw all UN troops from South Korea. The second issue was the boundary: the Communists demanded the restoration of the 38th parallel, but the United States insisted on the existing battle line. The third and most important issue was that of prisoners. The UN forces held 171,000 prisoners, 50,000 of them unwilling to return to their Communist countries. The Communists, not to lose face, were determined to have all prisoners back. On this matter the negotiations were deadlocked and did not resume until after the death of Joseph Stalin in March 1953. The United States administration under Dwight D. Eisenhower was inaugurated in early 1953 and, deeply concerned with balancing the budget, was determined to end the impasse even if this involved resumption of hostilities. On the other hand, the war weariness of the Communists was increasing. In April the first 6,670 Communists and the 684 UN personnel were exchanged at P'anmunjŏm. Soon the prospects for armistice negotiations seemed to improve. The Communists agreed to hand over to a neutral commission the UN-held prisoners of war who did not wish to be repatriated. But Syngman Rhee opposed any term that would leave Korea divided and demanded that the military offensive be resumed. On June 18 Rhee suddenly released 27,000 North Korean anti-Communist prisoners in defiance of the United Nations, whereupon the Communists broke off negotiations. On July 20 negotiations were resumed. Rhee gave in and agreed to support the armistice even though he would not sign it. In return the United States promised to extend economic aid and conclude a mutual-security pact to protect South Korea against further aggression.

Negotiating difficulties

Armistice signed

The armistice was signed on July 27, 1953. The United Nations had won most of its demands. The military line became the boundary between North and South Korea, and commissions were established to enforce the cease-fire regulations. A Neutral Nations Commission for Repatriation was entrusted with the repatriation of prisoners, 21,809 of whom—among them 7,582 Korean and 14,227 Chinese—chose to stay in South Korea or go to Taiwan.

The U.S. Army provided Korea with $181,200,000 during the occupation period of 1946–48. This money, which was provided under the assistance programs for occupied areas, was spent mainly on preventing hunger and disease. For the period of 1949–52, the U.S. provided $485,600,000 for economic aid and $12,500,000 for military aid. Following the war the UN Korean Reconstruction Agency (UNKRA) was established to carry out economic aid to South Korea, 34 member and five nonmember states contributing $148,500,000. The UNKRA came to an end in 1958, but meanwhile UN Emergency Relief also contributed $474,400,000 and other international voluntary agencies $85,000,000. Most of the UN contributions were provided by the United States, and total U.S. aid to South Korea from 1946 to 1978 exceeded $5,893,900,000.

Postwar South Korea. The political experience of the people of the Republic of Korea since 1948 represents a wide range of variation. They lived under 12 years of Syngman Rhee's authoritarian rule, tempered by the rise of a vocal opposition coalition under semicompetitive conditions; the collapse of this system after student demonstrations that brought the nation close to civil war in April 1960; nine months of multiparty liberalism under the cabinet system of Prime Minister Chang Myŏn that permitted competition among conservative and emergent socialist parties; and, after May 1961, a presidential system dominated by the armed forces, which had maintained a tight and effective rule over Korea.

Early disagreement over president's role

The First Republic. The first Korean republic, established in 1948, adopted a presidential system and elected Syngman Rhee as the first president. He was re-elected in August 1952 while the nation was still at war. Even before the outbreak of the Korean War there had been a serious conflict between Rhee and the opposition-dominated National Assembly that elected him in 1948. The dispute involved a constitutional amendment bill that the opposition introduced in an attempt to defeat Rhee by replacing the presidential system with a parliamentary cabinet system. The bill was defeated, but the dispute was carried to Pusan, the wartime provisional capital, where the National Assembly was reconvened.

When the opposition introduced another amendment bill in favour of a parliamentary system, Rhee counteracted by pushing through an amendment bill that provided for the popular election of the president. Later, in 1954, Rhee succeeded in forcing the National Assembly, then dominated by the ruling party, to pass a constitutional amendment bill providing a life-term presidency for himself. On the basis of the revised constitution, Rhee was able to run successfully for his third term of office in May 1956. Rhee's election for the fourth time, in March 1960, was preceded by a period of tension and violence, followed by student demonstrations resulting in many casualties. Rhee resigned under pressure and fled to exile in Hawaii, where he died in 1965 at the age of 90.

The Second Republic. The second Korean republic, which adopted a parliamentary system in place of the presidential system, lasted only nine months before it was overthrown by a military coup in May 1961. With a figurehead president elected by both houses of the legislature, power was shifted to the office of Prime Minister Chang Myŏn, who was elected by the lower house by a narrow margin of 10 votes.

In spite of some strenuous efforts to initiate reforms in a society already blemished by social and economic ills accumulated over a long period of time, the Chang regime was unsuited to deal with the explosive situation created as an aftermath of a violent political change. To make the situation worse, factionalism came to prevail in political life. As the ultimate source of authority now shifted to the office of the prime minister, constant efforts were made by conservative and more moderate factions to coalesce with a group of independents, either to keep or upset the majority within the legislature. Even before the Chang regime launched a full program of economic reform, the Democratic elite was seriously weakened by these factional struggles within its ranks.

Military junta of 1961

The military coup. With the military seizure of political power, postliberation Korean politics entered a new phase characterized by inflow of military officers into the government. The military junta that took over the government on May 16, 1961, dissolved the National Assembly and imposed a strict prohibition on political activity. The nation was then placed under martial law, and the Supreme Council for National Reconstruction (SCNR), headed by Maj. Gen. Park Chung Hee, took the reins and began to establish a new system of national government.

In November 1962 the SCNR made public a constitutional amendment bill that provided for a strong president and a weak, single-chamber National Assembly. The bill was approved by a national referendum held one month later. Then, in February 1963, Chairman Park of the SCNR issued a statement that he would not take part in the civilian government to be formed later in the year if civilian political leaders would uphold a nine-point "political stabilization proposal." On February 27, 1963, Chairman Park made a public pledge not to take part in the new government, for 53 political and military leaders had sworn to support his stabilization proposal.

In March, however, following bitter turbulence in the ruling junta itself and a chaotic situation created by the proliferation of minor political parties, Chairman Park proposed that military rule be extended for four years. The proposal met vigorous opposition from civilian political leaders, but some 160 military commanders, most of general officer rank, submitted a resolution to Park supporting the extension. Under considerable pressure, Park again changed course and announced, on April 8, 1963, a plan for holding elections toward the end of the year. In late May, Park was named presidential candidate of the Democratic Republican Party.

Election of Park Chung Hee

The Third Republic. The election for the president of the Third Republic took place on October 15, 1963. Park, by a narrow margin, defeated the opposition candidate, Yun Po-sun, former president of the Second Republic, who had remained in office (more or less as a figurehead) at the request of the junta to provide constitutional continuity for the military regime. When political activity was permitted to resume, Yun took the initiative in mustering opposition groups and became the presidential candidate of the Civil Rule Party. In May 1967 Park was elected to his second term of office, and the Democratic Republican Party won a large majority of the seats in the National Assembly. Members of the opposition New Democratic Party, whose candidate, Yun, had been defeated for the second time, claimed election fraud and refused to take their seats in the National Assembly until some months later.

During his second term, President Park was confronted with the constitutional provision that limited the president to two consecutive four-year terms. After considerable political turmoil and demonstrations by the opposition and students, Democratic Republican Party members in the legislature passed an amendment that provided for the incumbent president to become eligible for three consecutive four-year terms. The amendment was approved by a national referendum in October 1969. In the presidential elections held on April 27, 1971, Park defeated his opponent, Kim Dae-jung of the New Democratic Party; in elections for the eighth National Assembly, however, the New Democrats made impressive gains, securing 89 seats as compared with the 113 seats of the ruling Democratic Republicans.

The Yushin system. In December 1971, shortly after his election to a third term of office, Park declared a state of national emergency on the basis of the "dangerous uncertainties of the international situation," and ten months later, in October 1972, he suspended the constitution and dissolved the legislature. A new constitution, which extended the power and permitted the re-election of the president for an unlimited number of six-year terms, was promulgated in December.

National Conference for Unification

The institutional framework of the Yushin (Revitalization–Reform) system departed radically from the Third Republic. The National Conference for Unification was created as "a national organization based on the collective will of the people as a whole to pursue peaceful unification of the fatherland." The conference is a body of not less than 2,000 nor more than 5,000 members, who are directly elected by the voters for a six-year term. The president is the chairman of the conference; and until 1987 the conference was charged with the election of the president. Other functions of the conference include approving a list of one-third of the members of the National Assembly, who are appointed by the president for a three-year term (other members of the legislature are elected directly by the voters for a six-year term), and approving constitutional amendments proposed by the National Assembly. Under this arrangement, Park was elected without opposition in December 1972 and was reelected in December 1978.

Under the Yushin system, the Park government continued to achieve a high rate of economic growth. Despite the balance-of-payments difficulties in the mid-1970s, a consequence of the petroleum crisis of 1973, and a much less favourable international environment than expected, growth in gross national product averaged 11.2 percent a year during the third Five-Year Economic Plan (1972–76). Although the net growth in world trade in manufactured goods was sluggish during 1973–76, Korea's export volume more than doubled. At the same time, Koreans experienced great success in obtaining construction contracts in the Middle East. (B.-h.H.)

The assassination of Park. In October 1979 Korea faced a major national crisis that culminated in the assassination of President Park. Prime Minister Choi Kyu-hah became acting president under Article 48 of the Yushin constitution and was elected president in December. Although the initial impression was that a liberalization program was about to be put into effect, the country slipped quickly into stern military rule again. The military did away with all trappings of civilian government in May 1980, extending martial law, banning all political activity, and closing universities and colleges.

The Fifth Republic

In August 1980, following a period of internal upheaval, Chun Doo Hwan was elected president with the promise of a new constitution and a new government based on free elections. The new constitution was approved in October, ushering in the Fifth Republic; the powers of the president were reduced in favour of the National Assembly, and the president was limited to one seven-year term. In January 1981 martial law was lifted, and in February Chun won election with an overwhelming majority. Under Chun the nation's economy grew at a tremendous rate, and South Korea emerged as a strong competitor with Japan for foreign markets. The political climate became more relaxed, but Chun's administration had to endure several scandals and incidents—most notably a bombing in Rangoon (October 1983), which killed several members of the South Korean government—that shook government confidence.

By 1987 popular dissatisfaction with the government had become widespread. The government agreed to draw up another constitution, which was approved by national referendum in October. Among the new document's principal provisions were a reduction in the presidential term from seven to five years and the direct popular election of the president. Roh Tae Woo was elected president in December and took office in February 1988. Later in the year South Korea hosted the Summer Olympic Games in Seoul. (B.-h.H./Ed.)

For later developments in the history of South Korea, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

North Korea

The Democratic People's Republic of Korea (Chosŏn Minjujuŭi In'min Konghwaguk) occupies the northern section of the Korean peninsula, which juts out from the Asian mainland between the Sea of Japan (East Sea) and the Yellow Sea. The country is bordered by China and the Soviet Union to the north and by the Republic of Korea (South Korea) to the south.

North Korea has an area of 46,800 square miles (121,200 square kilometres), occupying about 55 percent of the peninsula. The national capital, P'yŏngyang, is a major industrial and transport centre near the west coast.

North Korea was created in 1948 as a result of the post-World War II military occupation by the United States in the south of the peninsula and by the Soviet Union in the north. In close alliance with China and the Soviet Union, the North Korean government has endeavoured to transform the prewar agricultural country into a self-sufficient industrial nation. Economic programs have been successful, but the industrialization has been at the expense of agriculture and of personal freedom. The government, run by the Korean Workers Party, has attempted to achieve strict production quotas with the use of forced labour, socialization of the entire economy, and ideological indoctrination.

Foreign relations are oriented toward the Communist countries and those non-Communist nations such as Japan that are active in the Afro-Asian solidarity movement. Close economic and political ties have been maintained with the Soviet Union and China, while a hostile attitude has been sustained toward the United States and, generally, South Korea.

PHYSICAL AND HUMAN GEOGRAPHY

The Kaema Plateau and Nangnim-sanmaek

The land. *Relief.* Mountains and valleys characterize most of the country. The Kaema Plateau in the northeast has an average elevation of 3,300 feet (1,000 metres) above sea level and forms the topographic roof of the entire Korean peninsula. Paektu-san (9,003 feet [2,744 metres]), the highest mountain in North Korea, rises at the northern edge of this plateau; it is an extinct volcano topped by a large crater lake. The Nangnim-sanmaek (Nangnim Mountain Range) runs from north to south through the middle of the country, forming a divide between the eastern and western slopes of the peninsula. The Kangnam, Myohyang, Ŏnjin, and Myŏrak mountains, with their roots in the Nangnim-sanmaek, extend parallel to each other toward the southwest. Large river valley plains have developed between the western mountains; they merge along the narrow, irregular coastal plain on the west coast. The Hamgyŏng mountains, extending from the Nangnim-sanmaek to the northeast, form a steep slope between the Kaema Plateau and the Sea of Japan. The T'aebaek-sanmaek of South Korea extends into North Korea; one peak, Kŭmgang-san (5,373 feet [1,638 metres]), is famous for its scenic beauty.

Drainage and soils. The longest river of North Korea is the Yalu, known as the Amnok-kang. It rises at Paektu-san (Paektu Mountain) and flows southwestward for 501 miles (806 kilometres) to its mouth on Korea Bay. The Tumen River also begins at Paektu-san but runs northeastward for 324 miles (521 kilometres) to the Sea of Japan. There are no large streams along the east coast except for the Tumen River, and all the significant rivers, such as the Yalu, Ch'ŏng-ch'ŏn, Taedong, Chaeryŏng, and the Yesŏng, drain to the Yellow Sea. The relatively large valley plains of the western rivers are major agricultural regions.

More than 60 percent of the soils are locally derived from the weathering of granitic rocks or various kinds of schists (crystalline rocks). The soils are generally brownish, abundant in sandy materials, and low in fertility. Well-developed reddish brown soils derived from limestone are found in North Hwanghae Province (Hwanghae-pukto) and the southern part of South P'yŏngan Province (P'yŏngan-namdo). The Kaema Plateau shows development of podzolic soils (ash-gray forest soil) as a result of its cold climate and coniferous forest cover. Although most of the soils are infertile and lacking in organic content, the valley plains have relatively rich alluvial soils.

Climate. North Korea has a generally cool continental climate. The winter season from December to March is long and cold; mean temperatures in January range between 21° F (−6° C) in the south and −8° F (−22° C) in the northern interior. The summer, from June to September, is warm, with mean July temperatures above 68° F (20° C) in most places. Accordingly, the annual range of temperatures is large—about 54° F (30° C) at P'yŏngyang and about 77° F (43° C) at Chung-gangjin, where the lowest temperature in the Korean peninsula, −46.5° F (−43.6° C), has been recorded. Because of ocean currents and the mountain ranges bordering the narrow coastal lowlands, winter temperatures on the east coast are about 4° or 5° F (2° or 3° C) higher than those of the west coast.

Most of the country receives around 40 inches (1,000 millimetres) of precipitation annually. The northern inland plateau, however, receives 24 inches (610 millimetres) and the lower reaches of the Taedonggang (Taedong River) Valley 32 inches (810 millimetres). The upper Ch'ŏngch'ŏn-gang area averages between 48 and 52 inches (1,220 and 1,320 millimetres) yearly. Approximately 70 percent of the annual precipitation falls in the four months from June to September; this heavy concentration of rainfall is related to the humid summer monsoon from the Pacific. Less than 5 percent of the total precipitation occurs in winter. There are about 200 frost-free days along the coast and fewer than 150 inland.

Plant and animal life. Vegetation on the Kaema Plateau, especially around Paektu-san, is composed of coniferous trees, such as the Siberian fir, spruce, pine, and Korean cedar. The western lowlands were originally covered by temperate mixed forests with many types of plants, but continuous deforestation has resulted in only remote patches of the original forests. Most of the lowlands are now cultivated, except for some of the hills that are covered with small pine groves mixed with oaks, chestnuts, and elders. Along streams that are subject to flooding, or where the ground is too stony for cultivation, reeds, sedges, wild mulberry trees, and Italian poplars are found. Common river fish include carps and eels.

Because of deforestation, the population of deer, mountain antelopes, goats, tigers, leopards, and panthers has greatly decreased and is restricted to the remote forests. In the plains, however, it is still possible to see wild pigeons, herons, cranes (which nest near human habitations), and many migratory water fowl, which alight in the rice fields.

Historic provincial divisions

Settlement patterns. Close examination reveals numerous distinct regions, each with a different natural environment and historical background. Of the eight Korean provinces of the Yi dynasty (1392–1910), North Korea contains the three provinces of P'yŏngan, Hwanghae, and Hamgyŏng and the northern parts of Kangwŏn and Kyŏnggi provinces. Each province was not only a political unit but also had characteristics of a cultural region in terms of dialect, customs, and a way of life. North Korea may also be divided into the two larger traditional regions of Kwanso to the west and Kwanbuk to the east, roughly divided by the Nangnim-sanmaek. Kwanso comprises the provinces of North P'yŏngan (P'yŏngan-pukto), South P'yŏngan (P'yŏngan-namdo), North Hwanghae (Hwanghae-pukto), South Hwanghae (Hwanghae-namdo), and Chagang, while Kwanbuk includes North Hamgyŏng, South Hamgyŏng, and Yanggang provinces.

Urbanization increased rapidly after 1953. Most of the rural population inhabits the eastern and western coastal lowlands and river valley plains. The inland areas of Chagang and Yanggang provinces are sparsely settled because of the lack of arable land and the cold climate, which is not suitable for rice culture. Villages in the lowlands and valley plains are usually clustered together on the southern foot of hills, which offer protection against the cold northwestern winter wind. Scattered fire fields are tilled by a small number of shifting cultivators in the Kaema Plateau, especially in Yanggang Province. The Upper Yalu and Tumen river valleys contain settlements associated with lumbering, and fishing villages are numerous along the coast, especially on the east.

Urban settlement

Cities that developed during the Japanese occupation (1910–45) were largely associated with the exploitation of natural resources, industry, and transportation. P'yŏngyang is the hub of the national railway system, Namp'o is a western port, Songnim contains an iron-ore refinery, and Sinŭiju contains factories for the production of electrical equipment, chemicals, textiles, and consumer goods. Mining cities include Aoji, Chaeryŏng, Iwŏn, Kilchu, Musan, and Pukchin. The Communist regime's heavy emphasis on manufacturing resulted in the continuous expansion of the early industrial centres and caused a population flow into the urban areas from the countryside. Most of the cities were destroyed during the Korean War (1950–53) and have since been rebuilt. There are no high-rise buildings, but seven- and eight-story buildings line the main streets of P'yŏngyang. Workers are expected to live in apartments rather than individual homes, and housing projects are supported almost solely by the government. Heating systems in the apartments and urban water supplies are inadequate. The streets are strangely empty of pedestrians, as the Koreans have few leisure hours.

The people. *Ethnic distribution.* Because they are a nation with a long history, Koreans believe that they belong to a single racial stock. Physical and cultural characteristics vary only slightly from one region to another. Koreans all have typical Mongoloid physical features; their average stature is a little shorter than that of the northern Chinese and slightly taller than that of the Japanese. All Koreans speak the Korean language, which is related to Japanese and contains Chinese loan-words. The Korean script, known in North Korea as Chosŏn muntcha (and in South Korea as Han'gŭl), is composed of phonetic symbols for the ten vowels and 14 consonants. In North Korea Chosŏn muntcha has been used exclusively without Chinese characters in newspapers and other publications since 1945.

The stress on industrialization since 1945 has promoted migration to the cities, and the farm labour shortage is severe. North Korea is not concerned with overpopulation because an abundant labour force is needed for its unusually high goals of economic achievement. There is a campaign to repatriate Koreans living in Japan. The coastlines are heavily populated, while the interior is only sparsely so.

Religion. The way of life and the value system of Koreans are based fundamentally on Confucian thought. To a lesser extent, Buddhism is also important. Roman Catholic and Protestant beliefs were introduced in the 17th and 19th centuries; Sŏn-ch'on and P'yŏngyang were major centres of Christian activities. World War II brought repression of Christianity, and all foreign missionaries were expelled from the country.

The monotheistic religion of Ch'ŏndogyo ("Society of the Heavenly Way"; originally known as Tonghak) was founded by the Confucian teacher Ch'oe Che-u in 1860. A combination of Buddhism, Confucianism, and Christianity, Ch'ŏndogyo played a leading role in the independence movement of 1919. Shamanism—the religious belief in gods, demons, and ancestral spirits responsive to a priest, or shaman—existed in Korea before the introduction of Buddhism and Confucianism and still prevails in rural villages.

Religious repression

The Communist regime has constitutionally confirmed freedom of religion but does not practice it, allegedly for fear that it will weaken the Communist Party and the government. Ch'ŏndogyo, however, is used as a means of propaganda. After the Korean War, churches and Buddhist temples were confiscated and looted, and many were converted to other purposes.

The economy. The means of production are socialized, and priorities and emphases in economic development are set by the government. The economy is self-sufficient except for industrial needs such as fuel and machinery. Like other Communist countries, North Korea places special emphasis on capital goods rather than consumer goods.

Resources. North Korea contains about 80 to 90 percent of all known mineral deposits on the peninsula. It is estimated that 200 minerals are of economic value, including gold, tungsten, graphite, magnesite (magnesium carbonate), barite (barium sulfate), and molybdenum (a metallic element used in hardening steel).

Major iron and coal deposits

Iron-ore reserves are estimated at about 2,400,000,000 tons; the deposits at Musan, North Hamgyŏng Province, are of low quality, while those of North Hwanghae Province and South P'yŏngan Province are of high grade. Rich deposits of anthracite (hard coal) occur along the Taedong-gang, not far from P'yŏngyang, and there are small amounts of lignite (brown coal) at Aoji and Anju.

The northern interior contains large forest reserves of larch, spruce, and pine trees. Most of the coastal slopes have been excessively deforested, however, and reforestation programs stress economic forestry. Hydroelectric-power resources were developed highly during the Japanese regime along the Yalu River and its upper tributaries, such as the Changjin, Pujŏn, and Hŏch'ŏn rivers. Power production is based mainly on hydroelectricity, but thermal electricity is becoming important because of the increasing demands of industrialization and the deficiency of hydroelectric power during the dry season.

Farm cooperatives

Agriculture. Despite the disproportionately small agricultural contribution resulting from the labour shortage and low productivity, there has been an increase in cultivated land, irrigation projects, chemical fertilizer supplies, and mechanization through the government program of socialization. By 1958 all farms were incorporated into more than 3,000 cooperatives; each cooperative is comprised of about 300 families on about 1,000 acres (405 hectares). The farm units are controlled by management committees, which issue orders to the work teams, set the type and amount of seed and fertilizer to be used, and establish production quotas. Produce is delivered to the government, which controls distribution through state stores. Each farmer is paid for his labour in money or in kind and is allowed to keep chickens, bees, fruit trees, and a garden.

There are also state and provincial model farms for research and development. The workers are paid in money and are allowed a garden. Livestock husbandry is concentrated on the state farms because the land is little suited to grazing, and few feed grains are grown.

The main food crops are grains, such as rice, corn (maize), millet, barley, and wheat. Although production has increased since the 1950s, grain must still be imported. Sweet potatoes, soybeans, and fruit trees are raised extensively. Industrial crops include tobacco, cotton, flax, and rape (an herb grown for its oilseeds).

MAP INDEX

Political subdivisions

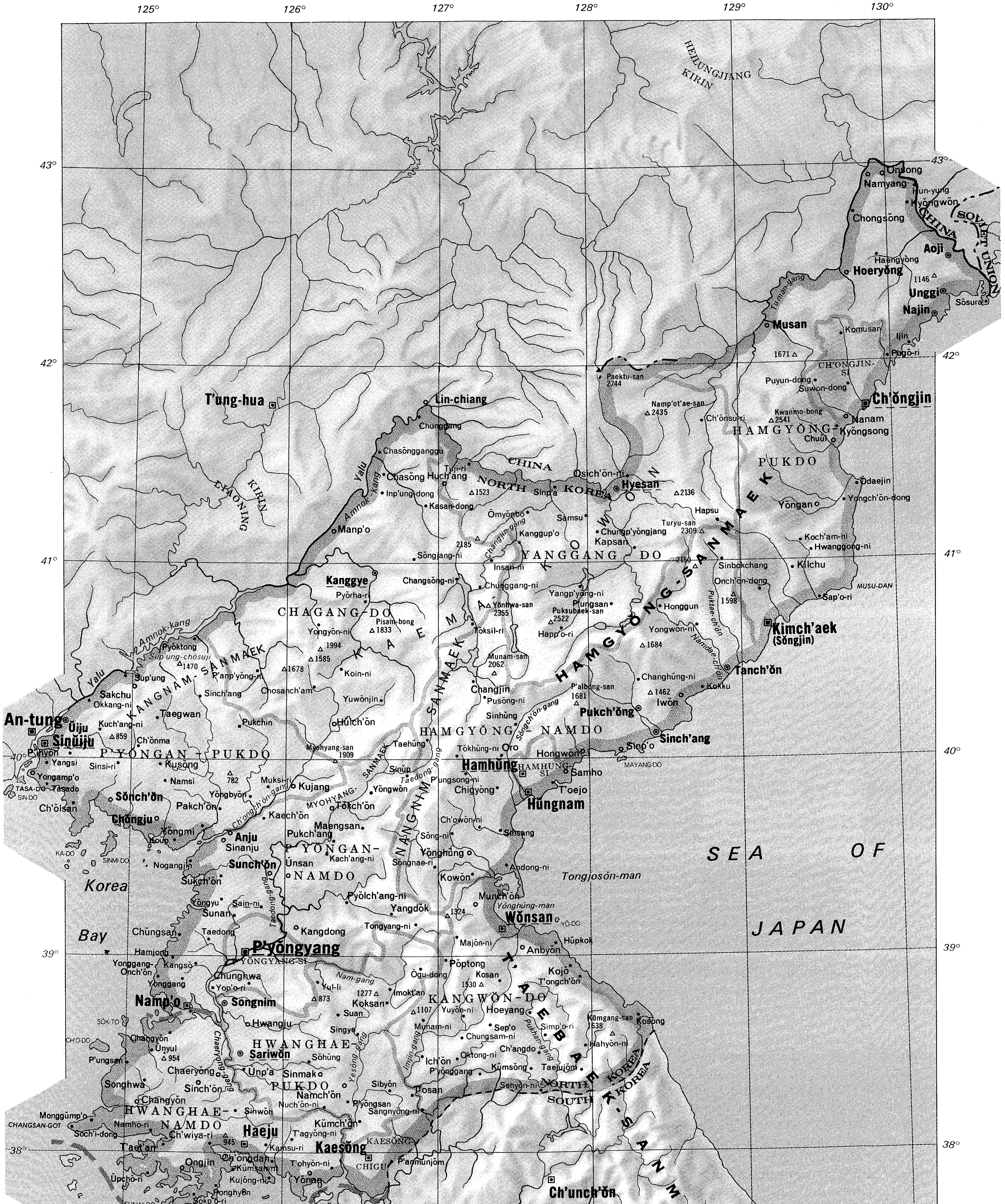
Chagang-do	40·50n	126·30e
Ch'ongjin-si	41·55n	129·40e
Hamgyŏng-namdo	40·00n	127·30e
Hamgyŏng-pukdo	41·45n	129·50e
Hamhung-si	39·55n	127·40e
Hwanghae-namdo	38·15n	125·30e
Hwanghae-pukdo	38·30n	126·25e
Kaesŏng-chigu	38·00n	126·38e
Kangwŏn-do	38·45n	127·35e
P'Yŏngan-namdo	39·20n	126·00e
P'yŏngan-pukdo	40·10n	125·20e
P'yŏngyang-si	39·05n	125·50e
Yanggang-do	41·15n	128·00e

Cities and towns

Anbyŏn	39·03n	127·32e
Andong-ni	39·28n	127·27e
Anju	39·36n	125·40e
Aoji	42·31n	130·23e
Chaeryŏng	38·24n	125·36e
Ch'angdo	38·30n	127·40e
Changhŭng-ni	40·24n	128·19e
Changjin	40·23n	127·15e
Changsŏng-ni	40·58n	127·04e
Changyŏn	38·37n	125·16e
Chasŏng	41·27n	126·37e
Chigyŏng	39·51n	127·26e
Ch'ŏlsan	39·46n	124·40e
Ch'ŏngdan	37·58n	125·56e
Ch'ŏngjin	41·47n	129·50e
Ch'ŏngju	39·41n	125·13e
Chongsŏng	42·46n	129·48e
Ch'ŏnma	40·03n	125·01e
Ch'ŏnsu-ri	41·43n	128·45e
Ch'osan	40·50n	125·47e
Chosanch'am	40·22n	126·11e
Ch'owŏn-ni	39·40n	127·17e
Chunggang	41·45n	126·52e
Chunggang-ni	40·52n	127·20e
Chunghwa	38·52n	125·47e
Chungsam-ni	38·34n	127·09e
Chŭngsan	39·06n	125·22e
Chuŭl	41·33n	129·34e
Ch'wiya-ri	38·03n	125·32e
Haeju	38·02n	125·42e
Haengyŏng	42·32n	129·56e
Hahyŏn-ni	38·33n	127·57e
Hamhŭng	39·54n	127·32e
Hamjong	38·59n	125·20e
Happ'o-ri	40·40n	127·49e
Hapsu	41·13n	128·51e
Hoeryŏng	42·27n	129·44e
Hoeyang	38·43n	127·36e
Honggun	40·46n	128·27e
Hongwŏn	40·02n	127·57e
Huch'ang	41·25n	127·03e
Hŭich'on	40·10n	126·17e
Hŭngnam	39·50n	127·38e
Hun-yung	42·53n	130·12e
Hŭpkok	39·04n	127·44e
Hwanggong-ni	40·03n	129·27e
Hwangju	38·42n	125·46e
Hyesan	41·23n	128·12e
Ich'on	38·30n	126·50e
Ijin	42·05n	130·08e
Imokt'an	38·50n	126·41e
Inp'ung-dong	41·25n	126·34e
Iwŏn	40·19n	128·39e
Kach'ang-ni	38·24n	126·11e
Kaech'ŏn	39·42n	125·53e
Kaesŏng	38·00n	126·30e
Kamsu-ri	38·03n	125·54e
Kangdong	39·09n	126·05e
Kanggup'o	41·07n	127·31e
Kanggye	40·58n	126·34e
Kangsŏ	38·58n	125·26e
Kapsan	41·04n	128·19e
Kasan-dong	41·18n	126·55e
Kilchu	40·58n	129·20e
Kimch'aek	40·41n	129·12e
Koch'am-ni	41·06n	129·23e
Koin-ni	40·28n	126·22e
Kojŏ	38·57n	127·51e
Kokku	40·22n	128·44e
Koksan	38·46n	126·40e
Komusan	42·08n	129·41e
Kosan	38·52n	127·24e
Kosŏng	38·40n	129·19e
Koŭp	39·36n	124·28e
Kowŏn	39·26n	127·14e
Kuch'ang-ni	40·09n	124·46e
Kujang	39·52n	126·01e
Kujŏng-ni	37·53n	125·54e
Kŭmch'ŏn	38·10n	126·29e
Kŭmsan-ni	37·55n	125·41e
Kŭmsŏng	38·25n	127·36e
Kusŏng	39·59n	125·15e
Kyŏngsŏng	41·35n	129·36e
Kyŏngwŏn	42·48n	130·09e
Maengsan	39·40n	126·30e
Majŏn-ni	39·06n	127·07e
Manp'o	41·09n	126·17e
Monggŭmp'o	38·09n	124·47e
Muksi-ri	39·52n	125·54e
Munam-ni	38·41n	126·54e
Munch'ŏn	39·16n	127·15e
Musan	42·14n	129·13e
Najin	42·15n	130·18e
Namch'ŏn	38·19n	126·23e
Namho-ri	38·07n	125·10e
Namp'o	38·45n	125·23e
Namsi	39·54n	124·36e
Namyang	42·57n	129·53e
Nanam	41·43n	129·41e
Nogangjin	39·30n	125·23e
Nuch'ŏn-ni	38·14n	126·16e
Ŏdaejin	41·34n	129·40e
Ogu-dong	38·57n	126·56e
Okkang-ni	40·18n	124·42e
Oktong-ni	38·27n	127·07e
Ŏmyŏnbo	41·16n	127·36e
Onch'ŏn-dong	40·51n	129·07e
Ongjin	37·57n	125·21e
Onsŏng	42·57n	129·52e
Oro	40·02n	127·26e
Osich'ŏn-ni	41·25n	128·16e
Pakch'ŏn	39·45n	125·35e
P'anmunjŏm	37·57n	126·40e
P'anp'yŏng-ni	40·28n	125·49e
P'ihyŏn	40·01n	124·37e
Ponghyŏn	37·49n	125·36e
Pŏptong	38·59n	127·05e
Pugŏ-ri	42·01n	129·59e
Pukch'ang	39·36n	126·17e
Pukchin	40·10n	125·43e
Pukch'ŏng	40·15n	128·20e
P'ungsan	38·28n	125·01e
P'ungsong-ni	39·56n	127·11e
Pusŏng-ni	40·19n	127·19e
Puyun-dong	41·55n	129·30e
Pyŏktong	40·35n	125·20e
Pyŏlch'ang-ni	39·17n	126·26e
P'yŏnggang	38·26n	127·16e
P'yŏngsan	38·19n	126·23e
P'yŏngyang	39·01n	125·45e
Pyŏrha-ri	40·48n	126·32e
Sain-ni	39·15n	125·51e
Sakchu	40·23n	125·01e
Samho	39·56n	127·53e
Samsu	41·19n	127·59e
Sangnyŏng-ni	38·14n	126·54e
Sap'o-ri	40·49n	129·31e
Sariwŏn	38·31n	125·44e
Sehyŏn-ni	38·20n	127·41e
Sep'o	38·39n	127·22e
Sibyŏn	38·19n	126·41e
Simp'o-ri	38·36n	127·41e
Sinanju	39·36n	125·36e
Sinbokchang	41·01n	128·54e
Sinch'ang	40·07n	128·28e
Sinch'ang	40·19n	125·27e
Sinch'ŏn	38·22n	125·28e
Singye	38·36n	126·30e
Sinhŭng	40·11n	127·34e
Sinmak	38·25n	126·14e
Sinp'a	41·24n	127·46e
Sinp'o	40·03n	128·12e
Sinsang	39·38n	127·25e
Sinsi-ri	39·59n	124·58e
Sinŭiju	40·05n	124·24e
Sinŭp	39·54n	126·47e
Sinwŏn	38·14n	125·47e
Soch'i-dong	38·06n	124·45e
Sŏhŭng	38·27n	126·10e
Sŏkp'o-ri	37·46n	125·27e
Sŏnch'ŏn	39·48n	124·55e
Songhwa	38·21n	125·08e
Sŏngjang-ni	41·02n	126·50e
Sonjin, see Kimch'aek		
Sŏngnae-ri	39·28n	126·59e
Sŏng-ni	39·38n	127·06e
Songnim	38·44n	125·38e
Sŏsura	42·16n	130·37e
Suan	38·42n	126·22e
Sukch'ŏn	39·24n	125·38e
Sunan	39·13n	125·41e
Sunch'ŏn	39·26n	125·54e
Sup'ung	40·27n	124·57e
Suwon-dong	41·54n	129·43e
Taedong	39·05n	125·32e
Taegwan	40·13n	125·12e
Taehŭng	40·06n	126·56e
Taejujŏm	38·24n	127·58e
T'aet'an	38·04n	125·18e
T'agyŏng-ni	38·04n	126·03e
Tanch'ŏn	40·27n	128·54e
Tasado	39·50n	124·27e
T'oejo-dong	39·54n	127·46e
Tohyŏn-ni	39·53n	124·52e
Tŏkch'ŏn	39·46n	126·19e
Tŏkhŭng-ni	40·02n	127·08e
Tŏksil-ri	40·44n	127·12e
T'ongchŏn	38·54n	127·54e
Tongyang-ni	39·08n	126·52e
T'osan	38·18n	126·43e
Tuji-ri	41·31n	127·12e
Ŭiju	40·12n	124·32e
Ŭnch'ŏn	38·34n	125·26e
Unggi	42·20n	130·24e
Unp'a	38·26n	125·45e
Ŭnsan	39·25n	126·01e
Ŭnyul	38·30n	125·10e
Ŭpchŏ-ri	37·53n	125·09e
Wŏnsan	39·09n	127·25e
Yangsi	39·59n	124·28e
Yangp'yŏng-ni	40·53n	127·58e
Yangdŏk	39·13n	126·39e
Yŏnan	37·55n	126·10e
Yongamp'o	39·55n	124·24e
Yŏngan	41·15n	129·30e
Yŏngbyŏn	39·49n	125·48e
Yongch'ŏn-dong	41·18n	129·40e
Yonggang	38·52n	125·21e
Yonggang-Onch'ŏn	38·53n	125·14e
Yŏnghŭng	39·32n	127·13e
Yŏngmi	39·39n	125·31e
Yŏngwŏn	39·49n	126·32e
Yongwon-ni	40·41n	128·42e
Yongyŏn-ni	40·44n	126·09e
Yŏngyu	39·18n	125·34e
Yop'o-ri	38·50n	125·33e
Yul-li	38·52n	126·13e
Yuwŏnjin	40·18n	126·37e
Yuyŏn-ni	38·42n	127·10e

Physical features and points of interest

Amnok-kang, see Yalu		
Chaeryong-gang, *river*	38·39n	125·28e
Changjin-gang, *river*	41·24n	127·45e
Changsan-got, *cape*	38·08n	124·39e
Ch'o-do, *island*	38·32n	124·50e
Ch'ŏngch'ŏn-gang, *river*	39·35n	125·28e
Hamgyŏng-sanmaek	41·50n	128·30e
Imjin-gang, *river*	37·47n	126·40e
Japan, Sea of	39·30n	129·30e
Ka-do, *island*	39·33n	124·40e
Kaema-kowŏn, *plateau*	40·40n	127·00e
Kangnam-sanmaek, *mountains*	40·30n	125·30e
Korea Bay	39·15n	125·00e
Kŭmgang-san, *mountain*	38·35n	128·10e
Kwanmo-bong, *mountain*	41·42n	129·13e
Mayang-do, *island*	40·00n	128·12e
Munam-san, *mountain*	40·22n	127·08e
Musu-dan, *point*	40·50n	129·43e
Myohyang-san, *mountain*	40·02n	126·17e
Myohyang-sanmaek, *mountains*	40·30n	127·00e
Namdae-ch'ŏn, *river*	40·46n	129·22e
Nam-gang, *river*	39·03n	125·52e
Namp'ot'ae-san, *mountain*	41·44n	128·24e
Nangnim-sanmaek, *mountains*	40·30n	127·00e
Paektu-san, *mountain*	42·00n	128·03e
P'albong-san, *mountain*	40·16n	127·57e
Pisam-bong, *mountain*	40·41n	126·34e
Pukhan-gang, *river*	38·20n	127·48e
Puksubaek-san, *mountain*	40·42n	127·44e
Puktae-ch'ŏn, *river*	40·28n	129·00e
Sin-do, *island*	39·48n	124·14e
Sinmi-do, *island*	39·33n	124·53e
Sŏjosŏn-man, *bay*	39·20n	124·50e
Sŏk-to, *island*	38·40n	125·00e
Sŏng-ch'ŏn-gang, *river*	39·48n	127·35e
Sunwi-do, *island*	37·44n	125·15e
Sup'ung-chŏsuji, *reservoir*	40·30n	125·05e
T'aebaek-sanmaek, *mountains*	38·35n	127·45e
Taedong-gang, *river*	38·42n	125·15e
Tasa-do, *island*	39·47n	124·25e
Tongjosŏn-man, *bay*	39·30n	128·00e
Tuman-gang, *river*	42·18n	130·41e
Turyu-san, *mountain*	41·10n	128·47e
Yalu (Amnok-kang), *river*	39·55n	124·22e
Yellow Sea	37·30n	125·00e
Yesŏng-gang, *river*	37·53n	126·24e
Yŏ-do, *island*	39·14n	127·38e
Yŏnghŭng-man, *bay*	39·15n	127·30e
Yŏnhwa-san, *mountain*	40·46n	127·23e



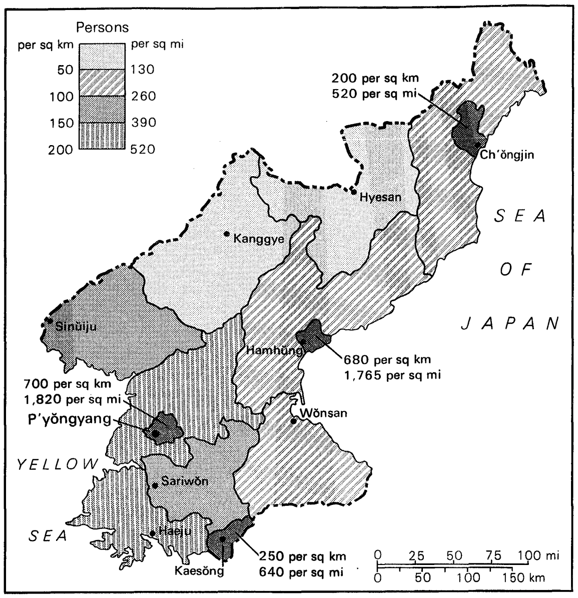
NORTH KOREA

Size of symbol indicates relative size of town ∘ ○ ◎ ▣ ▣ ■

Elevations in metres

0 25 50 100 150 km

0 25 50 100 mi



Population density of North Korea.

Forestry has declined since World War II. The sea is the main source of protein for North Koreans, and the government has continually expanded commercial fishing since the 1950s. Most fishing activity centres on the east coast and on Tasa-do (Tasa Island) in the mouth of the Yalu River. The annual marine catch includes anchovy, mackerel, pollack, tuna, crustaceans, and seaweed.

Mining is a state enterprise under the direction of a manager who is appointed by the government. The most emphasis is given to the extraction of coal and iron ore. There are more than 100 small coal mines, an iron mine at Sunhŭng, a nickel mine at Puyun, and a copper mine at Mandŏk.

Industry. Since World War II, North Korea has changed from an agricultural to an industrial nation. The three most important industries are concerned with the production of iron and steel, centred at Songnim and Ch'ŏngjin; of industrial and agricultural machinery at Kangsŏn, near P'yŏngyang; and of textiles, centred at P'yŏngyang and Hamhŭng. Other industrial products include chemicals, armaments, vehicles, cement, glass, ceramics, and some consumer goods.

Industrial development is related to the country's large supply of electric power. Production of electricity has increased steadily, although it has not kept pace with industry, and electricity is sold to China. Most power is provided by hydroelectric facilities, such as those at Sup'ung, Puryŏng, Kŭm'gang, and Kanggye. There has been expansion of thermal facilities through aid from the Soviet Union and China.

Finance and trade. The North Korean Central Bank is the sole bank of issue. It receives all national revenues and precious metals and provides government agencies with working capital. The Industrial Bank has branches in every farm cooperative; it administers the government insurance system, operates savings accounts for government and individuals, and is the only bank to grant loans. The Foreign Trade Bank handles all foreign transactions and currencies and is supervised by the Central Bank.

More than 80 percent of foreign trade is conducted with the Soviet Union and China. Since the 1960s, trade has been permitted with non-Communist countries, including Japan, France, Australia, West Germany. Hong Kong, The Netherlands, and the United Kingdom. Imports mainly consist of machinery (including machine tools and precision instruments), fuel and related oil, chemical, or rubber products; exports are pig iron, magnesia products, iron ore, and nonmetallic minerals.

Administration of the economy. The farm cooperative is not only an economic unit but also the basic unit of technical, ideological, and political control. A Cooperative Farm Management Committee was established in 1961 in each county (*kun*) in order to tighten control over the individual units. Since grain is marketed only through state-operated stores, the government controls production, pricing, and distribution of grain. The industrial sector is organized into state-owned enterprises and production cooperatives, the latter being confined largely to handicrafts, marine processing, and other small-scale operations.

Through the three national economic plans promulgated since 1954, the government has given high priority to manufactures—especially the chemical and heavy industries—at the expense of agriculture. In order to increase the low productivity of labour, the state adopted an independent accounting system and a mass-mobilization measure that is known as the Ch'ŏllima ("Flying Horse") campaign.

The nation's railways

Transportation. Railroads are the principal means of transportation. The basic railway pattern runs in a north–south direction roughly parallel to the coasts with branch lines to the river valleys. Because of the high mountains, there is only one east–west railway line between P'yŏngyang and Wŏnsan. The Kyŏng-Ui line on the west coast runs from Kaesŏng near the South Korean border to Sinŭiju on the Chinese border, connecting the major cities. From this major line a branch from P'yŏngyang southwestward to Namp'o connects centres of machine building and foundries. The Manp'o line, which runs northward from P'yŏngyang to Manp'o on the Yalu River, connects the western interior to China's Northeast (formerly Manchuria). The Wŏlla line is the major railway on the east coast; it runs from Wŏnsan northward to Najin and continues to Namyang on the Chinese border. Several branch lines serve the inland areas and mining centres.

Highway transportation is not as important as railroads because few motor vehicles are available. Major roads parallel the rail lines, and there are few east–west roads. Most roads are not paved.

River transportation plays an important role in transporting agricultural products and minerals. The most important rivers utilized for freight transportation are the Yalu, Taedong, and Chaeryŏng. Namp'o—the entry port to P'yŏngyang—Haeju, and Tasado are the major ports on the west coast, as are Wŏnsan, Ch'ŏngjin, and Najin in the east.

Air services are controlled by the air force. Flights are maintained between the major cities, and international services connect P'yŏngyang with Peking and Moscow. Sunan Airport, 10 miles north of P'yŏngyang, serves as the international airport; domestic airports are located at Hamhŭng, Ch'ŏngjin, and Wŏnsan.

Administrative and social conditions. *Government.* Constitutionally, the 45-member Cabinet, Supreme Court, and Supreme Procurator (an agency that maintains surveillance over all citizens) are responsible to the Supreme People's Assembly, which is the highest organ of state power. The actual source of authority, however, derives from the extraconstitutional political body of the Korean Workers Party. The government is highly centralized and totalitarian in nature and is often officially described as a transmission body of the party.

There are nine provinces (*do* or *to*); three special cities (*si*) of P'yŏngyang, Ch'ŏngjin, and Hamhŭng; and one special region (*chigu*) of Kaesŏng. These are further subdivided into cities, counties, and *ri,* the smallest administrative unit. Local people's assemblies elect the members of their people's committees, which execute administration duties and make local economic plans and budgets with the approval of higher authorities.

Control of political activity

There are a number of political parties and social organizations that serve to support the Korean Workers Party. All political activities are sponsored by the party or require its sanction and must closely follow the party line and policies. Elections provide a means whereby assent is registered for the policy and program of the party; they do not allow freedom of expression. There is seldom more than one candidate on the ballot for each constituency, and the electoral system is completely controlled by the party.

The judicial system consists of the courts and the procuracy. They are independent of each other, and actually the procurator's office functions as the fourth branch of government. The courts consist of the Supreme Court, whose judges are elected for three-year terms by the Supreme People's Assembly, and the provincial and people's courts, whose members are elected by local people's assemblies. Judges are usually party members or are controlled by the party.

Since 1966 there has been an emphasis on military preparedness, and economic plans have been altered to support military expenditures. The army is the largest force; there is also an air force and a navy. Both men and women are subject to conscription for three to four years service. There is also a paramilitary militia.

Education. Education is directly controlled by the party and serves as a process of indoctrination in Communist ideology and a means to supply skilled workers, technicians, and scientists to meet the government's economic goals. All students are required to engage in productive labour along with their studies, which emphasize science and technology. In 1967 education was made compulsory for those between the ages of seven and 16, later changed to between five and 16. The system comprises one year of preschool, four years of primary, and six years of secondary school. Institutions of higher education offer programs of two to six years in length; the most important school is Kim Il-sung University in P'yŏngyang. There is also a well-developed system of adult education, the major components of which are technical schools located in large industrial centres.

Health and welfare. Medical care is free, and there is at least one clinic in each village, but there is a shortage of physicians and medicine. Medical benefits are provided by social insurance for workers who are temporarily or permanently disabled and women during pregnancy and childbirth. There are also funeral benefits and old-age pensions. Homes for the aged in each province have been operative under the Ministry of Labour since 1964.

Reconstruction of houses after the Korean War was given high priority, and dwellings have improved considerably. Rural mud-walled, thatched-roofed huts have been replaced by brick buildings with tile or slate roofs. Urban housing is classified into five groups that range from one room and a half-sized kitchen to individual houses with gardens.

The national police and secret agencies under the Ministry of Social Security control people's movements and social activities even down to the household level. Because of the high priority for industrialization and defense, the provision of consumer goods and social services has long been inadequate. The material economy and the lot of the peasant have improved since World War II, however, with adequate supplies of such basic goods as food and clothing generally being available.

Cultural life. The compound religious thoughts of shamanism, Buddhism, and Confucianism have deep roots in Korean culture. Although the country has received continuous streams of foreign cultural influence mainly from China, Koreans have kept their identity and maintained and developed their unique language and customs. Westernization, begun in the late 19th century, was in harmony with Korean tradition and slowly transformed the culture without much conflict until the 1940s. After World War II, the occupying Soviets did not recognize the Korean traditional family system or Confucian philosophy; age-old lineage records were burned, and the kinship system was broken. Through education, people were molded to fit the pattern of party idealism, and private life and individual freedom became extremely limited.

Development plans since the Korean War have demanded almost superhuman patience and labour from the North Koreans. As a result, the people have had to lead an austere existence. The standard of living has improved, but leisure and cultural activities have continued to be regimented and geared toward organized group activities, such as rallies and museum tours. The government strongly believes in nationalism and is concerned with the maintenance and advancement of the traditional fine arts and other cultural features. The selection of cultural items is based on Communist ideology, and writers and artists attempt to enhance class consciousness and propagate the superiority and independence of Korean culture.

The state of the arts

All of the writers, artists, dancers, and musicians are assigned to government institutions, such as the National Theatre for the Arts, National Orchestra, and National Dancing Theatre in P'yŏngyang and to provincial organizations of music, ballet, and drama. Museums have been well sponsored by the government, and many archaeological sites have been excavated to promote the growth of a strong nationalistic feeling. There are more than a dozen museums, including the Korean Revolutionary Museum and the Korean Fine Arts Museum in the capital. Archaeological sites are located in Nangnang district of P'yŏngyang and at Kungsan, near Yonggang.

Of the daily newspapers, the Korean Workers Party Central Committee's *Nodong sinmun,* the government's *Minju Chosŏn,* and the General Federation of Korean Trade Union's *Nodongja sinmun* have the largest circulations. The Korean Central News Agency controls the dissemination of information, and all papers are strictly censored. The government considers radio and television to be important mass media, and they play a great role in ideological education. Radio broadcasts reach all parts of the country; most people in the rural areas listen to wire receivers given to every village by the government. Almost all North Korean households have access to radio broadcasts as a result of a government project to link household loudspeakers to village receivers. Television broadcasting in North Korea has been made available to all parts of the country, and the number of television sets, both imported and domestically produced, has increased annually.

For statistical data on the land and people of North Korea, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (C.Le.)

HISTORY

Unlike the U.S. forces in the South, the Soviet Army marched into the North in 1945 accompanied by an army of Korean Communists. By placing the latter in key positions of power, the Soviet Union easily set up a Communist-controlled government in the North. On August 25 the People's Executive Committee of South Hamgyŏng Province was created by the South Hamgyŏng Province Communist Council and other nationalists. The Soviet authorities recognized the committee's administrative power in the province, thus setting a pattern for the committee's role throughout the North Korean provinces. In this way the Soviet Union placed the North under its control without actually establishing a military government. In October the people organized the Bureau of Five Provinces Administration, a central governing body; but this was replaced in February 1946 by a Provisional People's Committee for North Korea. This new agency adopted the political structure of the Soviet Union.

Kim Il-sung, who arrived in P'yŏngyang in the uniform of a major of the Red Army, was introduced to the people as a national hero on October 14, 1945. Shortly after his public appearance, Kim was elected as the first secretary of the North Korean Central Bureau of the Communist Party. After the Provisional People's Committee for North Korea was organized, with Kim as its chairman, it took over the existing central administrative bureaus. A year later, in February 1947, a legislative body was established under the name of the Supreme People's Assembly, and with the strong support of the Soviet occupation authorities Kim began to consolidate his political power.

Rise of Kim Il-sung

In 1948, when the Democratic People's Republic of Korea was established, Kim became the first premier of the North Korean Communist regime, and in 1949 he became chairman of the Korean Workers Party (KWP), which had been created on August 23, 1946.

After 1956, as the Sino-Soviet conflict intensified, Kim was compelled to shift his positions vis-à-vis Moscow and Peking no fewer than three times: from pro-Soviet to neutral, from neutral to pro-Chinese, and from pro-Chinese to independent. The division between the two great powers was also reflected in conflicts within the North Korean leadership. The pro-Chinese group, known as the Yenan faction, was purged by Kim during 1956–58. At about the same time, he eliminated a pro-Soviet faction from the Central Committee of the Korean Workers Party.

In 1966, after a visit to P'yŏngyang by the Soviet prime minister, Aleksey N. Kosygin, Kim announced what became known as the independent party line in North Korea, in which he stressed the principles of "complete equality, sovereignty, mutual respect, and noninterference among the Communist and Workers' Parties." From this statement, the KWP ideologists worked out four principles: "*juche* [or *chuch'e;* 'autonomy,' or 'identity'] in ideology," "independence in politics," "self-sustenance in economy," and "self-defense in national defense."

In the late 1960s the Kim regime carried out a program of strengthening the military forces, turning the country into a fortress with a large standing army and much modern equipment. There was also a strong militia.

North Korea's emphasis on strengthening its military forces proceeded in parallel with its continued stress on the construction of a self-reliant economy. With aid from the Soviet Union, the People's Republic of China, and the countries of eastern Europe, North Korea implemented a series of economic development plans and made significant gains. When the Soviet Union, under Nikita S. Khrushchev, suspended its aid to North Korea, the Seven-Year Plan (1961–67) of the North Korean regime was seriously affected, as indicated by the extension of the plan for another three years. (B.-h.H.)

Two subsequent plans, a Six-Year Plan (1971–76, extended to 1977) and a Seven-Year Plan (1978–84), also failed to achieve their stated goals. Significant economic gains were made during those periods, but growth was hampered by the country's heavy expenditures on defense.

In 1980 the KWP held its first party congress in a decade. During the proceedings, Kim's son, Kim Chong Il, was named to three powerful party posts, thus making him the second most influential person in the government. During the 1980s the younger Kim consolidated his power and gradually assumed greater control of the day-to-day administration of the government.

The North Korean government continued to maintain its balanced diplomatic position between China and the Soviet Union, alternately favouring one and then the other. The country remained, however, one of the most isolated in the international community, a position that was reinforced by two incidents of terrorist violence against South Korea for which the North Korean government was widely held responsible: a bombing in Rangoon (October 1983) that killed several members of the South Korean government; and the destruction of a South Korean airliner (November 1987), presumably by a bomb, over the Thai–Burmese border. (Ed.)

For later developments in the history of North Korea, see the *Britannica Book of the Year* section of the BRITANNICA WORLD DATA ANNUAL.

BIBLIOGRAPHY. PYONG-DO YI, *Hankuksa Taegwan* (1948), a compendium of Korean history, revised several times; TAKASHI HATADA, *A History of Korea* (Eng. trans. 1969), a socio-economic study; KI-BAIK LEE, *Hankuksa Sillon* (1967), containing a unique periodization based on changes of ruling powers; WOO-KEUN HAN, *The History of Korea* (Eng. trans. 1971), an outline with emphasis on the modern period; CHIN-TAE SON, *Hankuk Minjoksa Kaeron* (1948), a social history covering the period up to the fall of the Silla dynasty; CHAE-WON KIM and PYONG-DO YI, *Hankuksa Kodaepyon* (1959), a history of the ancient period; PYONG-DO YI, *Hankuksa Chungsepyon* (1961), a history of medieval Korea; WON-YONG KIM, *Hankuk Kogohak Kaeron* (1966), valuable as an introductory work on Korean archaeology; and KI-JUN CHO, *Hankuk Kyongjesa* (1962), an economic history of Korea.

SANG-BAEK YI and SUN-GUN YI, *Hankuksa,* 4 vol. (1961–65), is a history of the Yi dynasty, with primary focus on political changes; KWAN-U CHON, "Hankuk Silhak Sasangsa," in *Hankuk Munhwasa Daege,* vol. 6 (1970), deals with the so-called practical thought that was highly developed in 18th-century Korea and with the nature of the several schools of thought and their developments; SANG-KI KIM, *Tonghak Kwa Tonghaknan* (1947), deals with Tonghak, or Eastern learning, developed in the early 1860s in reaction to Christian influence; HONG-YOL YU, *Hankuk Chonju-Kyohoesa* (1962), is a history of the Catholic Church in Korea; see also L. GEORGE PAIK, *The History of Protestant Missions in Korea, 1832–1910* (1929, reprinted 1970); HOMER B. HULBERT, *The Passing of Korea* (1906, reprinted 1969), gives a good picture of how Korea was deprived of its sovereignty toward the end of the 19th century amid the conflict of big powers in Korea; and ANDREW J. GRAD, *Modern Korea* (1944; reprinted 1979), critiques Japanese imperial rule in Korea.

CARL BERGER, *The Korea Knot: A Military-Political History* (1964), a concise but excellent study of the military and political history of Korea since the end of World War II; SOON SUNG CHO, *Korea in World Politics, 1940–1950: An Evaluation of American Responsibility* (1967), an analysis of events in Korea during the turbulent decade from World War II to the start of the Korean War—a significant contribution to the understanding of the twist and turn of events in postliberation Korean politics; GREGORY HENDERSON, *Korea: The Politics of the Vortex* (1968), containing rich historical data on both traditional and contemporary Korean society and an interesting comparison of China, Korea, and Japan, written by a former U.S. foreign service officer who spent a lengthy period in Korea; CHONG-SIK LEE, *The Politics of Korean Nationalism* (1963), a comprehensive study of Korean nationalism and the nationalist movement during the Japanese colonial rule that also treats the historical background of the Sino-Japanese War, the repercussion of the war on the Korean government, and the process by which the country lost its independence; GLENN D. PAIGE, *The Korean Decision, June 24–30, 1950* (1968), a detailed account of the United States decision to intervene in Korea in 1950, a significant contribution to the literature in foreign policy-making theories, and *The Korean People's Democratic Republic* (1966), an excellent analysis of the history of the North Korean Communist system, describing changes in economic, political, and ideological aspects; DAVID REES, *Korea: The Limited War* (1964), one of the best books dealing with the outbreak of the war; HAHN-BEEN LEE, *Korea: Time, Change and Administration* (1968), an imaginative and skillful application of time perception to administrative behaviour under conditions of rapid social change in Korea throughout the postliberation period; and SUNG-JOO HAN, *The Failure of Democracy in South Korea* (1974), a study of the causes of the collapse of Chang Myŏn's liberal democratic government in May 1961.

SHANNON MCCUNE, *Korea's Heritage: A Regional and Social Geography* (1956), and *Korea: Land of Broken Calm* (1966), provide a general description of Korea's geography, people, and culture; FREDERICA M. BUNGE (ed.), *South Korea, a Country Study,* 3rd ed. (1982), is a good source of general information on social, political, economic, and national security matters; and TAE-HUNG HA, *Guide to Korean Culture* (1968), is a survey of the varied phases of Korean culture. See also the *Korea Annual;* and REPUBLIC OF KOREA, NATIONAL BUREAU OF STATISTICS, ECONOMIC PLANNING BOARD, *Korea Statistical Handbook* (annual), *Korea Statistical Yearbook* (annual), *Monthly Statistics of Korea,* and *Social Indicators in Korea 1981,* which include current facts and statistics on government, foreign relations, the economy, social affairs, education, and culture. REPUBLIC OF KOREA, OVERSEAS INFORMATION, MINISTRY OF CULTURE AND INFORMATION, *A Handbook of Korea,* 4th ed. (1982), includes a detailed discussion of and extensive bibliography on the land, race, history, culture, arts, customs, government, foreign policy, and social developments.

FREDERICA M. BUNGE (ed.), *North Korea, a Country Study,* 3rd ed. (1981); and TAI-SUNG AN, *North Korea: A Political Handbook* (1983), are comprehensive and objective studies of all aspects of the country; ROBERT A. SCALAPINO (ed.), *North Korea Today* (1963), is a comprehensive collection of essays on various subjects; and RYU HUN, *Study of North Korea* (1966), is a detailed study that is based on North Korean materials and sources.

(K.-b.L./K.-r.L./B.-h.H./C.Le./Ed.)

Korean Literature

Although Korea has had its own language for several thousand years, it has had a writing system only since the mid-15th century, when the Korean alphabet, Han'gŭl, was invented. As a result, early literary activity was in Chinese characters. Korean scholars were writing poetry in the traditional manner of classical Chinese at least by the 4th century AD. A national academy was established shortly after the founding of the Unified Silla dynasty (668–935); and, from the institution of civil-service examinations in the mid-10th century until their abolition in 1894, every educated Korean had read the Confucian Classics and Chinese histories and literature. The Korean upper classes were therefore bilingual in a special sense: they spoke Korean but wrote in Chinese.

By the 7th century a system, called *idu,* had been devised that allowed Koreans to make rough transliterations of Chinese texts. Eventually, certain Chinese characters were used for their phonetic value to represent Korean particles of speech and inflectional endings. A more extended system of transcription, called *hyangch'al,* followed shortly thereafter, in which entire sentences in Korean could be written in Chinese. In another system, *kugyŏl,* abridged versions of Chinese characters were used to denote grammatical elements and were inserted into texts during transcription. Extant literary works indicate, however, that before the 20th century much of Korean literature was written in Chinese rather than in Korean, even after the invention of Han'gŭl.

In general, then, literature written in Korea falls into three categories: works written in the early transcription systems, those written in Han'gŭl (Hankul), and those written in classical Chinese.

This article is divided into the following sections:

TRADITIONAL FORMS AND GENRES

Poetry. There are four major traditional poetic forms: *hyangga* ("native songs"); *pyŏlgok* ("special songs"), or *changga* ("long poems"); *sijo* ("current melodies"); and *kasa* ("verses"). Other poetic forms that flourished briefly include the *kyŏnggi*-style, in the 14th and 15th centuries, and the *akchang* ("words for songs") in the 15th century. The most representative *akchang* is *Yongbi ŏch'ŏn ka* (1445–47; *Songs of Flying Dragons*), a cycle compiled in praise of the founding of the Yi dynasty. Korean poetry originally was meant to be sung, and its forms and styles reflect its melodic origins. The basis of its prosody is a line of alternating groups of three or four syllables, which is probably the most natural rhythm to the language.

The oldest poetic form is the *hyangga,* poems transcribed in the *hyangch'al* system, dating from the middle period of the Unified Silla dynasty to the early period of the Koryŏ dynasty (935–1392). The poems were written in four, eight, or 10 lines; the 10-line form—comprising two four-line stanzas and a concluding two-line stanza—was the most popular. The poets were either Buddhist monks or members of the Hwarangdo, a school in which chivalrous youth were trained in civil and military virtues in preparation for state service. Seventeen of the 25 extant *hyangga* are Buddhist in inspiration and content.

The *pyŏlgok,* or *changga,* flourished during the middle and late Koryŏ dynasty. It is characterized by a refrain either in the middle or at the end of each stanza. The refrain establishes a mood or tone that carries the melody and spirit of the poem or links a poem composed of discrete parts with differing contents. The theme of most of these anonymous poems is love, the joys and torments of which are expressed in frank and powerful language. The poems were sung to musical accompaniments chiefly by women entertainers, known as *kisaeng.*

Sijo and *kasa*

The *sijo* is the longest enduring and most popular form of Korean poetry. Although some poems are attributed to writers of the late Koryŏ dynasty, the *sijo* is primarily a poetic form of the Yi dynasty (1392–1910). *Sijo* were still being written in the second half of the 20th century. They are three-line poems in which each line has 14 to 16 syllables, and the total number of syllables seldom exceeds 45. Each line consists of groups of four syllables. *Sijo* may deal with Confucian ethical values, but there are also many poems about nature and love. The principal writers of *sijo* in the first half of the Yi dynasty were members of the Confucian upper class (*yangban*) and *kisaeng.* In the latter part of the Yi dynasty, a longer form, called *sasŏl sijo* ("narrative *sijo*"), evolved. The writers of this form were mainly common people; hence, the subject matter included more down-to-earth topics, such as trade and corruption, as well as the traditional topic of love. In addition, *sasŏl sijo* frequently employed slang, vulgar language, and onomatopoeia.

The *kasa* developed at about the same time as the *sijo.* In its formative stage, *kasa* borrowed the form of the Chinese *tz'u* (lyric poetry) or *fu* (rhymed prose). The *kasa* tends to be much longer than other forms of Korean poetry and is usually written in balanced couplets. Either line of a couplet is divided into two groups, the first having three or four syllables and the second having four syllables. The history of the *kasa* is divided into two periods, the division being marked by the Japanese invasion of 1592–97. During the earlier period the poem was generally about 100 lines long and dealt with such subjects as female beauty, war, and seclusion. The writers were usually *yangban.* During the later period the poem tended to be longer and to concern itself with moral instruction, travel accounts, banishment, and the writer's personal misfortunes. The later writers were usually commoners.

Akchang poems

Immediately after the founding of the Yi dynasty at the end of the 14th century and the establishment of the new capital in Seoul, a small group of poetic songs called *akchang* was written to celebrate the beginning of the new dynasty. In its earliest examples the form of *akchang* was comparatively free, borrowing its style from early Chinese classical poetry. Whereas the early *akchang* are generally short, the later *Yongbi ŏch'ŏn ka* consists of 125 cantos.

Prose. Korean prose literature can be divided into narratives, fiction, and literary miscellany. Narratives include myths, legends, and folktales found in the written records. The principal sources of these narratives are the two great historical records compiled during the Koryŏ dynasty: *Samguk sagi* (1146; "Historical Record of the Three Kingdoms") and *Samguk yusa* (1285; "Memorabilia of the Three Kingdoms"). The most important myths are those concerning the Sun and the Moon, the founding of Korea by Tangun, and the lives of the ancient kings. The legends touch on place and personal names and natural phenomena. The folktales include stories about animals; ogres, goblins, and other supernatural beings; kindness rewarded and evil punished; and cleverness and stupidity. Because the compiler of the *Samguk yusa* was a Zen master, his collection includes the lives of Buddhist saints; the origin of monasteries, stupas, and bells; accounts of miracles performed by Buddhas and bodhisattvas; and other tales rich in shamanist and Buddhist elements. The compilations made in the Koryŏ period preserved the stories of prehistoric times, of the Three Kingdoms, and of the Silla dynasty and have remained the basic sources for such material. Later compilations made during the Yi dynasty served as a major source of materials for later Yi dynasty fiction.

Korean fiction can be classified in various ways. First, there is fiction written in Chinese and that written in Korean. Second, there are the short works of one volume, "medium" works of about 10 volumes, and long works of more than 10 volumes. Third, there are works of *yangban* writers and those of common writers. In respect to the last classification, however, there is also a group of fictional works in which the viewpoints of the *yangban* and the commoner are combined. Most of this fiction was based on the narratives mentioned above, the author adding incidents and characters to the original story. It is not possible to assign definite dates or authors to most of these works. The stories are generally didactic, emphasizing correct moral conduct, and almost always have happy endings. Another general characteristic is that the narratives written by *yangban* authors are set in China, whereas those written by commoners are set in Korea.

The literary miscellany consists of random jottings by the *yangban* on four broad topics: history, biography, autobiography, and poetic criticism. Like fiction, these jottings were considered to be outside of the realm of officially sanctioned Chinese prose (*e.g.,* memorials, eulogies, and records), but they provided the *yangban* with an outlet for personal expression. Thus, their portrayal of the customs, manners, and spirit of the times in which they were composed make these writings an essential part of Korean prose.

Oral literature. Oral literature includes all texts that were orally transmitted from generation to generation until the invention of Han'gŭl—ballads, legends, mask plays, puppet-show texts, and *p'ansori* ("story singing") texts.

In spite of the highly developed literary activity from early in Korean history, song lyrics were not recorded until the invention of Han'gŭl. These orally transmitted texts are categorized as ballads and are classified according to singer (male or female), subject matter (prayer, labour, leisure), and regional singing style (capital area, western, and southern). The songs of many living performers, some of whom have been designated as "intangible national treasures" by the South Korean government, are still being recorded.

Legends include all those folk stories handed down orally and not recorded in any of the written records. These legends were for long the principal form of literary entertainment enjoyed by the common people. They deal with personified animals, elaborate tricks, the participation of the gods in human affairs, and the origin of the universe.

Mask plays

The mask plays are found in Hahoe, Chinju, T'ongyŏng, Kimhae, and Tongnae in North and South Kyŏngsang provinces; Yangju in Kyŏnggi Province; Pongsan in Hwanghae Province; and Pukch'ŏng in south Hamgyŏng Province. The most representative plays are the *sandae kŭk* genre of Yangju, the *pyŏlsin kut* of Hahoe, and the *okwangdae nori* (five-actor play) of Chinju. Although the origin of these plays is uncertain, they are generally presumed to have developed from primitive communal ceremonies. Gradually, the ceremonial aspect of the plays disappeared, and their dramatic and comic possibilities were exploited. The dialogue was somewhat flexible, the actors being free to improvise and satirize as the occasion demanded. The plays were not performed on a stage, and there were no precise limits as to the space or time in which the performances took place. The audience also traditionally responded vocally to the play as well as passively watching it. The organization of the mask plays—through repetition and variety—achieves a remarkable effect of dramatic unity.

Only two puppet-show texts are extant, *Kkoktukaksi nori* (also called *Pak Ch'ŏmjikuk;* "Old Pak's Play") and *Mansŏk chung nori.* Both titles are derived from names of characters in the plays. No theory has been formulated as to the origin and development of these plays. The plots of the puppet plays, like those of the mask plays, are full of satiric social criticism. The characters—Pak Ch'ŏmji, governor of P'yŏngam, Kkoktukaksi, Buddhist monk, and Hong Tongji—dance and sing, enacting familiar tales that expose the malfeasance of the ruling classes.

The final type of folk literature is found in the texts of *p'ansori* of the Yi dynasty. These texts were first recorded in the 19th century as verse, but the written forms were later expanded into *p'ansori* fiction, widely read among the common people. This transformation from poetry to narrative fiction was easily accomplished, since *p'ansori* were always narrative. Originally the entire *p'ansori* performance repertoire consisted of 12 *madang* ("titles"). Although all 12 remain as narrative fiction, only five of them are sung today. The texts evolved gradually from the legends, which provided their sources and were altered and expanded as they were passed from one performer to another.

HISTORY

The earliest literature: before 57 BC. From the earliest times, poetry and music have played an important part in the daily life of the Korean people. This love for song and dance impressed the ancient Chinese, whose observations are found in their early records. Ancient Korean songs, closely allied to the religious life of the people, were performed at such rites as the worship of heaven in the north and the sowing and harvest festivals in the south. These songs were transmitted orally and were thought to have magical properties.

Early songs

Three songs are handed down in Chinese translation: "Kuji ka" (or "Yŏng singun ka"; "Song for Welcoming the Gods," in the *Samguk yusa*), "Hwangjo ka" (17 BC; "Song of Orioles," in the *Samguk sagi*), and "Kong mudoha ka" (or "Konghuin"; "A Medley for the Harp," in the *Haedong yŏksa*). The "Kuji ka" is related to the myth of the founding of the Karak state, but it appears to have been a prayer sung at shamanist rituals. Some have interpreted it as being a song of seduction sung by women. The "Hwangjo ka," attributed to King Yuri, seems to be a fragment of a love song. The hero of "Kong mudoha ka" is thought to have been a shaman who drowned himself while in a trance. Perhaps the poem indicates the loss of the shaman's efficacy and authority when ancient Korea was transformed into a structured state. The story also includes other characters such as the sailor, his wife, and her friend. Another song, the "Tosol ka" (AD 28), is mentioned in the *Samguk sagi* as the beginning of secular poetry, but the poem itself has not survived.

Literature of the Three Kingdoms: 57 BC–AD 668. In contrast to the literature of the earliest ages, which is characterized by collective artistic activity, that of later ages shows the effects of political, economic, and cultural changes as the peninsula increased in wealth and widened its contacts with other areas. The introduction of Buddhism and Chinese characters to the Three Kingdoms enriched their literature and changed their worldview greatly. In consequence, their artistic activity advanced far beyond collective singing and dancing to the direct expression of individual feelings. The heroes of this literature were human beings with individual personalities in contrast to the more idealized tribal heroes of earlier times.

The three kingdoms of this period were Koguryŏ, in the north; Paekche, in the southwest; and Silla, in the southeast. The writers of Koguryŏ, the geographical location of which provided close contact with the Chinese mainland, seem to have retained something of the original pioneer spirit from the times when Koreans came from the northern regions and settled on the peninsula; their poems tended to be heroic tales in epic form. The foundation myth of Koguryŏ concerns the migration of King Tongmyŏng and his people into the region. The stories of Ondal, King Mich'ŏn, Prince Hodong, the heir apparent Yuri, and others that had their origin in Koguryŏ are still used today as the bases for dramas and motion pictures.

In contrast to that of Koguryŏ, the literature of Paekche and Silla tended to be lyrical, perhaps because of the milder climate and easier life in the south. Although little literature from Paekche has survived, the legends and songs contained in the *Samguk sagi* give a hint of its original extent and richness. For example, "Chŏngŭpsa" ("Song of Chŏngŭp")—in which the wife of an itinerant merchant asks the Moon to protect her husband—was passed down from Paekche through the Koryŏ and Yi dynasties and is still appreciated in the 20th century.

Silla led the other two kingdoms both politically (as proved by its subsequent unification of Korea) and artistically, in spite of the fact that it was farthest removed from contact with Chinese culture. The geographical and cultural distance from China, however, seems to have been an advantage, since the culture of Silla was able to create a true synthesis of native and foreign elements.

Absorption of Chinese culture

Literature of Unified Silla: 668–935. After the mid-7th century Silla absorbed Koguryŏ and Paekche and created a stable political system covering most of the Korean peninsula. During the Unified Silla dynasty many students were sent at government expense to study in T'ang China. The consequent absorption of Chinese culture and the flourishing of Korean Buddhism both contributed to the remarkable artistic flowering of Silla. In particular, the spiritual life of the Silla nobility—the monks and the chivalrous Hwarangdo—was dominated by Buddhism, and Buddhism thus became the driving force behind virtually all artistic activity.

The *hyangga* was the crown of Silla's literary achievement. Although the term *hyangga* is used generally to distinguish Korean songs from Chinese poetry, it more specifically denotes the 25 extant poems transcribed in the newly devised *hyangch'al* system in the Unified Silla and early Koryŏ periods. The texts of 14 *hyangga* are preserved in the *Samguk yusa* and those of 11 devotional poems by the Buddhist monk Kyunyŏ in *Kyunyŏ chŏn* (1075; "Life of Kyunyŏ"). A large collection, *Samdaemok,* compiled by the monks Taegu and Wi Hong in 888, has not survived. The poems that remain reveal a delicate and elegant style. Two examples written in the 8th century include "Ch'an Kip'arang ka" ("Ode to the Knight Kip'a"), which praises a member of the Hwarangdo, and "Che mangmae ka" ("Song of Offerings to a Deceased Sister"), a funeral hymn.

At the same time, a great body of prose narratives was also being written in classical Chinese. These include hundreds of volumes of commentaries on Buddhist scriptures by such monks as Wŏnhyo, Ŭisang, Wŏnch'ŭk, Taehyŏn, and Kyŏnghŭng; stories of miracles performed by eminent monks, tales of the efficacy of Buddhist statues, and the origins of Buddhist monasteries; stories of valour by members of the Hwarangdo; and stories inspired by the Chinese narrative form *ch'uan-ch'i* ("tales of marvels"). The last three types of narratives in particular became the basis of classical fiction in later dynasties.

Literature of Koryŏ: 935–1392. The last master of the *hyangga* was the monk Kyunyŏ, who wrote voluminous commentaries on, and was a great popularizer of, Buddhism. He composed his poems in Korean, transmitted them orally, and encouraged his followers to chant and memorize them. The poems in his *Kyunyŏ chŏn,* based on the 10 vows of the bodhisattva Samantabhadra, were transcribed frcm this oral transmission. The new poetic form that flourished during the period was the *pyŏlgok,* which was of folk origin. The *pyŏlgok* was intended for large-scale performances on festive occasions, especially the Harvest Festival and the Lantern Festival. Many *pyŏlgok* were written and performed by women, and such poems as "Tongdong" ("Ode on the Seasons") and "Isanggok" ("Winter Night") are among the most moving love lyrics in the Korean language.

The Koryŏ dynasty was a time of social instability. Internal and external crises abounded, the result of a factious and oppressive nobility and army, constant border harassment by the Khitan and Juchen peoples, and the invasions of the Mongols. Under such conditions established scholarly writers tended to be introspective or hedonistic. Consequently, the new intellectuals who arose toward the end of the dynasty began to adopt Confucian and Taoist dualistic thought as their philosophy. They were dissatisfied with *pyŏlgok* and sought a different form of poetic expression. This was the genesis of the *sijo,* which became a popular poetic form in the Yi dynasty.

Prose narratives underwent much development during the Koryŏ period. These included myths, legends, folklore, Buddhist stories and lives of saints, and literary miscellany. One notable class of tales is that in which the hero is represented by a personified inanimate object, such as wine, paper, a cane, ice, or a coin, or by an animate object, such as bamboo or a turtle. Representative of this form is *Kongbang chŏn* ("Tale of the Square-Holed Coin"), by Im Ch'un. Another major style is heroic narrative poetry, of which the masterpiece is the "Tongmyŏng wang p'yŏn" (1193; "Lay of King Tongmyŏng"), by Yi Kyubo, written in an old, pentasyllabic style. A work in a similar vein is the *Chewang ungi* (1287; "Rhymed History of Emperors and Kings"), by Yi Sŭnghyu, written in lines of five and seven syllables. A notable example of hagiography is the *Haedong kosŭng chŏn* (1215; "Lives of Eminent Korean Monks"), by Kakhun. The first collection of essays on poetry and other current subjects written in Korea is the *P'ahan chip* (1260; "Jottings to Break Up Idleness"), by Yi Inno (or Yi Illo). In addition to poetic criticism, the random jottings of Yi Inno contain autobiographical information in diary form; biographical notes on his friends and associates, including their life-styles and literary tastes; and remarks on contemporary manners and mores. The *P'ahan chip* inaugurated a long tradition of similar works written in the late Koryŏ and Yi dynasties.

Literature of the early Chosŏn period: 1392–1598. The literature of the Yi dynasty falls naturally into two periods, with the end of the Japanese invasion (1597) serving as a dividing line. The early period is notable for its poetry; the later, for its prose. Inheriting the tradition of Silla and Koryŏ, the writers of the early Yi dynasty raised Korean literature to new heights.

Invention of Korean alphabet

The early Yi dynasty also marks the initiation of a new era in Korean literary history with the invention of Han'gŭl in 1443–44, during the reign of King Sejong. This important event finally enabled Korean writers to record works in their native language.

The extraordinary king Sejong was not only the motivating force behind the invention of Han'gŭl but also had his scholars compile *Yongbi ŏch'ŏn ka* to praise the founding of the Yi dynasty, especially the valour and virtue of his father and grandfather. He himself compiled *Wŏrin ch'ŏngang chigok* (1447; "Songs of the Moon's Reflection on a Thousand Rivers") in praise of the life of the Buddha. Both works helped test and demonstrate the practicality of Han'gŭl as a means of literary expression and were the prototype of the new *akchang* form. Scholar-officials used the form to justify the founding of the new dynasty and to praise the virtues of its founder and the beauty of the new capital. As a literature of the privileged class, the popularity of the *akchang* was always limited, and it was soon eclipsed by the most important forms of the Yi dynasty—*sijo* and *kasa.*

These forms owe their popularity to two factors. First, their style of expression was rich and natural and was widely appreciated by readers. Second, they were popular with writers because together the forms provided ideal outlets for the two sides of the Confucian temper: the brief and simple *sijo* were perfect vehicles for intense lyrical expression, whereas the longer *kasa* gave writers an opportunity to expound at greater length on the more practical aspects of Confucian thought.

The expressive content of the *sijo* ranges from the idealistic union of man and nature (often coupled with the poet's pride in his poverty) to the longing for sovereigns by subjects in exile (allegorical pieces in which an analogy is drawn between fidelity and romantic love) to the deeper exploration of human problems. Writers of *sijo* include Maeng Sasŏng, Yi Hyŏnbo, Yi Hwang, and Yi I. Representative poets of *kasa* include Chŏng Ch'ŏl and Hŏ Nansŏrhŏn.

Even after the invention of Han'gŭl, prose continued to be written in Chinese. The five stories contained in the *Kŭmo sinwha* ("New Stories from Golden Turtle Mountain") by Kim Sisŭp, for example, are in the tradition of the *ch'uan-ch'i.* Subject material includes love affairs between mortals and ghosts and dream journeys to the underworld or to the Dragon Palace. Two collections of literary miscellany, the *P'aegwan chapki* ("The Storyteller's Miscellany") by Ŏ Sukkwŏn and the *Yongjae ch'onghwa* ("Miscellany of Yongjae") by Sŏng Hyŏn, were written in Chinese and influenced the growth and development of vernacular prose in the later Yi dynasty.

Shift from poetry to prose

Literature of the later Chosŏn period: 1598–1894. The shift in emphasis from poetry to prose after the Japanese invasion represents a significant step in the evolution toward modern literature. It also reflects a basic change in the philosophical outlook of Korean society. The Yi dynasty had suffered from the rigid formalism of Confucian officials, whose doctrine was based on the principles of the 12th-century Chinese philosopher Chu Hsi. This Neo-Confucian philosophy was gradually replaced by the Sirhak, or Silhak ("Practical Learning"), school, which was based on reason and the scientific spirit of criticism. The introduction of Roman Catholicism from the West and of new scientific ideas from China also stimulated the reform measures advocated by the champions of the new school.

Practical Learning gave impetus to literary activity and awakened the self-consciousness of the common people. Poetry, which had been the monopoly of the lettered class, came to be written by the common people. Women also were admitted into the literary world as the principal audience for traditional fiction. The later active compilation of *sijo* and prose narratives reveals the awakening interest in rediscovering and reappraising the past.

Prose. The traditional vernacular fiction—commonly called *sosŏl* ("small talk")—that emerged during this period consisted of stories, romances, and fables. The 15th-century *Kŭmo sinwha,* written in Chinese, was an important precursor, but the first work of the genre was *Hong Kiltong chŏn* ("Tale of Hong Kiltong"), written in the early 17th century by the scholar Hŏ Kyun. Kim Manjung, building on this style, wrote two major works: *Kuun mong* (1687–88; "Dream of Nine Clouds"), the story of a Buddhist monk's search for Enlightenment, and *Sassi namjŏng ki* (*c.* 1689–92; "Story of Lady Sa's Journey to the South"), a satire against the institution of concubinage. The most popular stories of the 18th century were all anonymous: *Ch'unhyang chŏn* ("Story of Spring Fragrance"), *Shim Ch'ŏng chŏn* ("Story of Shim Ch'ŏng"), *Changhwa hongnyŏn chŏn* ("Tale of Rose Flower and Pink Lotus"), and *Hŭngbu chŏn* ("Story of Hŭngbu"). These stories were written in a simple and natural style, their characters being modeled on common people, and they have become deeply rooted in Korean consciousness.

Court literature

Stories set at court and written by women also flourished during this period. Memorable works of court literature include the *Hanjung nok* (1795–1805; "Record of Sorrowful Days"), the tragic story of a succession dispute written by Lady Hong, princess of Hyegyŏng Palace; *Kyech'uk ilgi* ("The Diary of Kyech'uk"), the anonymous record of Queen Inmok's confinement after the assassination of her son; and *Inhyŏn wanghu chŏn* ("Tale of Queen Inhyŏn"), an anonymous account of the rivalry between the Queen and the King's concubine. All three of these works described events that had actually taken place. Other prose works written by women in Han'gŭl include diaries, travel records, letters, and portraits. These works, written in prose that verged on lyricism, could easily be chanted and memorized by a growing female readership.

Poetry. During the later Yi dynasty there was also a great flowering of poetry by scholar-officials and commoners. The most gifted poet of the period was Yun Sŏndo. His 77 *sijo* poems, including *Ŏbu sasi sa* (1651; *The Angler's Calendar*), a cycle of 40 poems on the theme of the fisherman as sage, show his mastery of topics and techniques of the *sijo.* Gradually, the *sijo* was superseded in popularity by the *sasŏl sijo.* The growth of this new form, together with the rise of fiction, drama, genre painting, and *p'ansori,* reflects the rise of the middle class and changes in the approach to life.

Pak Inno, the master of *kasa* in the 17th century, wrote in a style that combined erudition and lyricism. He produced seven pieces between 1598 and 1636; the theme of his first two *kasa* was the Japanese invasion, during which he served in the navy. The desire to reevaluate the past and to re-create the world of literature led to changes in the *kasa,* as exemplified in the anonymous *kasa* by women and commoners. Women writers of *kasa* were mainly from the southern regions of Korea. They expressed their joys, angers, griefs, and pleasures, and discoursed on the etiquette for entertaining guests, religious rites, and the principles of being a wise mother and a good wife. *Kasa* written by the commoners were marked by the same style as those written by women and played a similar role in the literary activity of the general masses. There are two other forms of *kasa* written by the literati: travel records and accounts of life in exile. To the first belong *Iltong changyu ka* (1764; "Song of a Grand Trip to Japan"), written by Kim In'gyŏm upon his return from an official trip to Japan; and the *Yŏnhaeng ka* (1866; "Song of a Journey to Peking"), written by Hong Sunhak upon his return from an official trip to Peking. The second includes *Pukkwan kok* ("Song of the Northern Pass"), written by Song Chusŏk, who in 1675 accompanied his grandfather, Song Siyŏl, to his place of exile in the northeast; the *Manŏnsa* ("Song of Ten Thousand Words"), written by An Towŏn (or An Chohwan) during his banishment on the lonely island of Ch'uja, off the southeast coast of Korea; and the *Pukch'ŏn ka* (1853; "Song of a Northern Exile"), written by Kim Chinhyŏng, depicting the life of exile in the northeast.

Oral literature. Another feature of the later Yi dynasty was the formation, by the common people, of *p'ansori* texts. *P'ansori* seem to have originated during the reign of Sukchong (1675–1720), when old folktales were first sung. Their style and form were fixed by the *kwangdae,* or professional singers, and a group of amateurs in Chŏlla and Ch'ungch'ŏng provinces. Six of the original 12 titles were revised by the master *p'ansori* writer Sin Chaehyo, of which five are still performed.

Mask plays and puppet shows

The representative mask play is *Sandae kŭk.* Of unknown origin, it was usually performed on a makeshift, open-air stage in 12 scenes, or acts. The masked actors followed a script that presented a story in dialogue interspersed with dances and songs. As the puppets of the *Kkoktukaksi nori* show were made of *pak* (a gourd, rhyming with the Korean name Pak), it was also called *Pak Ch'omji kŭk* ("Old Pak's Play"). Through keen satire presented in a unique and distinguished style, the contents of the masked drama and the puppet show strongly reflect the environment and the feelings of the common people of the later Yi dynasty.

Transitional literature: 1894–1910. By the time of the 1894 reforms, enough social and intellectual change had occurred to suggest the beginnings of a division between traditional and modern literature. But, just as conservatism did not favour sudden changes in the political and social structure, literature, too, faced a period of transition toward its modern transformation. Schools were established by the educational ordinance of 1895, and the organization of learned societies and "enlightenment" movements followed soon after. Vernacular publications, the *Tongnip sinmun* ("Independent") and the *Cheguk sinmun* ("Imperial Post"), along with the establishment of the Korean Language Institute and the scientific study, consolidation, and systematization of Korean grammar, also helped open the way for the modern literary movement.

The first literary forms to appear after the 1894 reforms were the *sinsosŏl* ("new novel") and the *ch'angga* ("song"). These transitional literary forms were stimulated by the adaptation of foreign literary works and the rewriting of traditional stories in the vernacular. The *ch'angga,* which evolved from hymns sung at churches and schools in the 1890s, became popular upon the publication of the "Aeguk ka" ("National Anthem"), by Yi Yongu, and "Tongsim ka" ("A Boy's Mind"), by Yi Chungwŏn, in an issue (1896) of the *Tongnip sinmun.* Songwriters still used such traditional verse forms as the *sijo* and *kasa* or a song form, the predominant pattern of which (seven and five syllables) showed the influence of popular Japanese songs (*shōka*). Most songs denounced corruption in the government and stressed independence, patriotic fervour, and modernization.

The *sinsosŏl*

Three distinctly traditional elements were inherited by the *sinsosŏl*. First was the basic moral stance of reproving vice and rewarding virtue. Owing to the prevailing atmosphere of the "enlightenment" period, advocates of modernization were cast as virtuous, while the wicked were conservative. Second, the development of the plot was governed by coincidence, and events that lacked causality were nevertheless arbitrarily connected. Finally, the dialogue and the accompanying narrative were fused into one expository structure. The pioneering aspects of the *sinsosŏl,* however, were that it was written wholly in prose, whereas a considerable part of traditional fiction had been in verse; and it tried to depict a plausible human existence with backgrounds and events that more closely resembled reality than was the case in traditional fiction, which tended to follow certain model stories with their established plot lines and stereotyped characterizations. Writers of *sinsosŏl* also tried to unify the spoken and written language. Typical writers and their works are Yi Injik, *Kwi ŭi sŏng* (1907; "A Demon's Voice"); Yi Haejo, *Chayujong* (1910; "Liberty Bell"); and Ch'oe Ch'ansik, *Ch'uwŏlsaek* (1912; "Colour of the Autumn Moon"). In their works these writers advocated modernization, a spirit of independence, contact with advanced countries, study abroad, the diffusion of science and technology, and the abolition of conventions and superstition.

Modern literature: 1910 to the present. The modern literary movement was launched by Ch'oe Namsŏn and Yi Kwangsu. In 1908 Ch'oe published the poem "Hae egeso pada ege" ("From the Sea to Children") in *Sonyŏn* ("Children"), the first literary journal aimed at producing cultural reform. Inspired by Byron's *Childe Harold's Pilgrimage,* Ch'oe celebrates, in clean masculine diction, the strength of the young people who will carry out the necessary social and literary revolution. The poem's inventions include the use of punctuation marks, stanzas of unequal length, and reference to the sea and children, hitherto little mentioned in classical poetry. Neither Ch'oe nor his contemporaries, however, could escape the bounds of traditional prosody or succeed in modernizing traditional forms of speech and allusion. In his stories, which dealt with the enlightened pioneers who championed Western science and civilization, Yi Kwangsu adopted a prose style that approximated the everyday speech of common people. Yi's reputation was established by *Mujŏng* (1917; "The Heartless"), the first modern Korean novel.

In 1919, shortly before the unsuccessful movement for independence from Japan, translations of such Western poets as Paul Verlaine, Rêmy de Gourmont, and Stéphane Mallarmé began to exert a powerful influence on Korean poetry. The indirection and suggestiveness of French Symbolist literature were introduced by Kim Ŏk, the principal translator. Against the didacticism of the age Kim set Mallarmé, and against its rhetoric and sentimentality he set Verlaine, concluding in the process that free verse was the supreme creation of the Symbolists. Kim's fascination with the Symbolist movement culminated in the publication of *Onoe ŭi mudo* (1921; "Dance of Anguish"), the first Korean collection of translations from Western poetry. The exotic and melancholy beauty of autumn and expressions of ennui and anguish appealed to poets who sought to vent their frustration and despair at the collapse of the independence movement.

Naturalism

The movement for literary naturalism was launched in the 1920s by a group of young writers who rallied around a new definition of universal reality. Yŏm Sangsŏp, the first to introduce psychological analysis and scientific documentation into his stories, defined naturalism as an expression of awakened individuality. Naturalism's purpose, Yŏm asserted, was to expose the sordid aspects of reality, especially the sorrow and disillusionment occurring as authority figures are debased and one's idols are shattered. Many works of naturalist fiction were first-person narratives in which writers presented themselves as the subjects of case studies. The disharmony between the writer and his society often induced the writer to turn to nature; the land and simple folk furnished themes and motifs for some of the better stories in the Zolaesque tradition, among them "Pul" (1925; "Fire") by Hyŏn Chingŏn and "Kamja" (1925; "Potato") by Kim Tongin.

The 1920s produced several major poets, Han Yongun published *Nim ŭi ch'immuk* (1926; "The Silence of Love"), comprising 88 meditative poems. Han sought insight into the reasons why he and his country had to endure Japanese occupation, and he found Buddhist contemplative poetry the lyric genre most congenial to this pursuit. The nature and folk poet Kim Sowŏl used simplicity, directness, and terse phrasing to good effect. Many of his poems in *Chindallaekkot* (1925; "Azaleas") were set to music.

The Mukden, or Manchurian, Incident (1931) and the Japanese invasion of China in 1937 induced the Japanese military authorities to impose wartime restrictions. The grinding poverty of the lower classes at home and abroad, especially in the Korean settlements in southern Manchuria, was the chief concern of the writers of the "new tendency" movement, which opposed the romantic and "decadent" writers of the day and later became proletarian in spirit. Writers of the class-conscious Korean Artist Proletariat Federation (KAPF), organized in 1925, asserted the importance of propaganda and regarded literature as a means to establish socialism.

Modern Korean literature attained its maturity in the 1930s through the efforts of a group of talented writers. They drew freely upon European examples to enrich their art. Translation of Western literature continued, and works by I.A. Richards, T.S. Eliot, and T.E. Hulme were introduced. This artistic and critical activity was a protest against the reduction of literature to journalism and its use as propaganda by leftist writers.

The first truly successful poet of modern Korea was Chŏng Chiyong, who was influenced by William Blake and Walt Whitman. *Paengnoktam* (1941; "White Deer Lake"), his second book of poetry, symbolically represents the progress of the spirit to lucidity and the fusion of man and nature. A poetry of resistance, voicing sorrow for the ruined nation with defiance but without violence or hatred, was produced by Yi Yuksa and Yun Tongju. In Yi's poem "Chŏlchŏng" (1939; "The Summit"), he recreates the conditions of an existence in extremity and forces the reader to contemplate his ultimate destiny. The poetry of Yun Tongju, a dispassionate witness to Korea's national humiliation, expresses sorrow in response to relentless tyranny.

Literature since World War II

Korean fiction of the 1930s took shape in the void created by the compulsory dissolution of KAPF in 1935. Barred from all involvement with social or political issues, some writers returned to nature and sex; others retreated to the labyrinth of primitive mysticism, superstition, and shamanism; still others sympathetically portrayed characters born out of their time, defeated and lonely. In the early 1940s, the Japanese suppressed all writings in Korean. Censorship, which had begun with the Japanese annexation of Korea in 1910, was intensified. Korea was liberated in August 1945, and the Republic of Korea (South Korea) was established three years later. The literary scene experienced the revival of the controversy between left and right that had raged in the late 1920s and early 1930s. There were frantic groupings and regroupings, and most of the hardcore leftist writers, such as Yi Kiyŏng and Han Sŏrya, were in North Korea by 1948.

The liberation of 1945 produced a flowering of poetry of all kinds. Some poets were determined to bear witness to the events of their age; some sought to further assimilate traditional Korean values, while others drew variously on Western traditions to enrich their work. Sŏ Chŏngju and Pak Tujin are known for their lifelong dedication and contributions to modern Korean poetry. Considered to be the most "Korean" of contemporary poets, Sŏ is credited with exploring the hidden resources of the language, from sensual ecstasy to spiritual quest, from haunting lyricism to colloquial earthiness. Pak is capable of a wide range of moods, and his language and style impart a distinctive tone to his Christian and nationalistic sentiments. Marked by sonorific intricacies and incantatory rhythms, Pak's poems are imbued with a strong historical and cultural consciousness that bears testimony to contemporary reality.

The single overwhelming reality in Korean fiction since the Korean War has been the division of the country. The 38th parallel torments the conscience of every fictional protagonist, for it is a symbol not only of Korea's trials but also of the division of mankind and of the protagonist's alienation from himself and his world. Some have attempted to capture the images of the people in lyrical prose; others have delved into the conscience of the war's lost generation or into the inaction, self-deception, and boredom of the alienated generation of the 1960s. Some have studied the defeat and disintegration of good people; others have investigated the ways in which modern society negates freedom and individuality. Outstanding among writers of the roman-fleuve is Pak Kyŏngni, the mother-in-law of the poet Kim Chiha. Pak's multivolume *T'oji* (1969; "Land") has been acclaimed for its commanding style and narrative techniques.

In the last quarter of the 20th century a host of talented writers have been perfecting the art of being themselves. The poet Hwang Tonggyu, for example, has drawn material not only from his own experiences but also from the common predicament of the Korean people, expressing what others know but do not think of saying or cannot say. The novelist Yun Hŭnggil is another example of a writer who has cultivated fiction as an instrument of understanding himself and others. In his *Changma* (1973; "The Rainy Spell"), for example, Yun says that ideological differences imposed upon the Korean people by history can be overcome if they delve into the native traditions that have given them cohesion.

Drama

The "new" drama movement, which began in 1908, saw the rise and fall of small theatre groups, such as the T'owŏrhoe, organized in 1923, and finally the Kŭk Yesul Yŏnguhoe ("Theatrical Arts Research Society"), organized in 1931. Through their experimental theatre, the members of the society staged contemporary Western plays and encouraged the writing of original plays, such as Yu Ch'ijin's *T'omak* (1933; "Clay Hut"). The paucity of first-rate playwrights and actors, the dearth of plays that satisfy dramatic possibilities, the general living standards of the audience, as well as the lack of government support have limited dramatic activity. Domestic plays and historical pieces, however, have continued to be written and staged.

BIBLIOGRAPHY

Poetry: PETER H. LEE (comp. and ed.), *Anthology of Korean Literature from Early Times to the Nineteenth Century* (1981), collects representative poetic and prose works written in Chinese and Korean and supplies commentary and criticism; his *Lives of Eminent Korean Monks* (1969) is an annotated translation of KAKHUN, *Haedong kosŭng chŏn* (1215), with an introduction. RICHARD RUTT (ed. and trans.), *The Bamboo Grove* (1971), introduces *sijo* arranged by themes. WON KO (trans. and comp.), *Contemporary Korean Poetry* (1970), is another collection. PETER H. LEE (ed.), *The Silence of Love: Twentieth-Century Korean Poetry* (1980), contains translations of 16 major modern poets. DAVID R. MCCANN (trans.), *The Middle Hour: Selected Poems of Kim Chi Ha* (1980), contains 40 poems. Another selection of poems, prose pieces, and a play by the same author is presented in CHONG SUN KIM and SHELLY KILLEN (eds.), *The Gold Crowned Jesus and Other Writings* (1978).

Prose: RICHARD RUTT and CHONG-UN KIM (trans.), *Virtuous Women* (1974, reprinted 1979), contains translations of "Dream of Nine Clouds," "Tale of Queen Inhyŏn," and "The Song of a Faithful Wife, Ch'un Hyang." IN-SŎP CHŎNG (ed. and trans.), *Folk Tales from Korea* (1952, reprinted 1969), is a representative selection. DUK-SOON CHANG *et al.* (eds.), *The Folk Treasury of Korea: Sources in Myth, Legends, and Folktale*, trans. by TAE-SUNG KIM (1970), is a collection of oral literature. SOUN KIM, *The Story Bag* (1955), collects 30 folktales. For modern prose see KEVIN O'ROURKE (comp.), *Ten Korean Short Stories* (1973, reissued 1981); PETER H. LEE (ed.), *Flowers of Fire: Twentieth-Century Korean Stories* (1974, rev. ed. 1986); CHONG-WHA CHUNG (ed.), *Modern Korean Short Stories* (1980); CHONG-UN KIM (ed.), *Postwar Korean Short Stories*, 2nd ed. (1983); and JI-MOON SUH (trans.), *The Rainy Spell and Other Korean Stories* (1983).

Literary criticism: PETER H. LEE, *Korean Literature: Topics and Themes* (1965), is an introduction to Korean literature, and his *Songs of Flying Dragons: A Critical Reading* (1974) is an annotated translation of *Yongbi ŏch'ŏn ka* (1445–47). W.E. SKILLEND, *Kodae Sosŏl: A Survey of Korean Traditional Style Popular Novels* (1969), is a catalog of Korean fiction.

(B.-W.C./P.H.L.)

Kyōto

The capital of Japan for more than 1,000 years (from 794 to 1868), Kyōto (literally, "Capital City") has been called a variety of names through the centuries—Heian-kyō ("Capital of Peace and Tranquillity"), Miyako ("The Capital"), and Saikyō ("Western Capital"), its name after the Meiji Restoration (1868) when the Imperial Household moved to Tokyo. The contemporary phrase *sekai no Kyōto* ("the world's Kyōto") reflects the reception of Japanese culture abroad and Kyōto's own attempt to keep up with the times. Nevertheless, Kyōto is the centre of Japanese culture and of Buddhism, as well as of fine textiles and other traditional Japanese products. The deep feeling of the Japanese people for their culture and heritage is represented in their special relationship with Kyōto—all Japanese try to go there at least once in their lives, with almost a third of the country's population visiting the city annually.

Kyōto is located on the island of Honshu some 29 miles (47 kilometres) northeast of the industrial city of Ōsaka and about the same distance from Nara, another ancient centre of Japanese culture. Gently sloping downward from north to south, the city averages 180 feet (55 metres) above sea level, and it covers an area of about 236 square miles (611 square kilometres). It is the capital of Kyōto-fu (Kyōto Urban Prefecture), which is at the centre of the Kinki-chihō (Kinki Region). The city is also one of the centres (with nearby Ōsaka and Kōbe) of the Keihanshin Industrial Zone, the second largest urban and industrial agglomeration in Japan.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. Designated as the site of the new capital by the emperor Kammu, Kyōto was laid out in 794 on the model of Ch'ang-an (modern Sian), the capital of the T'ang dynasty in China. The plan consisted of a rectangular enclosure with a grid street pattern, 3.2 miles north to south and 2.8 miles east to west. The Imperial Palace, surrounded by government buildings, was in the north-central section of the city. Following Chinese precedent,

City layout

care was taken when the site was selected to protect the northern corners, from which, it was believed, evil spirits could gain access. Thus, Hiei-zan (Mt. Hiei; 2,782 feet) to the northeast and Atago-yama (Mt. Atago; 3,031 feet) to the northwest were considered natural guardians. Hiei-zan especially came to figure prominently in the fate of the city: between the 11th and 16th centuries warrior-monks from its Tendai Buddhist monastery complex frequently raided the city and influenced politics. The Kamo and Katsura rivers—before joining the Yodo-gawa (Yodo River) to the south of the city—were, respectively, the original eastern and western boundaries. But the attraction of the eastern hills kept the city from filling out to its original western border until after World War II. Kyōto is actually cradled in a saucer of hills on three sides that opens to the southwest toward Ōsaka.

Climate. During spring and fall Kyōto is in its prime; the rainy season (June–July), lasting three to four weeks, is unavoidable; and summers are hot and humid. Winter brings two or three light snows and a unique "chilling from below" (*sokobie*), which must be tolerated. The yearly mean temperature of Kyōto is about 59° F (15° C); the highest monthly mean, 80° F (27° C), is in August, and the lowest, 38° F (3° C), is in January. The average yearly rainfall is about 62 inches (1,574 millimetres).

The city layout. The original grid pattern of the streets has been retained. Numbered avenues run east and west, Shijō-dōri ("Fourth Street") being the busiest. Karasuma-dōri, running north from the Japanese National Railways station, divides the city roughly into halves. Under it is the single line of the municipal subway. Kyōto was the first city in Japan to have electric streetcars (starting in 1895), which eventually made it necessary to widen the major thoroughfares to allow for citywide service.

Kyōto is a city with few large factories or businesses, a fact reflected in the look of the inner city—shops and workshops, residences, and offices all standing side by side. Stringent building codes limit the height of buildings in order to preserve the overall look of the historic city. Characteristic of the architecture are tiled roofs and wood weathered to dark brown, but telephone poles (now made of concrete) and a forest of television antennas protrude at every turn. A typical Kyōto house presents a narrow and low front to the street, but as it recedes it gains in height and embellishment—all this a reflection of its past history and character: wariness of the marauding monk, the zealous revenue collector, or the curious neighbour. Rarely does one enter a home beyond the front vestibule; if one is invited in, it is good form to demur.

Architecture

Because of earthquakes and conflagrations, the attacks of monks from Hiei-zan, and the Ōnin War (1467–77), which utterly destroyed the city, Kyōto's historical architecture rarely predates the 17th century. Replacements and renovations, of course, followed previous plans, but the single shining example of Heian-period architecture remaining is the soaring Hōō-dō ("Phoenix Hall") of the Byōdō-in ("Byōdō Temple"), located a few miles southeast of the city on the banks of Uji-gawa.

Temples and shrines

Buddhist temples and Shintō shrines abound. Their grounds and those of the Kyōto Gosho ("Kyōto Imperial Palace") and Nijō-jo ("Nijō Castle") give Kyōto more green areas than most Japanese cities. Kyōto claims some 1,660 Buddhist temples, more than 400 Shintō shrines, and even some 90 Christian churches. Major Buddhist institutions include Higashi and Nishi Hongan-ji ("East and West Hongan" temples), the former with the world's largest wooden roof of its kind and the latter containing a repository of the grandeur of Toyotomi Hideyoshi; Ryōan-ji, with its famous rock-and-sand Zen garden; Tenryū-ji, in the Arashiyama district to the west; Kiyomizu-dera, built on stilts on the side of the eastern hills; and Kinkaku-ji, the "Golden Pavilion" (burned down by a deranged student in 1950 but rebuilt exactly), and Ginkaku-ji, the "Silver Pavilion," both of which were products of the Ashikaga shoguns' attraction to Zen. The great Shintō shrines are Kitano, Yasaka, and Heian, the last built in 1894 to commemorate the 1,100th anniversary of Kyōto's founding.

The buildings of the Kyōto Gosho, originally located farther west, date from 1855 and replaced in the same "monumental" Japanese style the earlier structures that were destroyed by fire. Nijō-jo, built by the Tokugawa shogunate, is a "token" castle, but it contains many cultural treasures; it is known for its "chirping floors" (to signal the approach of an intruder) and elaborate wall paintings of the Kanō school. The two foremost examples of traditional Japanese landscape architecture are the Katsura Rikyū ("Katsura Detached Palace") in the southwest corner of the city and the Shūgakuin Rikyū set in the northeast hills. Katsura underwent a complete renovation using perfectly matched modern materials; its buildings are models of Japanese architectural proportions. Shūgakuin contains three gardens, the third with an artificial lake. From there one can view the entire expanse of the city stretching out to the south.

THE PEOPLE

Kyōto is one of the largest cities in Japan. Its population—which includes a sizable foreign community comprising mainly Koreans (many brought there forcibly during World War II), Chinese, and Americans—has remained relatively stable for a number of years. Most of the city's residents live in the central districts, but increasingly people are moving to outlying and suburban areas.

A major item remaining on the municipal agenda has been how to assimilate the thousands of *burakumin,* the historical outcaste group, who live in segregated communities in the city. This has been a continuing social problem largely in the older urban areas of western Japan, particularly Kyōto, Ōsaka, and Kōbe. Despite the fact that the last discriminatory legal bars were removed in 1969, social and occupational progress has lagged.

THE ECONOMY

Industry. Kyōto is a city of thousands of medium and small industries, many of them family owned and operated. Traditional handicrafts abound, and their manufacture for the tourist trade is an important element of Kyōto's economic life. The central part of the city is crowded with small workshops, which produce such typical Japanese goods as fans, dolls, Buddhist altar fittings, and lacquer ware. Antipollution measures have forced the once-thriving Kiyomizu pottery kilns to move to nearby Yamashina.

Silk weaving

For centuries silk weaving, centred in the north-central Nishijin district, has been Kyōto's major industry. Along with the geisha and entertainment industry, the fine textiles, delicate fabrics, and embroidery represent a continuity of Kyōto's traditional role as the centre of Japanese culture. In addition, the Fushimi district in southern Kyōto, favoured with excellent water, produces some of Japan's finest sake. Also located in southern Kyōto are several industries established after World War II. The most important of these produce industrial ceramics, women's garments, and medical instruments.

Commerce. Kyōto is mainly a consumer city. It is the national centre of silk and fine textile wholesaling, but its main commercial activity is retail trade. The Gion and Pontocho districts, famed for their geisha and *maiko* (apprentice geisha), offer a variety of traditional and foreign food and drink. During the summer *yuka* (platforms on stilts) are set up on the banks of the Kamo-gawa in the heart of town, and strolling troubadours pass below as a reminder of how Kabuki originated. Traditional Japanese inns (*ryōkan*) abound, and many Western-style hotels cater to the wedding, tourist, and convention trades. A large conference centre near the foot of Hiei-zan hosts major industrial exhibitions and international conferences.

Transportation. Most of Japan's east–west traffic must come through Kyōto. During the Tokugawa period (1603–1867) the city was the western terminus of the Tōkaidō, the road that connected Kyōto to Edo (now Tokyo). River traffic to Ōsaka favoured the Yodo-gawa. Today the numerous, high-speed "bullet" trains of the Shinkansen give reliable service to major cities east and west. Interurban lines between Kyōto and Ōsaka–Kōbe and Nara provide fast and frequent local service. Kyōto itself finally abandoned streetcars in the 1970s. The Meishin Expressway links Kyōto to Ōsaka and Nagoya.

The main house of Katsura Detached Palace, Kyōto.
Camera Tokyo

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Kyōto Urban Prefecture, which extends to the Japan Sea, is under the administration of an elected governor, while the city is administered by an elected mayor and city council of 72 members.

Education. Kyōto was traditionally organized into extended neighbourhoods, called *machi*, and after the Meiji Restoration it was found that this was a good base by which to quickly establish general public education; as such, the city preceded the national effort to systematize primary education. Kyōto is surpassed only by Tokyo in its number of institutions of higher learning, but it claims several Nobel Prize laureates that Tokyo, with none, is reminded of from time to time. The city's relatively calm atmosphere, its distance from the hurly-burly of national government, and its numerous cultural and religious institutions and facilities are cited as prime reasons for its educational advantages. There are more than 40 two-year and four-year colleges and universities with a total annual enrollment of more than 100,000 students. The state-run Kyōto Daigaku (Kyōto University), established in 1897, is the nation's second most prestigious school. Dōshisha Daigaku, the leading private institution, was founded in 1875 by Niijima Jō (also called Joseph Hardy Neesima), the first Japanese to graduate from a Western college (Amherst College in 1870). Major Buddhist universities include Ryūkoku, Ōtani, and the smaller Hanazano.

Higher education

CULTURAL LIFE

The millennium as the nation's capital and residence of the Imperial family has meant that Kyōto devolved as the preserver of the Japanese "spirit." This is exemplified in its varied and unique cultural institutions: the schools of tea ceremony (*cha-no-yu*) and flower arranging (*ikebana*); the theatrical arts of Nō, Kabuki, and traditional dance; or the masterpieces of calligraphy, painting, sculpture, and architecture that can be found everywhere in the city. Kyōto is the repository of hundreds of designated "national treasures" and "important cultural objects," representing a significant proportion of the national total. Included among these are individuals who have been named "living national treasures" (*ningen kokuhō*) in recognition of their superior skills in the traditional arts and crafts.

Most of the important works of art are housed in Kyōto's temples and shrines, many of which are themselves listed as national treasures. Even institutions that do not normally display their collections periodically have public "airings" at which their treasures can be viewed. Kyōto also has numerous museums, including Kyōto Kokuritsu Hakubutsukan (Kyōto National Museum; founded 1889), containing national treasures; Kyōto-shi Bijitsukan (Kyōto Municipal Art Museum; 1933); and Kyōto-shi Dentō Sangyō Kaikan (Kyōto Municipal Commercial and Crafts Museum; 1976).

Museums

The birthplace of traditional Japanese drama, Kyōto maintains an active theatrical life. Several Nō stages offer frequent performances, and the annual opening performance (*kaomise*) at the Minami-za is the customary inauguration of the national Kabuki season. A traditional form of humorous pantomime, *Mibu kyōgen,* is performed faithfully by troupes of amateurs.

The three major festivals (*matsuri*)—Aoi in May, Gion in July, and Jidai in October—are almost national events. The Jidai-matsuri ("Festival of the Ages") is a parade depicting, in period costume, Japan's entire history. The Gion-matsuri dates from the 9th century and features more than 30 elaborate, carefully preserved, hand-drawn floats, some decorated with French Gobelin tapestries imported through Nagasaki during Tokugawa times. The northern hills—Hiei-zan with its scenic drive and the Takao district for its fall foliage—are famed for their well-tended stands of Japanese cedar (*sugi*).

History

The ancient city

Kyōto as the national capital dates from 794, although the area was settled earlier by Korean immigrants who brought with them the skills of sericulture and silk weaving. As noted above, the planned city was between the Katsura and Kamo rivers, but it soon extended beyond the eastern banks of the Kamo. The powerful Fujiwara family dominated the Heian period. Excessive Buddhist influence at the old capital of Nara had occasioned the removal of the government to Nagaoka and then to Kyōto, where the building of Buddhist temples was proscribed. As an exception, Rashōmon, the great southern gateway, was flanked by Tō-ji on the east and Sai-ji on the west; Sai-ji was short-lived, but the handsome, five-tiered pagoda of Tō-ji is a classic landmark.

Following the decline of the Fujiwara and the ascendance of the Minamoto in the late 12th century, political and military leadership was vested in a shogun ("generalissimo"), the first of whom, Minamoto Yoritomo, chose to administer the expanding domains from Kamakura to the east. It was during the Kamakura period (1192–1333) that many of the Buddhist temples were established, and indigenous sects of Buddhism, together with Zen from the continent, appeared. During the ensuing Muromachi period (1338–1573), the Ashikaga shogunate moved the government back to Kyōto. The aristocratic culture of the Heian era blended with the culture of Zen that had developed under the samurai (warriors), resulting in the refinement of the Nō theatre, the tea ceremony and flower arranging, and pottery making.

By the mid-16th century, however, the city had been so devastated that St. Francis Xavier, on a pilgrimage to Kyōto, could not even locate the Imperial court, much less seek an Imperial audience. The city's fortunes revived under the regimes of the national unifiers Oda Nobunaga and Toyotomi Hideyoshi. Buddhists, especially the Tendai monks on Hiei-zan, were such an anathema to Nobunaga that he set fire to the entire monastery complex; but under Hideyoshi, an ardent patron of the arts, Kyōto flourished. One of his tea parties was attended by thousands of people and went on for days.

With the ascendance of the Tokugawa shogunate at the beginning of the 17th century, the political centre again moved, this time to Edo (modern Tokyo). The Imperial court was left to pursue its ceremonial functions, and access to it was carefully monitored. Only after the arrival of Matthew Perry in 1853 and the collapse of the Tokugawa did Kyōto again come to the fore. At the Nijō-jo in 1867 the last Tokugawa shogun finally turned back to the Imperial court his mandate to rule the nation, marking the first time in more than 200 years that a ruling Tokugawa had set foot in Kyōto.

Shortly after the proclamation of the Meiji Restoration, however, the young Meiji emperor took up residence in the new capital, Tokyo—a move that has not been forgotten in Kyōto. Kyōto busied itself in outbidding Ōsaka to become in 1872 the site of an annual exhibition that was held for more than 30 years. During World War II U.S. Secretary of War Henry L. Stimson, recalling his visits to Kyōto, struck the city from the list of targets for aerial bombing. Its cultural treasures intact, it maintains a special place in the hearts of the Japanese and, increasingly, in the eyes of the world.

The modern city

BIBLIOGRAPHY. On the history of Kyōto, see JOHN WHITNEY HALL, *Japan from Prehistory to Modern Times* (1970); EDWIN O. REISCHAUER and JOHN K. FAIRBANK, *East Asia: The Great Tradition* (1960); JOHN K. FAIRBANK, EDWIN O. REISCHAUER, and ALBERT M. CRAIG, *East Asia: The Modern Transformation* (1965); and EDWIN BAYRD, *Kyoto* (1974). Descriptive works include GOUVERNEUR MOSHER, *Kyoto: A Contemplative Guide* (1964, reprinted 1978); DONALD KEENE, *Landscapes and Portraits: Appreciations of Japanese Culture* (1971, reissued as *Appreciations of Japanese Culture,* 1981); YOSHIKAZU IZUMOJI, *Kyoto,* 34th ed. (1983; originally published in Japanese, 1963); and TADASHI ISHIKAWA, *Palaces of Kyoto* (1968; originally published in Japanese, 1962), and *Imperial Villas of Kyoto* (1970). OTIS CARY, *Mr. Stimson's "Pet City": The Sparing of Kyoto, 1945* (1975), details Kyōto's survival during World War II. On social life and customs, see RUTH L. GAINES, *City-Royal: A Memory of Kyōto* (1953). HERBERT E. PLUTSCHOW, *Introducing Kyoto* (1979), is a good guidebook.

(O.C.)

Lakes

A lake is a body of slowly moving or standing water that occupies an inland basin. Definitions that precisely distinguish lakes, ponds, swamps, and even rivers and other bodies of nonoceanic water are not well established. It may be said, however, that rivers and streams are relatively fast moving; marshes and swamps contain relatively large quantities of grasses, trees, or shrubs; and ponds are relatively small in comparison to lakes. Geologically defined, lakes are temporary bodies of water.

This article treats lake basins and sedimentation; the physical and chemical properties of lake waters; lake currents, waves, and tides; and the hydrologic balance of lakes. For information on related systems, see the article RIVERS. The place of lakes within the hydrologic cycle is further dealt with in HYDROSPHERE; as are certain aspects of lake sedimentation and water chemistry. See ECOSYSTEMS for information on lacustrine life-forms.

This article is divided into the following sections:

General considerations

OCCURRENCE

Within the global hydrologic cycle, freshwater lakes play a very small quantitative role, constituting only about 0.009 percent of all free water, which amounts to less than 0.4 percent of all continental fresh water. Saline lakes and inland seas contain another 0.0075 percent of all free water. Freshwater lakes, however, contain well over 98 percent of the important surface waters available for use. Apart from that contained in saline bodies, most other continental waters are tied up in glaciers and ice sheets and the remainder is in groundwater.

Four-fifths of the 125,000 cubic kilometres (30,000 cubic miles) of lake waters occur in a small number of lakes, perhaps 40 in all. Lake Baikal, in central Asia, contains about 22,000 cubic kilometres (5,000 cubic miles) of water, and Lake Tanganyika (19,000 cubic kilometres [4,500 cubic miles]) and Lake Superior (12,000 cubic kilometres [3,000 cubic miles]) are the next largest. The Great Lakes of North America contain a total of about 25,000 cubic kilometres (6,000 cubic miles) of water and, together with other North American lakes larger than 10 cubic kilometres (two cubic miles), constitute about one-fourth of the world's lake waters.

Distribution

Although lakes are to be found throughout the world, the continents of North America, Africa, and Asia contain about 70 percent of the total lake water, the other continents being less generously endowed. A fourth of the total volume of lake water is spread throughout the world in uncounted numbers of small lakes. Anyone who has flown over much of the Canadian plains area cannot help but be struck by the seemingly endless skein of lakes and ponds covering the landscape below. Though the total volume of water involved is comparatively small, the surface area of lake water is substantial. The total surface area of all Canadian lakes has been estimated to exceed the total surface area of the province of Alberta. The state of Alaska has over 3,000,000 lakes with surface areas greater than eight hectares (20 acres).

The larger, deeper lakes are a significant factor within the cycle of water—from rain to surface water, ice, soil moisture, or groundwater and thence to water vapour. These lakes receive the drainage from vast tracts of land, store it, pass it on seaward, or lose it to the atmosphere by evaporation. On a local basis, even the smaller lakes play an important hydrologic role. The relatively high ratio of exposed surface area to the total water volume of these lakes accentuates their effectiveness as evaporators. In some cases the efficiency of lakes in losing water to the atmosphere is locally undesirable, because of demands for it by public and industrial requirements. In some basins, lakes are the terrestrial end point of the hydrologic cycle. With no outflow downstream toward the oceans, these closed lakes swell or recede according to the balance of local hydrologic conditions.

USES AND ABUSES OF LAKES

Importance to man

In today's industrial societies, requirements for water—much of which is derived from lakes—include its use for dilution and removal of municipal and industrial wastes, for cooling purposes, for irrigation, for power generation, and for local recreation and aesthetic displays. Obviously, these requirements vary considerably among regions, climates, and countries.

In another vein, it is convenient to use water to dilute liquid and some solid wastes to concentrations that are not intolerable to the elements of society that must be exposed to the effluent or wish to use it. The degree of dilution that may be acceptable varies from situation to situation and is often in dispute. In some cases, dilution is used purely to facilitate transport of the wastes to purification facilities. The water may then be made available for reuse.

Lake water is also used extensively for cooling purposes. Although this water may not be affected chemically, its change in thermal quality may be detrimental to the environment into which it is disposed, either directly, by affecting fish health or functions, or indirectly, by causing an excessive plant production and ultimate deoxygenation due to biological decay. Both fossil- and nuclear-fuelled power plants are major users of cooling water. Steel mills and various chemical plants also require large quantities.

Concern with thermal pollution of surface waters is concentrated principally on rivers and small lakes. With power requirements in modern societies increasing by about 7 percent per year, however, some apprehension has been expressed about the future thermal loading of even the largest lakes. It has been predicted that thermal inputs to each of the North American Great Lakes will increase by nearly 11 times by the year 2000. In terms of energy to be disposed in this fashion, the numbers are staggeringly large. These lakes have such large volumes, however, and such large surface areas (from which much of the heat goes into the atmosphere) that there is some question about the nature and magnitude of the actual effects.

The economic importance of waterways as communication links is enormous. In the earliest times, when travel by many societies was substantially by water, travel routes became established that resulted in relationships between cultural factors and surface hydrology networks. Today, river and lake systems serve as communication links and play an important role in shipping because of the large cargo capacities of merchant vessels and the still fairly uncongested condition of inland waterways. Oceanic shipping lanes play the major role, but river and lake systems, which link inland ports with the oceans, have been key factors in the rates of economic growth of many large inland ports.

Commercial fisheries and other food industries reap great harvests from the major lakes of the world. The quality of the fish catch has steadily decreased, however, as a result of pollution in many lakes, with the more desirable species becoming less plentiful and the less desirable species gradually dominating the total. Other commercial harvesting from lakes includes waterfowl, fur-bearing mammals, and some plant material, such as rice.

Each of the uses described has associated with it the means for abuse of the very characteristics of lakes that make them desirable. Wise management of natural resources has never been man's forte. Municipalities and industries have polluted lakes chemically and thermally, the shipping that plies large inland water bodies leaves oil and other refuse in its wake, water used for irrigation often contains chemical residues from fertilizers and biocides when it is returned to lakes, and the populace that so desperately demands clean bodies of water for its recreation often ignores basic sanitary and antipollution practices, to the ultimate detriment of the waters enjoyed.

Major problems

Among the major problems affecting the optimum utilization and conservation of lake waters are eutrophication (aging processes), chemical and biological poisoning, and decreases in water volumes. In the former case, discussed in more detail later, the enrichment of lakes with various nutrients supports biological productivity to an extent in which the ultimate death and decay of biological material places an excessive demand on the oxygen content, resulting in oxygen depletion in the worst cases. Phosphates and nitrates are two of the types of nutrients that are most important in this connection, particularly since they are often introduced in critical quantities in waste effluents from human sources. Other examples of chemical pollution of lakes include the introduction of DDT and other pesticides and heavy metals such as mercury. Bacteriological contamination of lake waters resulting in levels that constitute a hazard to health is another common result of man's inhumanity to his natural environment.

Water-quantity problems are complex, being related to natural vagaries of supply and levels of consumptive utilization of water. In the latter case, the percentage of water returned to the source after utilization varies with the use. The largest losses are due to actual water diversions and processes that result in evaporative losses. The use of large quantities of lake water for cooling purposes by industry and utilities, for example, may raise lake temperatures near the effluents sufficiently to cause increased evaporation. The use of certain types of cooling towers results in even larger losses. Some of the water evaporated will stay within the lake basin, but some will be lost from it.

Another example of this type of loss is connected with the possible application of weather-modification techniques to alleviate the heavy lake-effect snowfalls experienced along the lee shores of large lakes in intermediate latitudes. Redistribution of precipitation always raises the possibility of redistribution of water among various basins.

Lake-effect snowfall is just one example of the influence of lakes on local climate. The ability of large bodies of water to store heat during heating periods and to lose it more gradually than the adjacent landmasses during cooling periods results in a modifying influence on the climate. Because of this propensity, a lake cools air passing over it in summer and warms air passing over it in winter. Consequently, the predominantly downwind side of a lake is more influenced by the ameliorating effects of a lake.

In most instances, moisture is also passed to the atmosphere. In summer, lake cooling serves to stabilize the air mass, but winter heating tends to decrease stability. The moisture-laden, unstable winter flows off lakes produce so-called snowbelts, which affect downwind cities. The snowbelts are usually of limited extent, often within about a kilometre of the lake shore. (R.K.L./Ed.)

Lake basins

CLASSIFICATION OF BASINS

Limnology

The name given to the study of lakes is limnology. Limnologists have used several criteria for the development of systems for classifying lakes and lake basins but have resorted particularly to the mechanisms that have produced lake basins. These have been summarized and examined in *A Treatise on Limnology,* by the American limnologist G.E. Hutchinson, which includes treatment of tectonism, volcanism, landslides, glaciation, solution, river action, wind action, coastline building, organic accumulation, animal activity, meteoritic impact, and human activity.

Basins formed by tectonism, volcanism, and landslides. Tectonism—or movements of the Earth's crust—have been responsible for the formation of very large basins. During the late Miocene (the Miocene Epoch includes the time interval from 26,000,000 to 7,000,000 years ago), broad, gentle earth movements resulted in the isolation of a vast inland sea across southern Asia and southeastern Europe. Through most of the Tertiary Period (from 65,000,000 to 2,500,000 years ago), sub-basins developed that gradually were characterized by a great range of salinities. Resumption of communication with the oceans occurred later, and there is evidence of considerable variation in water levels. The present remnants of these inland bodies of water include the Caspian Sea and the Aral Sea, along with numerous smaller lakes. The Black Sea, which was also once part of this large inland basin, is now in direct communication with the oceans.

In some cases, elevated land areas may already contain depressions that eventually form lake basins. Lake Okeechobee, Florida, is cited as being such a basin, formed by uplift of the ocean floor.

Tectonic uplift may interfere with natural land-drainage patterns in such a way as to produce lake basins. The Great Basin of South Australia, some of the lakes in Central Africa (*e.g.,* Lakes Kioga and Kwania), and to some extent Lake Champlain, in the northeastern United States, are examples of this mechanism. Land subsidence, due to earthquake activity, also has resulted in the development of depressions in which lakes have evolved. Many such cases have been reported within the past 300 years.

The damming of valleys as a result of various tectonic phenomena has resulted in the formation of a few lake basins, but faulting, in its great variety of forms, has been responsible for the formation of many important lake basins. Abert Lake, in Oregon, lies in the depression formed by a tilted fault block against the higher block. Indeed, many lakes in the western United States are located in depressions formed through faulting, including Lake Tahoe, in the Sierra Nevada, California. Great Salt Lake, Utah, and other nearby salt lakes are remnants of a large Pleistocene lake, Lake Bonneville, which was formed at least partly by faulting activity (the Pleistocene Epoch includes the interval from about 2,500,000 to about 10,000 years ago).

In other parts of the world, faulting has also played an important role in basin formation. Lake Baikal and Lake Tanganyika, the two deepest lakes in the world, occupy basins formed by complexes of grabens (downdropped faulted blocks). These lakes are among the oldest of modern lakes, as are other graben lakes, particularly those within the East African rift system, which extends through the East African lake system and includes the Red Sea (see further CONTINENTAL LANDFORMS: *Rift valleys*).

Crater lakes and lava damming

Basins formed from volcanic activity are also greatly varied in type. The emanation of volcanic material from beneath the surface can be explosive, or it can issue in a gentle and regular manner. This range of activity and the variation of types of material which may be involved result in the possible development of many different types of basins.

One broad category includes those occupying the actual volcanic craters or their remnants. Crater lakes may occupy completely unmodified cinder cones, but these are rare. Craters caused by explosions or by the collapse of the roofs of underground magma (molten silica) chambers and those caused by explosion of new volcanic sources and that are built of nonvolcanic material are other examples. The latter are termed maars, following the local name for such forms in Germany. They are found, however, in several locations, including Iceland, Italy, and New Zealand. The maars of the volcanic district of Eifel, West Germany, are among the best known of these formations.

The collapse of magma chambers and the development of very large surface craters called calderas is an important source of lake basins. Crater Lake, Oregon, is a typical example, exhibiting characteristically great depth and a high encircling rim. Some caldera basins evolved with gently sloping sides, however, due to the deposition of material from a series of explosions and a gentler collapse of the structure. Secondary cones may develop within calderas, as shown by Wizard Island, in Crater Lake. The largest caldera in the world, which contains Lake Toba, in Sumatra, was formed through a combination of volcanic action and tectonic activity. Lake Toba's basin is contained in a strike-slip fault belt along the entire length of the Barisan Mountains of Sumatra. A vast, initial eruption of lava under gas pressure collapsed the magma reservoir, forming a depression that filled with water, producing the lake. Renewed volcanic activity subsequently led to the formation of an island in the centre, but a second collapse later cut it in two. Additional tectonic activity has further modified the lake's configuration.

Lake basins may also arise from the action of lava flows that emanate from volcanic fissures or craters. Lake Mývatyn, in Iceland, was formed in a basin arising from the collapse of the interior part of a large lava flow. Other basins have formed as the result of volcanic damming. This usually happens where a lava flow interrupts the existing drainage pattern.

Lake basins also may form following the blockage of a drainage depression by landslides. These may be temporary in nature because of the eroding action of the lake on the damming material. Lake Sārez in the Pamirs is stable, being dammed by a rockslide.

Basins formed by glaciation. The basin-forming mechanism responsible for the most abundant production of lakes, particularly in the Northern Hemisphere, is glaciation. The Pleistocene glaciers, which seem to have affected every continent, were especially effective in North America, Europe, and Asia. The retreat of ice sheets produced basins through mechanical action and through the damming effect of their ice masses at their boundaries.

In some cases, lakes actually exist in basins made of ice. In other cases, water masses may form within ice masses. Such occurrences are rare and are not very stable. Damming by ice masses is a more common phenomenon but is also likely to be relatively temporary. Glacial moraine (heterogeneous sedimentary deposits at glacier margins) is also responsible for the occurrence of dammed lake basins. The Finger Lakes of New York State are dammed by an endmoraine.

Ice-scour lakes, cirque lakes, and fjords

Ice sheets moving over relatively level surfaces have produced large numbers of small lake basins through scouring in many areas. This type of glacial rock basin contains what are known as ice-scour lakes and is represented in North America by basins in parts of the high Sierras and in west central Canada (*e.g.,* near Great Slave Lake). Tens of thousands of these lakes are found in the ice-scoured regions of the world. Many of them are interconnected with short streams and may contain narrow inlets. Characteristically, they may be dotted with numerous islands and sprawling bays. Many are comparatively shallow. Where they are particularly abundant, they may cover up to 75 percent of the total surface, as in Quetico canoe country of Minnesota.

Glacier scouring associated with the freezing and thawing of névé (granular snow adjacent to glacier ice) at the head of a glaciated valley may produce a deepened circular basin termed a cirque. These are found in widely scattered mountain locations. The action of glaciers in valleys can produce a similar type of basin, often occurring in series and resembling a valley staircase. Ice movement from valleys through narrow openings has produced another type of rock basin, known as glint lake basins, particularly in Scandinavian regions.

Piedmont and fjord (*i.e.*, a river valley that has been "drowned" by a rise of sea level) lakes are found in basins formed by glacial action in long mountain valleys. Excellent examples are found in Norway, the English Lake District, the European Alps, and the Andes. In North America, several regions contain this type of lake basin. In British Columbia, many good examples exist, the largest of which are the Okanagan and Kootenay systems. These are long, narrow lakes of substantial depth. In northwestern Canada, some of the largest lakes, including Lake Athabaska, Great Slave Lake, and Great Bear Lake, are of this type, although they are not found in the same type of mountainous terrain. These lakes, as well as the North American Great Lakes, resulted from the movements of large ice sheets that deepened existing valleys.

The Wisconsin (latest stage of Pleistocene glaciation) ice sheet was responsible for shaping the present Great Lakes system, which drains mainly eastward to the Atlantic through the St. Lawrence River, during its retreat. The principal stages in the history of these lakes have received much study, and several stages of retreat and advance of the ice sheet have been identified. Behind the lobes of the ice sheet, ice lakes developed that drained according to the modifications of preexisting valleys for glacial action. As the mass of ice retreated far to the north, glacial rebound (uplift of the Earth's crust in response to removal of the loading by ice) caused a general tilting of the land surface; the new lake basins also contributed to the subsequent changes through their own erosional action.

The material comprising glacial moraines or glacial outwash may provide dams that confine postglacial waters. The Finger Lakes, in New York State, constitute one interesting group of this type. These lakes were formed through glacial scouring of existing valleys, which were blocked at both the northern and southern ends by morainic deposits.

A variety of basin types have been formed in the different types of glacial drift deposits, including basins in morainic material, kettle lakes, channels formed by water movement in tunnels beneath the ice masses, and lake basins formed by thawing in permafrost. An interesting example of glacial action is the formation of giant's kettles; these are glacial potholes in the form of deep cylindrical holes. Their origin is still uncertain. Sand, gravel, or boulders are sometimes found at their bottom. The kettles vary from a few centimetres to a metre or more in diameter. Good examples are found in the Alps, Germany, Norway, and in the United States.

Formation of natural dams

Basins formed by fluvial and marine processes. Fluvial action in several forms can produce lake basins; the most important processes include waterfall action, damming by sediment deposition from a tributary (fluviatile dams), sediment deposition in river deltas, damming by tidal transport of sediments upstream, changes in the configuration of river channels (*e.g.*, oxbow lakes and levee lakes), and solution of subsurface rocks by groundwater.

This last mechanism has produced the well-known formations in the Karst region, in Yugoslavia, which include subterranean and surface cavities and basins in limestone. The term karstic phenomena is applied to similar cases in many parts of the world (see further CONTINENTAL LANDFORMS: *Caves and karst landscape features*). Solution lakes in Florida (*e.g.*, Deep Lake) are also of this origin, as are Lünersee and Seewlisee, in the Alps. Other rock types susceptible to solution basin formation include gypsum and halite. Mansfeldersee in Saxony was formed in this manner.

In some coastal areas, longshore marine currents may deposit sufficient sediment to block river outflows. This damming action may be of varying intensity, and it may also occur in lake regions, where such current action causes sediment deposition that leads to the formation of multiple lakes. Accumulation of organic plant material can also result in structures that produce lake basins; Silver Lake, Nova Scotia, evolved from damming by plant material. Structural formations of coral are another potential cause of damming.

Basins formed by wind action, animal activity, and meteorites. Wind action may lead to dam or dune construction or erosion and thus can play a role in lake-basin formation. The latter case has been demonstrated in North America; a number of basins in Texas and northward, on the plains east from the Rocky Mountains, are thought to have originated from wind erosion—at least in part. Moses Lake in Washington state was formed by windblown sand that dammed the basin.

Mammals have constructed lake-forming dams; the American beaver is highly skilled at this, and its activities in this connection have established it as a symbol of industriousness. Man has also been busy in this regard and is fully capable of producing lakes that would rival the largest of the more natural variety. Plans once proposed for the damming of the Yukon River in Alaska would, if carried through, result in the formation of a lake larger than Lake Erie in surface area. Other human activities, such as quarrying and mining, also have produced cavities suitable for lake formation.

The last major mechanism of basin formation is that due to meteoritic impact. Meteorite craters are best preserved in arid climates and are often dry for this reason. A few lakes are known in craters, however, including Ungava Lake, in Quebec. In many other cases, it has not been possible to definitely confirm that basins that have the appearance of meteorite craters have, indeed, been produced by meteorite impact. Controversial ones include the bay lakes of southeast North America.

TOPOGRAPHY OF BASINS

Lakes meet with both the atmosphere and the underlying material of their terrestrial basins and interact with each. The topography and configuration of the lake bottom and the nature of the bottom materials vary considerably. They are of sufficient importance to most lake processes to warrant recognition as basic lake characteristics.

The surface area of a lake can easily be determined by cartographic techniques, but lake-volume determinations require knowledge of lake depths. Throughout the world, lakes important enough to warrant study have been sounded, and many nations have completed comprehensive programs to determine the bathymetry of large numbers of lakes. Lake sounding involves traversing a lake to collect either point or continuous measurements of depth until an accurate survey is made. Modern sounding devices measure the time taken for emitted sound to return after reflection from the bottom, relying on a knowledge of the speed of sound in water. The more sophisticated of these also provide for detection of the depths of stratification in sedimentary materials on the lake bottom. The employment of laser devices from aircraft is a recent development that is based on the transmission of light beams with wavelengths that will penetrate water.

For more practical purposes, lake morphology is a stable characteristic. Shore erosion, sediment deposition and transfer, and other processes, however, including dredging by man, may significantly alter a lake's bottom topography and thus affect navigation, currents, and ecological factors, such as fish spawning grounds.

SEDIMENTS AND SEDIMENTATION

Lake sediments are comprised mainly of clastic material (sediment of clay, silt, and sand sizes), organic debris, chemical precipitates, or combinations of these. The relative abundance of each depends upon the nature of the local drainage basin, the climate, and the relative age of a lake. The sediments of a lake in a glaciated basin, for example, will first receive coarse clastics, then finer clastics, chemical precipitates, and then increasingly large amounts of biological material, including peats and sedges.

Clues to lake history

Geologists can deduce much about a lake's history and the history of the lake basin and climate from the sedimentary records on its bottom. A sediment core contains such clues as ripple marks caused by current or wave action, carbonaceous layers, and alternations of strata that include cold- and warm-water species of fossils, pollen, and traces of chemicals of human derivation. These data provide the basis for extensive documentation of lake history (paleolimnology). Some well-known historical events, such as major volcanic eruptions, the clearing of North American forests by early settlers as revealed by pollen concentrations, the first extensive use of certain heavy metals by industry, and nuclear explosions, provide reference points in the sediment record.

Many of the materials that are detrimental to the ecology of a lake—*e.g.,* excessive quantities of nutrients, heavy metals, pesticides, oil, and certain bacteria—are deposited in lake sediments by chemical precipitation or the settling of particulate matter. These materials are potentially available for regeneration into the lake water and must be considered in any planning for measures to abate lake pollution. Within the uppermost lake sediments, large volumes of interstitial water are often present. This water may have high concentrations of nutrients and other constituents and enhance the exchange potential with the lake proper.

Clastic sediments. Waters draining into a lake carry with them much of the suspended sediment that is transported by rivers and streams from the local drainage basin. Current and wave action along the shoreline is responsible for additional erosion and sediment deposition, and some material may be introduced as a result of wind action. Rivers and streams transport material of many different sizes, the largest being rolled along the riverbed (the bed load). When river water enters a lake, its speed diminishes rapidly, bed-load transport ceases, and the suspended load begins to settle to the bottom, the largest sizes first. Lake outlets carry with them only those materials that are too small to have settled out from the inflows or those that have been introduced adjacent to the outflow. Because dynamic processes that keep materials suspended are generally more active near the shore, lake sediments are usually sorted by size; the rocks, pebbles, and coarse sands occur near shore, whereas the finer sands, silts, and muds are, in most cases, found offshore.

Clastic material over most of a lake basin consists principally of silts and clays, especially away from shores and river mouths, where larger material is deposited. Clays exist in a variety of colours, black clays containing large concentrations of organic matter or sulfides and whiter clays usually containing high calcium carbonate concentrations. Other colours, including reds and greens, are known to reflect particular chemical and biological influences.

Organic sediments are derived from plant and animal matter: *Förna* is recognizable plant and animal remains, *äfja* finely divided remains in colloidal suspension, and *gyttja* is a deposit formed from *äfja* that has been oxidized. Rapid accumulation of organic matter in still lakes is not uncommon; in the English Lake District, five metres (15 feet) of lake sediment of organic origin accumulated in 8,000 years. Pollen analysis has been used to accurately decipher climatic conditions of the lake in the past.

Varved deposits are the product of an annual cycle of sedimentation; seasonal changes are responsible for the information. Varves are a common feature in many areas and especially so where the land has received meltwaters from ice sheets and glaciers. The deposits consist of alternating layers of fine and coarse sediments.

Coarse clastic materials seldom are larger than boulders (25 centimetres [10 inches]), and the type of material in sizes larger than silt and clay often reveals its source. Materials along lakeshores can usually be traced back to a particular eroded source within the local drainage basin, and the distribution of this material provides evidence of the predominant current or wave patterns in the lake.

Volcanic ash is deposited downwind from its source. Ash from volcanic activity during the Pleistocene Epoch can often be dated and used as a stratigraphic marker. Lakes throughout the northwestern United States contain some of the best examples (the Mazama ash), and one deposit in the central United States, called the Pearlette ash deposit, occurs in beds as thick as three metres (10 feet).

Deposition of salts

Chemical precipitates. The major chemical precipitates in lake systems are calcium, sodium, and magnesium carbonates and dolomite, gypsum, halite, and sulfate salts. Calcium carbonate is deposited as either calcite or aragonite when a lake becomes saturated with calcium and bicarbonate ions. Photosynthesis can also generate precipitation of calcium carbonate, when plant material takes up carbon dioxide and bicarbonate and raises the pH above about 9 (the term pH is a measure of the acidity or alkalinity of water; acid waters have a pH of less than 7, and the pH of alkaline waters range from 7 to 14).

Dolomite deposition occurs in very alkaline lakes when calcium carbonate and magnesium carbonate combine. Recent dolomites have been found in Lake Balkhash, in the Soviet Union. In many saline lakes, gypsum deposition has occurred; Lake Eyre, Australia, is estimated to contain more than 4,000,000,000 tons of gypsum. For gypsum to be deposited, sulfate, calcium, and hydrogen sulfide must be present in particular concentrations. Hydrogen sulfide occurs in deoxygenated portions of lakes, usually following the depletion of oxygen resulting from decomposition of biological material. Bottom-dwelling organisms are usually absent.

Lakes that contain high concentrations of sodium sulfate are called bitter lakes, and those containing sodium carbonate are called alkali lakes. Soda Lake, California, is estimated to contain nearly 1,000,000 tons of anhydrous sulfate. Magnesium salts of these types are also quite common and can be found in the same sediments as the sodium salts. Other salts of importance occurring in lake sediments include borates, nitrates, and potash. Small quantities of borax are found in various lakes throughout the world. Mono Lake, in California, due to its high alkalinity is devoid of any life.

The gradual increase of sediment thickness through time may threaten the very existence of a lake. When a lake becomes shallow enough to support the growth of bottom-attached plants, these may accelerate the extinction of a lake. In several European countries, steps are being taken to restore lakes threatened by choking plant growth. Lake Hornborgesjön, Sweden, which has been prized as a national wildlife refuge, became the subject of an investigation in 1967 that may lead to a program involving the cutting and removal of plants that have seriously reduced the area of exposed water. It is expected that the lake level may be raised at least one metre above the present mean summer level. Lake Trummen, also in Sweden, has been treated by dredging its upper sediments. In Switzerland, Lake Wiler (Wilersee) has been treated by the removal of water just above the sediments during stagnation periods.

Lake waters

CHEMICAL COMPOSITION

Although the chemical composition of lakes varies considerably throughout the world, due to the varying chemistry of the erosion products of different lake basins, in most cases the principal constituents are similar. Human influences have also contributed greatly to the chemical makeup of lakes, and, although industrial effluents vary from lake to lake, many of the chemical effects of human activities are similar throughout the world. Another source in the chemical balance of lakes is the dissolved and suspended material contained in precipitation. Again, human activities have been responsible for steadily increasing concentrations of this input.

Concentration of salts in lake water

Salinity, nutrients, and oxygen. Salinity is the total concentration of the ions present in lake water and is usually computed from the sodium, potassium, magnesium, calcium, carbonate, silicate, and halide concentrations. Several important bodies of inland waters, often called inland seas, have very high salinities. Great Salt Lake, in Utah, has a salinity of about 200,000 milligrams per litre, as compared to Lake Superior's value of about 75 and an estimated mean for all rivers of 100 to 150. These ions are steadily introduced to lakes from rivers and rainwater, where they concentrate because of the evaporative loss of relatively pure water.

Where inflowing rivers erode igneous rocks, lake salinity values are relatively low, but, where soluble salts are available for erosion, salinities are relatively high. In general, it has been found that, of the cations (positively charged ions), calcium concentrations are highest, followed by magnesium, sodium, and potassium, in that order. Of the major anions (negatively charged ions), carbonate is usually the most abundant, followed by sulfate and chloride.

Other inorganic ions, though present in smaller concentrations, are of great importance. In particular, the nutrients (especially phosphate, nitrate, and silicate), heavy metals (*e.g.,* mercury, manganese, copper, lead), and polychlorinated hydrocarbons (DDT, for example) have attracted recent interest because of their role in ecological problems. Although sources of nutrients and mercury exist that are not directly related to human activities, budget studies and studies of the historical records available in sediment cores clearly reveal the great impact of human disposal of these constituents in lakes. Rainfall and dry fallout are small but significant chemical inputs to lakes. The release of gases and particulate matter into the atmosphere from factories and similar sources has increased dramatically in recent years, with consequent alterations in the chemistry of rainwater. It has been estimated, for example, that 16,000 tons of nitrogen, about 8 percent of the total from all sources, is introduced annually to Lake Erie from atmospheric action.

Factors affecting oxygen concentration within lakes

The substance of most interest in lakes is oxygen; once introduced to the lake water, its concentration is subject to factors within the water. Biological production (photosynthesis) releases oxygen into the water, while biological decay consumes it. Various chemical reactions within the lake system also affect the concentration of dissolved oxygen. The main source is the passage of oxygen through the air-water interface, which is affected principally by the lake temperatures; at low temperatures the partial pressure of dissolved oxygen in water is reduced. Consequently, during cold seasons, especially when vertical mixing is greatly enhanced due to lack of thermal structure and increased wind stirring, lakes are replenished with oxygen. In the warmer seasons, although surface waters may remain more or less saturated and even supersaturated, the concentrations are lower. Beneath the surface, oxygen consumption due to biological decay may cause serious depletion. Oxygen depletion also occurs near the bottom due to processes at the mud-water interface, many of which are still inadequately explained.

In winter months, a rapid formation of ice or the establishment of strong winter thermal stratification may significantly inhibit the replenishment of oxygen. Where ice cover lasts for long periods, a loss of oxygen at the mud-water interface may have repercussions for the whole lake, particularly if density currents cause significant vertical transport.

In tropical regions, where the winter replenishment mechanism (turnover) is absent, there is great reliance on the occasional occurrence of cold spells or on significant nighttime cooling to promote oxygen replenishment. Deep lakes in these regions are often anoxic (lacking in oxygen) in the deeper portions.

At any particular time, lake waters or waters entering a lake may have a biological or chemical potential for oxygen utilization. Measurements of this are termed BOD (biological oxygen demand) or COD (chemical oxygen demand). These concepts are used as partial indicators of the quality of waters being introduced to a lake.

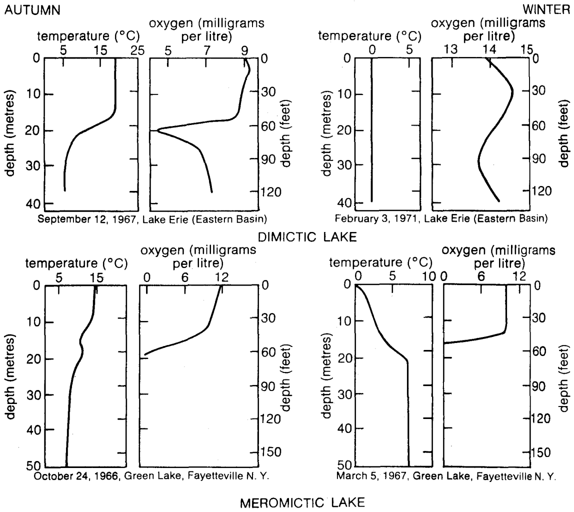


Figure 1: Seasonal variation of temperature and dissolved oxygen in dimictic (having two circulation periods annually) and meromictic (undergoing incomplete circulation at the fall overturn) lakes (see text).

Lakes that have a vertical salinity gradient strong enough to prevent winter turnover will usually be deoxygenated at depths where the vertical diffusion of oxygen is less than the oxygen demand. Such lakes are termed meromictic.

Carbon dioxide. Another gaseous substance of great importance that is exchanged with the atmosphere at the surface is carbon dioxide. Photosynthesis requires the presence of carbon dioxide, and it is released during biological breakdown.

Relation of carbon dioxide to pH and carbonate solubility

Carbon dioxide is very soluble in lake water; it forms carbonic acid, which dissociates and raises the concentration of hydrogen ions (lowering the pH). The relative proportions of bicarbonate, carbonate, and free carbon dioxide depend upon the pH. At high values of pH, carbonate ions will predominate; at low values, free carbon dioxide and carbonic acid will predominate.

Various carbonates (particularly sodium, calcium, potassium, and magnesium) are important to the carbon dioxide system. Increased pressure of carbon dioxide in the system increases the solubilities of these carbonates. In some cases, photosynthetic activity results in precipitation of certain carbonates. The entire carbon dioxide system and its behaviour at various pH values is very complex but can be interpreted from historical knowledge of lake sediments.

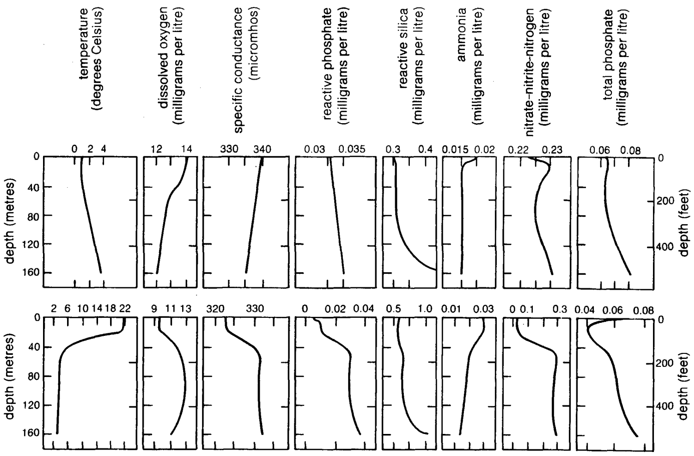


Figure 2: Distribution of some physical and chemical variables in central Lake Ontario in (bottom) July 1969 and (top) June 1970.

In waters that are neither very acidic (pH much less than 7) nor very basic (pH much greater than 7 but less than 14), the carbon dioxide system serves as a buffer, because, within limits, a change in pH will cause a shift within the system that ultimately serves to offset the pH change. Consequently, most lakes have a pH between 6 and 8. Some volcanic lakes are extremely acid, however, with pH values below 4, and some lakes with very high pH values, such as Lake Nakuru, Kenya, also occur in nature.

Sulfates, nitrates, and phosphates. Sulfate usually occurs as a principal ion in lake waters. Under anaerobic conditions in which bacteria persist in the oxidation of biological material, hydrogen sulfide is produced. When anoxic conditions exist in the deep waters just above the sediments, and the water is acidic enough to precipitate the iron present, hydrogen sulfide occurs. The characteristic and unpleasant odour of this gas is often popularly identified with the "death" of a lake. Big Soda Lake, Nevada, is extremely rich in this substance.

Nitrogen and its various compounds form another complex system in lakes, appearing as free nitrogen in solution, organic compounds, ammonia, nitrite, and nitrate. Sources of nitrogen compounds include influents to the lake (the most important source), fixation in the lake, and precipitation. Losses are experienced mainly through effluents but also by denitrification, sediment formation, and loss to the atmosphere.

Orthophosphate and various organic phosphates are the most important phosphorous compounds in lakes. Phosphates and nitrates are heavily consumed in the upper portion of lakes during periods of high productivity of phytoplankton. Increased concentrations occur in deeper portions due to decay of falling biological material and regeneration from the sediments, especially during anoxic conditions or stormy periods in shallow lakes. As limiting nutrients in many lake productivity cycles, phosphates and nitrates are often identified as controllable elements in situations where abatement is necessary to control eutrophication. Carbon is also a necessary constituent for production and in some cases can be the limiting component. Because carbon is less easily controlled and not often limiting, however, phosphates are most frequently named as substances to be reduced in effluents from industry and municipalities.

Silica also is present in lake waters, and, as with the other nutrients, it is introduced in influents and to some extent from the sediments. The production of diatom blooms is a major process for reducing silicate concentrations. Within this context, silica can also be regarded as a limiting nutrient.

THERMAL PROPERTIES

Basic physical data

Pure water freezes at 0° C (32° F), boils at 100° C (212° F), and has a latent heat of evaporation of 539.55 calories per gram, a latent heat of sublimation (ice) of 679 calories per gram, and a specific heat of 1.01 calories per gram, per ° C, at 0° C. The temperature of maximum density at atmospheric pressure occurs at 3.94° C (39.09° F). At the freezing point, ice has a lower density than water. For natural waters with high salinities, such as the oceans and inland seas, each of the values above is significantly altered. In most lakes, however, they are quite representative.

The density of water increases at pressures above one atmosphere (the pressure at sea level). Thus, pure water at 10°C (50° F) has a density of 0.9995 at one atmosphere and 1.0037 at the pressure existing at a lake depth of 1,000 metres (3,000 feet). Water raised from great depths to conditions of lower pressure experiences adiabatic cooling (without significant heat exchange with surrounding water), but there are very few lakes in which this factor can be of much significance.

Vertical mixing and overturn. It is useful to know how the temperature of maximum density changes with depth (*e.g.*, from 3.94° C at the surface to 3.39° C at 500 metres depth [38.10° F at 1,500 feet]). The fact that the temperature of maximum density of most lake waters is close to 4° C (39° F), whereas ice forms at temperatures close to 0° C in response to surface cooling, vertical mixing takes place. When density increases with depth, the lake is said to be stable. Unstable conditions exist when density decreases with depth. Cooling at the surface to temperatures below 4° C establishes stability based on a negative thermocline (a positive thermocline is a vertical decrease in temperature with depth), because density will increase with depth. Ice then forms at the surface, enabling liquid water to exist beneath the ice in lakes, unless they are shallow enough to freeze to the bottom.

During the warming season, after ice has melted, heating increases the density of the surface waters, causing them to sink until stability is achieved. When surface heating proceeds above the temperature of maximum density, this process ceases, and the vertical thermal structure maintains and strengthens its stable condition, based on a positive thermocline. Turnovers tend to be seasonal.

Dimictic, monomictic, holomictic, and meromictic lakes

Mixing due to cooling or warming processes that increase the density of surface waters sufficiently to cause them to sink results in what is termed circulation, or overturn, of lake waters. Lakes that cool to below 4° C in winter experience two turnover periods, as just described, and are called dimictic lakes. Most lakes in temperate regions fall into this category. Lakes that do not cool to below 4° C undergo overturn only once per year and are called warm monomictic. Lakes that do not warm to above 4° C also experience only one overturn period per year and are called cold monomictic. There are many examples of the former, including lakes in the tropical regions and generally as far north as about 40°. The cold monomictic type, however, is less common but can be found at high latitudes and at high altitudes (in the Alps, for example).

All the types described that circulate at least once throughout are called holomictic. It is possible, however, for lakes to be stable despite the thermal processes that normally induce overturn due to the existence of a positive salinity gradient with depth (chemocline). This type is called meromictic, and, in those cases where stability is permanent in at least part of the lake, the deep waters do not experience overturn and consequently are deoxygenated. Three principal origins of meromixis have been recognized. Ectogenic meromixis results from either the intrusion of seawater into a lake, as in the case of flooding from an unusually high sea level (*e.g.*, Hemmelsdorfersee, in Germany), or the introduction of fresh water through land drainage and precipitation to a saline lake (*e.g.*, Soda Lake, Nevada). Crenogenic meromixis is due to the introduction of saline water by springs, and biogenic meromixis is due to the uptake of salts from the lake sediments. North American examples include Lake Mary, Wisconsin, and Sodon Lake, Michigan.

A strong vertical salinity gradient that exists in the upper portion of a lake will affect the thermal structure by inhibiting the downward mixing of heat. In holomictic lakes, however, the downward mixing of heat due primarily to wind action usually compresses or concentrates the thermocline until it essentially separates an upper layer (epilimnion) from a lower layer (hypolimnion), each possessing weak or nonexistent vertical thermal gradients. The thermocline normally begins to grow at the beginning of the warming season. As summer passes and autumn commences, it intensifies and deepens. The onset of the cooling sees the beginning of the decay of the thermocline from above, although it usually continues to deepen until it is completely destroyed. The process just described is commonly found in lakes in temperate regions and is a seasonal phenomenon. During any period of strong warming, one or more shallower thermoclines may be observed to develop and move downward to the seasonal thermocline.

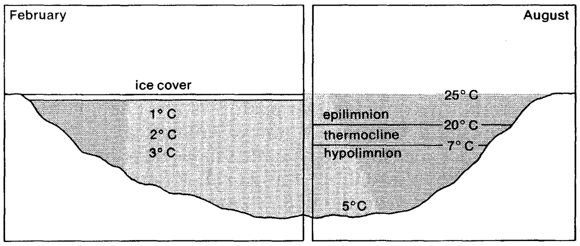


Figure 3: Temperature distribution in lakes in the middle latitudes of North America in winter and summer.

The heat budget of lakes. The heat budget of a lake includes several major factors: net incoming solar radiation, net exchange of long-wave radiation emitted by the lake surface and the atmosphere, transfer of sensible heat at the surface interface, and latent-heat processes. Those processes that are usually of much smaller importance include net inflow and outflow of heat advected by streamflow, precipitation, and groundwater flow, conduction from terrestrial heat flow, and dissipation of kinetic energy. In some cases, however, river inflow may be of more importance, such as where flow is from a nearby glacier or where the volume inflow is a significant fraction of the lake volume. Within a large lake the heat-budget considerations for a particular location must also take into account the local advection of heat within the lake by currents.

Seasonal variations in incoming solar radiation

Incoming solar energy varies seasonally and with the latitude and is greatly influenced by cloud cover. The fraction that is reflected away from the lake surface depends upon the solar angle, the turbidity of the atmosphere, and the wave state, or surface roughness. In middle latitudes this ranges from about 6 percent in summer months to about 14 percent in winter.

The amount of radiation emitted by the lake surface is proportional to the fourth power of the surface temperature, whereas the radiation emitted by the clouds and atmosphere overlying the lake depends primarily upon the amount and height of the clouds and the temperature and moisture content of the atmosphere near the lake surface.

The fluxes of sensible heat and moisture at the lake surface are of great importance yet are still poorly understood. They depend upon the vertical gradients of temperature and vapour pressure above the water, respectively, and upon the factors that influence the transfer processes, such as wind and atmospheric stability. The transfer of sensible heat may be either into or out of the lake surface, usually on a seasonal basis but also sometimes on a diurnal basis. It is also possible but less likely for condensation to occur on a lake surface.

Heat flow through the bottom of lakes is normally of small significance, but exceptions exist. In a very deep lake where low rates of heating are important, such as Lake Baikal, Soviet Union, the results may be detectable. In some ice-covered lakes where other sources of heating are small, heat flux through sediments also has been shown to be significant.

The dissipation of wind energy that has been transferred to water movements is quite insignificant, as is the effect of heat transfer due to chemical and photosynthetic processes.

In latitudes and altitudes where ice is a factor, the latent heats of fusion and of evaporation of ice must also be considered within any heat-budget considerations. Heat-balance studies have been performed for lakes that are always ice covered. Solar radiation is often an important factor where ice thickness and consistency permits penetration. The heat balance of the ice is often difficult to assess, as long-wave radiation and evaporation factors are not easily measured and are very important. The exchange of sensible heat may not be large during summer months in these cases but is likely to be significant in the colder months. Several lakes that are ice covered have been shown to be meromictic; two examples are Lake Tuborg, Ellesmere Island and Lake Bonney, Antarctica.

Heat balance of two representative lakes

Heat-balance measurements or estimates have been made for many lakes throughout the world. Results show that the difference between the highest and lowest heat content for each lake varies from around 5,000 calories per square centimetre for high and low latitudes to around 45,000 calories for some midlatitude lakes.

The relative importance of each of the major terms of the heat budget is shown by data for two North American lakes: Lake Ontario, a large, deep, middle-latitude lake; and Lake Hefner, a relatively small, shallow lake in Oklahoma. The energy unit frequently employed is the langley (one gram calorie per square centimetre), and the figures given are approximate monthly means of langleys per day. Net solar radiation input to Lake Ontario varies from 80 to 600 (Lake Hefner varies from 200 to 600), midwinter to midsummer. Net losses due to long-wave radiation from Lake Ontario are nearly 100 throughout the year (Lake Hefner varies from 100 to 200). Evaporation losses for Lake Ontario vary from 250 in midwinter to slightly negative values in early summer (Lake Hefner varies from 450 in late summer to 150 in spring). Conduction of heat from the surface of Lake Ontario varies from 250 in winter to about minus 100 in summer (Lake Hefner varies considerably from 80 to -80 for the same time interval).

Heat added to a lake at the surface is usually mixed mechanically downward as a result of wind action. This process keeps the upper portion of a lake relatively uniform thermally. Consequently, a thermal gradient (thermocline) becomes established between the upper mixed layer (epilimnion) and the deep portion of the lake (hypolimnion). In shallow lakes or shallow portions of large lakes, the thermocline will eventually intercept the lake bottom so that no hypolimnion exists. Normally, as the heating season progresses, the thermocline intensifies and deepens. Secondary thermoclines may develop in the epilimnion, and these will migrate downward to the main seasonal thermocline. On very warm, still days, a thin surface layer may store heat before a mixing episode transfers heat downward. When the cooling season commences, the mixing that tends to destroy the thermocline is enhanced by vertical convection. If the cooling continues until the entire thermocline is eliminated, the lake becomes essentially isothermal and no longer exhibits the characteristics of a two-layered system.

When a lake is stratified, the most important process for downward transfer of heat to the hypolimnion is through eddy conduction. The coefficient of eddy conductivity is determined empirically and varies substantially from lake to lake. Mixing processes are generally more active in coastal areas, so that isotherms can be expected to slope downward toward shore. In large, relatively unprotected lakes, wind stress at the surface causes convergence or divergence or both of shallow waters along coastlines. Isotherms will slant upward toward the shore, and hypolimnion water may even become exposed at the surface. These occurrences are of great importance with regard to the distribution of heat within stratified lakes.

Thermal pollution in lakes

Heat introduced to lakes in large quantities, as a waste product of cooling processes in power-generating plants and other industrial concerns, is presently viewed with some concern as a pollutant, especially in small lakes. If the heat is injected at the surface it will spread initially according to the momentum of the influent and the speed and direction of ambient surface currents. When the initial momentum is sufficiently dissipated, the heat will spread mainly as a consequence of turbulent mixing processes. Throughout these events, substantial losses of heat to the atmosphere may occur, so that the full effects of the thermal input are not borne solely by the lake. Temperature values at the surface, adjacent to the influent-heat source, may be raised to a very high level—as much as several centigrade degrees. Under certain conditions fish-activity tolerances may be exceeded, and undesirable algae and plankton production may be stimulated.

If waste heat is not released at the surface but is diffused over a large depth range or injected at depth, the large local-surface-temperature problem is avoided. Losses to the atmosphere in this case, however, are also greatly reduced, and the net heat input to the lake as a whole is much greater. Over a long period, this may prove to be more detrimental to the general ecology than near-surface injection.

Lake hydraulics

CURRENTS

The principal forces acting to initiate water movements in lakes are those due to hydraulic gradients, wind stress, and factors that cause horizontal or vertical density gradients. Lake water movement is usually classified as being turbulent.

Hydraulic effects are frequently the result of inflows and outflows of water. These may be substantial and continuous or weak and sporadic; in terms of the ratio of the volume of the inflow or outflow to the lake volume, the latter is the most frequently observed situation.

The stress of wind moving over the lake surface causes a transport of water within the lake, as well as the movement of energy downwind through the mechanism of surface waves. The wind is therefore one of the most important external forces on a lake. It can be relatively consistent in speed and direction, or it can be highly variable in either or both.

Pressure gradients. Water movements can occur as a result of internal pressure gradients and from density gradients caused by variations in temperature, sediment concentration, or the concentration of dissolved substances. Surface water in lakes can become denser than underlying water either by cooling or heating, because the temperature of maximum density for pure lake water is about 4° C (39° F). Water entering a lake from rivers with a high concentration of dissolved substances will sink to a lake level of similar density. These movements are both horizontal and vertical, but the net effect is downward, if not vertical, motion.

Horizontal pressure gradients can result from many different processes that act to produce density gradients.

One example is the situation of solar heating in a shallow nearshore region, where the heat is committed to the warming of a relatively small volume of water. This produces a water of lower density than the near-surface water of an adjacent deep region, where the heat is spread throughout a greater volume. Consequently, the pressure gradient force will act to move the warmer water offshore and to replace it from below with cooler water.

Influence of the Coriolis effect

Lake currents are the result of complex interactions of forces, but in many cases a small number of particular forces dominate. In the case of horizontal flow in the absence of horizontal pressure gradients, assuming no friction, water set in motion will curve to the right in the Northern Hemisphere because the Earth rotates from west to east. This effect is called the Coriolis force, and it will continue to influence water motion until there is a balance with the centrifugal force. This movement causes free-floating markers to move in an elliptical manner with a period that depends upon the latitude. In Lake Ontario, for example, it is about 17 hours. Where a dominating pressure gradient exists, the balance of the pressure-gradient force with the Coriolis force results in the so-called geostrophic flow, at right angles to the pressure gradient, with low pressure on the left (Northern Hemisphere). These conditions are most nearly realized only in very large lakes and in the oceans.

In those small lakes where hydraulic effects dominate, steady flow conditions may be achieved through balance with friction. This situation is commonly encountered in rivers, and relationships exist between mean current speed and the slope and mean depth of the river or narrow lake. These are called gradient currents and occur following situations where the wind or atmosphere pressure gradient causes a tilting of the lake surface (denivellation). In cases where the Coriolis force is a significant factor, the flow down a lake will tend to move toward the right (in the Northern Hemisphere). The development of a deeper countercurrent to the left will occur to compensate for the piling up of water on the right side.

Horizontal pressure gradients will be important in lakes where there are significant inflows of water with markedly different density from ambient lake density or where significant differential surface heating occurs.

Wind stress. Currents resulting from wind stress are the most common in lakes. Considerable research is still underway into the mechanism of transfer of wind momentum to water momentum. The stress on the lake is proportional to some power of wind speed, usually taken to be 2, although it evidently varies with wind speed, wave conditions, and atmospheric stability. In large, deep lakes, away from the boundaries, where wind-stress effects may be balanced by Coriolis-force effects, theory suggests that the surface current will move in a direction 45° to the right of the wind and that deeper currents are progressively weaker and directed farther to the right. The depth at which flow is opposite to the wind direction is effectively the depth below which there is no influence from the wind. This depth, designated *D*, can theoretically occur at about 100 metres (300 feet) in large, deep, midlatitude lakes. Observations show varying degrees of fidelity to theory because of complications from coastal effects and thermal stratification.

In coastal regions, if water depth is a significant fraction of or greater than *D*, winds blowing parallel to the shore will transport water either onshore or offshore. In the latter case, where the coast is to the left of the wind flow (Northern Hemisphere), the water driven offshore is replaced by cooler, deeper water (upwelling).

Internal waves and Langmuir circulation. Under stratified conditions a strong thermocline will essentially separate a lake into two layers. Shearing forces that develop between these layers cause a motion, termed internal waves, that may serve to directly dissipate a substantial portion of a lake's kinetic energy and act as a coupling between motion in the epilimnion and hypolimnion. A great range of periodicities is observed in the oscillations of the thermoclines, particularly in large lakes. Internal seiches, which are responsible for relatively long-period internal waves, are discussed later.

Effect of shearing forces

A small-scale circulation phenomenon that has aroused considerable attention on lakes is Langmuir circulation. On windy days, parallel "streaks" can be observed to develop on the water surface and exhibit continuity for some distance. These streaks may be caused by convergence zones where surface froth and debris collect. Langmuir circulation thus appears to be a relatively organized mixing mechanism wherein sinking occurs at the streaks and upwelling occurs between the streaks. Under favourable circumstances, this appears to be a key process for mixing heat downward in lakes.

SURFACE WAVES

Wind blowing over a calm lake surface first produces an effect that may appear as a widely varying and fluctuating ruffling of the surface. The first wave motion to develop is relatively regular, consisting of small, uniformly developed waves called capillary waves. These are quite transient, dissipating rapidly if the wind dies away or developing to the more commonly observed and more persistent gravity waves.

Energy will be continually fed to the waves by the frictional drag of the air moving over the water and by the direct force of the wind on the upwind face of the waves. The latter effect occurs only while the waves move more slowly than the wind. Pressure differences at the air–water interface also contribute energy to surface waves. Energy losses occur due mainly to turbulence in the water and, to a smaller extent, to the effects of viscosity.

Growth and movement of waves

Waves will continue to grow as long as there is a net addition of energy to them. Their height will increase as a function of wind speed and duration and the distance over which it blows (fetch). Most lakes are so small that fetch considerations are unimportant. Studies in larger lakes, however, have shown that the height of the highest waves are related to the fetch. In these lakes, waves as high as several metres are common, although waves of about seven metres (23 feet) are the highest to be expected. Wave heights in a given portion of a lake may vary considerably, due to interactions that suppress some waves and amplify others. As waves develop, their lengths increase, even after their height has stopped increasing. The phenomenon of swell, commonly observed in the oceans, is not truly realized, even in the largest lakes.

Waves travel in the same direction as the wind that generated them and at right angles to their crests. If they meet a solid object rather than a sloping beach, much of their energy will be reflected. If they enter shallow water obliquely, they are refracted. Wave speed, for waves longer than four times the depth of the water, is approximately equal to the square root of the product of the depth and the gravitational acceleration. For waves in relatively deep water, the wave speed is proportional to the square root of the wavelength.

As wave height increases, the sharpening of the wave crest may result in instability and a breaking off of the crest, a process hastened by the wind. This results in the familiar whitecaps. Waves that run ashore break up in surf. The wave height first decreases slightly, then increases, and the speed decreases, and eventually the wave form disappears as it crumbles into breakers. These can be plunging forms, in which the top curls right over the forward face, or of the spilling type, in which the crest spills down the forward face. A particular wave may break several times before reaching shore.

SEICHES

Cause and characteristics. If a denivellation, or tilting of a lake's surface, occurs as a result of a persistent wind stress or atmospheric pressure gradient, the cessation of the external forcing mechanism will result in a flow of water to restore the lake level. The flow would be periodic and uniform with depth, except for the damping effects of the lake-bottom friction and internal turbulence. Because of this, each successive tilt of the lake surface in the opposite direction occurs at a level slightly less than the previous one. The oscillation proceeds, moving the water back and forth until damping levels the water or until wind and pressure effect another tilt. This process is seiching; the lake oscillation is a seiche. The basic seiche has a single node, but harmonics of the oscillation occur, with several nodes being possible.

Seiche periods

The period of the uninodal seiche can be estimated from a formula that equates it to twice the length in the direction of the tilt, divided by the square root of the product of the mean lake depth and the gravitational acceleration.

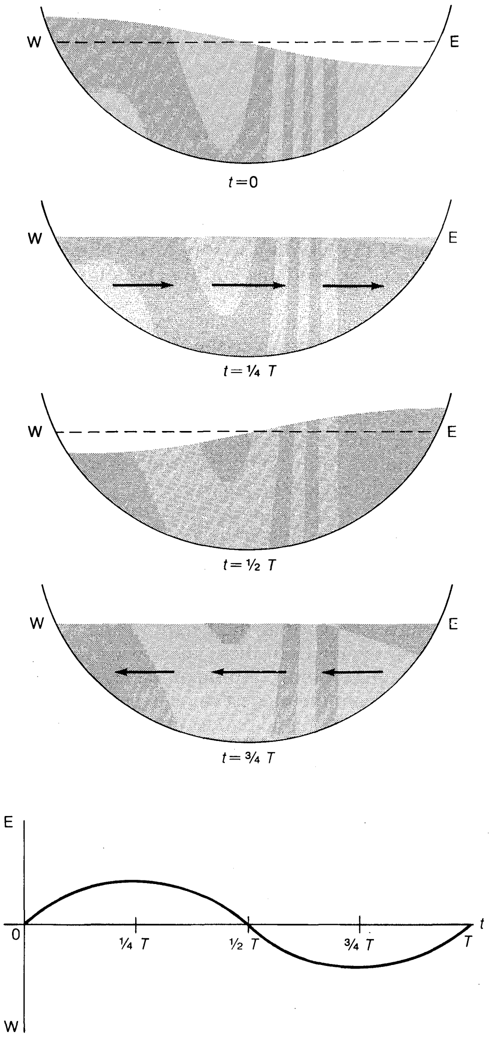


Figure 4: Four phases of elevations and currents. Below, time graph of current at centre of lake (*t* is time and *T* is seiche period; see text).

Seiches have been noted, recorded, and studied for hundreds of years. Lake Geneva, Switzerland, was one of the first lakes to be studied in connection with seiching; it has an observed uninodal period of about 74 minutes and a binodal period of about 35 minutes. The observed uninodal periods of Loch Treig and Loch Earn, Scotland; Lago di Garda, Italy; Lake Vetter, Sweden; and Lake Erie, North America, are approximately nine, 14.5, 43, 179, and 880 minutes, respectively.

Long, relatively narrow lakes that are exposed to a predominance of wind flow along their major axes are most likely to exhibit so-called longitudinal seiches. Transverse seiching can occur across the narrower dimension of a lake; that observed in Lake Geneva, for example, has a period of about 10 minutes.

The height of the denivellation depends upon the strength and duration of the forcing mechanism, as well as on the lake size and dimensions. In small lakes, level changes of a few centimetres are common, whereas, in the Great Lakes, intense storms can produce changes as great as two metres (seven feet). If the disturbance causing the tilting moves across the lake at close to the speed of the shallow-water wave speed, a profound amplification can occur, with possible disastrous consequences.

True tides that result from the gravitational effects of the Moon and Sun are rarely measurable in lakes, but small values of tidal components occasionally have been discerned.

Internal seiches. Internal seiching results from thermal stratification. The layers separated by the thermoclines oscillate relative to one another. Observed uninodal periods for Loch Earn, Lake Geneva, Lake Baikal, and Lake Cayuga (New York) are approximately 16, 96, 900 (binodal), and 65 hours, respectively.

Because hypolimnion water is very different from epilimnion water with regard to both thermal and biological characteristics, the massive movements of water and the turbulent exchanges that can occur during internal seiching are very important. Substantial portions of the bottom of shallow lakes can experience periodic alternation of exposure to hypolimnetic and epilimnetic water, and hypolimnetic water can be periodically exposed to the surface.

EFFECTS OF WAVE AND CURRENT ACTION

Shore erosion and coastal features. In a lake's early stages of existence, its shore is most susceptible to changes from wave and current action. As these changes occur, there is a tendency over time to an equilibrium condition—a balance between form and processes that depends upon the nature of the materials present (*e.g.*, the size of sand and gravel present). The effectiveness of waves in the erosion process depends in part upon the depth and slope of the lake bottom. Where the shore consists of a sheer cliff adjacent to deep water, wave energy will be reflected away without much erosional effect. The refraction of waves in zones of irregular coastline tends to concentrate wave energy at some locations and dilute it in others. Thus, features extended out into the lake will receive more wave energy, and the tendency is to smooth out an irregular coastline. Other net effects of shore erosion are an increase in the surface area of a lake and a reduction in its mean depth.

Changes in the coastline

As erosion takes place, the distribution of erosion products results in transport of finer material offshore. The resulting terrace is called the beach in its above-water manifestation and the littoral shelf where it is below water. Landward, beyond the beach, a wave-cut cliff is usually found. The steeper slope that often separates the littoral shelf from the benthos (bottom) zone in the central part of the lake is called the step-off by some limnologists.

Water movement directed at an angle to the coastline will result in the generation of currents along the shore. Erosion products will then be transported down the coast and may be deposited in locations where transport energy is dissipated due to movement around a bend or past an obstruction. A buildup of such material is called a spit. If a bay becomes completely enclosed in this way, the spit is called a bar.

Water in very shallow lakes that are subjected to strong winds may be piled against the lee shore to such an extent that countercurrents will develop from along the lee shore around each side of the lake. The cutting effects of these currents are known as end-current erosion and may characteristically alter the shape of a lake frequently subjected to winds from a particular direction.

Bottom morphology. The bottom morphology of a lake can be greatly influenced by deposition of sediment carried by inflowing rivers and streams. Although this process can be modified by wave and current action, most lakes are sufficiently quiet to permit the formation of substantial deltas. In very old lake basins the relief may become so extensively decreased due to the great buildup of deltaic deposits and the long-term effects of river widening, that deposition on the outer portions of a delta will fail to balance the effects of wave erosion. A delta, in these circumstances, will begin to shrink in size (see further RIVERS: *River deltas*).

Problems with artificial lakes

It is very important to understand lake processes that affect the basin morphology and to be able to predict their trends and their impact on human activities. Increasingly, man is imposing his ability to change natural events in lakes, and he has often encountered problems by not anticipating a lake's reaction to his projects. The actual creation of a lake by damming a river is a major undertaking of this type. One fairly recent example is Lake Diefenbaker, in Saskatchewan. In this region of prairie farmland, the banks of the new lake are extremely vulnerable to erosion, and planners have had to contend with the consequences of bank cutting and infilling of the basin. There are many examples of lesser engineering undertakings that have had to face the consequences of a lake's reaction. The building of jetties or breakwaters, for example, may interfere with natural circulation features. In some cases this has resulted in the reduction of flow past a harbour and increase in flow past a previously stable shoreline, with the result that the harbour has filled in or been blocked by sediment deposition, while the stable shoreline has become badly eroded.

The hydrologic balance of the lakes

THE WATER BUDGET

The role of lakes within the global hydrologic cycle has been described earlier. Lakes depend for their very existence upon a balance between their many sources of water and the losses that they experience. This so-called water budget of lakes is important enough to have warranted considerable study throughout the world, with each lake or lake system possessing its own hydrologic idiosyncracies. Aside from being of scientific interest, water-budget studies serve to reveal the dependence of each lake on particular hydrologic factors, thus enabling better management practices. These may include restrictions on water utilization during drought conditions, dike construction and evacuations prior to flooding, control of water levels to ensure efficient power production, and major decisions associated with diversions of watercourses in order to enhance water-quantity- and water-quality-management activities.

Importance of water-budget studies

Often, man is able to react to predicted imbalances in the hydrologic budget, although he is usually unable to influence the basic natural factors that cause the imbalances. Precipitation and evaporation, for the most part, are uncontrollable, although some advances have been made in evaporation suppression from small lakes through the use of monomolecular surface films. Groundwater flow is not controllable, except where highly restricted flow can be tapped. Rivers and streams, however, can be subjected to regulation by well-established practices through the use of dams, storage reservoirs, and diversions. It is mainly through these controls that efforts are made to make the most efficient usage of water as a resource.

When man takes steps to alter elements of a basin's water budget, careful consideration must be given to the consequences of the hydrology and ecology of the entire watershed. Dredging operations for the purpose of harbour clearance or improvements to a navigable channel, for example, may increase the outflow from an upstream lake, increase shore erosion, or regenerate undesirable sedimentary constituents into the lake or river water. The damming of a river or a lake outlet to increase local water storage may also result in undesirable effects, such as an increased evaporation from the larger surface area, the restriction of fish movement, or changes in the thermal climate of the downstream flow. Diversions and dam-site construction may also result in flooding of important bird-breeding areas or a lowering of other lakes in the system, resulting in undesirable consequences.

Water input. The usual major input of water to a lake derives from streams and rivers, precipitation, and groundwater. In some cases inflow may come directly from glacier melt. The relative importance of each of the major sources varies from lake to lake.

Stream and river flow are usually seasonally variable, depending upon precipitation cycles and snowmelt. At low altitudes some rivers exhibit a peak during a high precipitation period in winter and then a second peak associated with a subsequent spring snowmelt that feeds the nearby high-altitude tributaries. In regions where precipitation can occur in great quantities at high rates, streams swell quickly and water is delivered in relatively large volumes to downstream lakes.

A great deal of work has been done to improve the ability to measure and record streamflow. Consequently, it is usually the most accurately known of the inflow terms in the water budget. Most frequently, the height of the river level (stage) correlates well with the water discharge. In other cases, direct river-flow measurements are taken periodically with flow meters.

Precipitation over lake basins

Precipitation reaching a lake's surface directly may be the major input; this is true of Lake Victoria, Kenya–Tanganyika. In other cases, where the lake basin is large with well-developed drainage to a deep lake of small surface area, precipitation may be a small component. Precipitation that falls elsewhere in the lake basin may reach the lake through either surface or groundwater flow, or it may be lost due to evapotranspiration.

Measurements or estimates of precipitation for a basin are difficult to achieve. Even where elaborate networks of rain gauges exist or where these are supplemented by meteorological radar installations, total basin-precipitation data are still considered to be poor. Measurements of direct precipitation over lakes are exceedingly rare; this situation is especially serious in the case of a large lake for which nearby land data are not necessarily representative of conditions over the lake. Each climatic region throughout the world has its typical precipitation pattern, and the lakes within the regions are affected accordingly.

Groundwater controls on lake level

Groundwater reaches lakes either through general seepage or through fissures (springs). Groundwater is taken to be water in that zone of saturation that has as its surface the water table. The depth of the water table can be determined by digging a well into the saturated zone and noting the level of water—unless the water is under pressure, in which case it will rise in the well to a level above the water table. Clearly, it is possible for a lake level to coincide with the water table. In fact, unless impermeable material intervenes, the water table will drop to, rise to, or lie level with a lake surface. Groundwater that is lost from the saturated zone to a lake is termed groundwater discharge. Groundwater introduced to the saturated zone from a lake is termed recharge. The rate at which groundwater is exchanged between a lake and the saturated zone depends mainly upon the level of the water table and the pressure conditions within the saturated zone.

In permeable materials the zone above the water table is called the zone of aeration, and water within it is called soil moisture. Soil moisture is classified into three types: hygroscopic water adsorbed on the surface of soil particles; water held by surface tension in capillary spaces in the soil and moving in response to capillary forces; and water that drains through the soil under gravitational influence. The latter will most significantly contribute to groundwater recharge and to the water balance of a lake. The second category will generally be subject to loss due to transpiration by plants.

Water output. Lakes that have no outlets, either above or below surface, are termed closed lakes, whereas those from which water is lost through surface or groundwater flows are called open lakes. Closed lakes, therefore, lose water only through evaporation. In these cases, the loss of water that is less saline than the source water results in an increasing lake salinity.

Evaporation losses from lake surfaces

The process of evaporation results from a vertical gradient of vapour pressure over the water surface. Next to the water surface, saturation conditions exist that are a function of the temperature at the interface. The vapour pressure in the air above the surface is calculated from the temperature of the air and the wet-bulb temperature. The rate at which evaporation occurs also depends upon the factors that affect the removal of the saturated air above the surface (for example, wind speed and thermal convection).

Studies of evaporation must surely constitute a sizable proportion of all hydrological and oceanographic work. The principal categories of evaporation studies are water budget, energy budget, bulk aerodynamic techniques, and direct measurements of vapour flux (see further HYDROSPHERE).

The so-called aerodynamic technique is based upon Dalton's formula, which correlates evaporation with the product of the vapour pressure gradient and the wind speed. Studies during the past 20 years have produced a host of variations of this equation, determined empirically using independent measurements of evaporation. One of the most often used of these was developed in a study of Lake Hefner, and even this work has been subsequently modified to suit other climates and conditions. Few workers are satisfied with the present state of the art in the use of the aerodynamic equations. Nevertheless, once an equation of this type is satisfactorily developed for a particular lake, having been checked with independent methods, it is attractive because it usually employs data that can be routinely observed.

The direct measurement of vapour fluxes is an extremely intricate proposition, as motions over a water surface are usually turbulent, and instruments capable of measuring rapidly changing vertical motions and humidities are required. Not the least of the difficulties is the likelihood that the kind of turbulence over large bodies far from land is significantly different from that over land. Recent advances in theoretical developments and instrumentation continue to encourage this type of study. In turn, successes in this field offer the opportunity for the refinement of empirical techniques more practically suited for general lake investigators.

In many lake studies, data from evaporation pans have been used to determine lake evaporation. Pans have even been developed for flotation on lakes. Pans cannot truly simulate lakes, however, as they constitute a different type of system (they are not exposed to the atmosphere in the same way, they exchange heat through their sides, and they do not store heat in the same way as lakes).

Some examples of evaporation estimates include annual totals of between 60 and 90 centimetres (two and three feet) for Lake Ontario (using different techniques and for different years); about 75 centimetres (2.5 feet) for Lake Mendota, Wisconsin; over 210 centimetres (seven feet) for Lake Mead, Arizona and Nevada; about 140 centimetres (4.5 feet) for Lake Hefner; about 660 millimetres (26 inches) for the IJsselmeer, in The Netherlands; and about 109 millimetres (4.25 inches) for Lake Baikal.

Water output from a lake in the form of surface-water outflow generally depends upon the lake level and the capacity of the effluent channel. Although lakes often have many surface inflows or at least several incoming streams or rivers, they generally have but one surface effluent.

Water-level fluctuations. The net water balance for a particular lake will vary according to the periodic and nonperiodic variations of the inputs and outputs and is reflected in the fluctuations of the lake level. Because the prime influencing factors are meteorological, the periodicity of seasonal events are often seen in water-level records.

Seasonal variations of level

Lake-level rises generally coincide with or closely follow seasons of high precipitation, and falls of level generally coincide with seasons of high evaporation. Complications are introduced by a variety of factors, however. The storage of heavy winter precipitation as snowpack is one example. The release of this water during the spring thaw may also be hampered by the presence of river ice, resulting in late-spring or summer peaks. In large drainage basins the full effects of heavy precipitation may not be immediately realized in the lake-water balance because of the time required for basin drainage. Where glacier melt is a major input to a lake, the changes in level respond to seasonal heating as well as seasonal precipitation.

Although artificial controls, in the form of diversions, river dredging, and dams, affect the levels of the Great Lakes, the latter provide good examples of seasonal variations because of the lengthy record of levels available. The rivers draining to these large lakes are relatively stable; that is, the ratio of maximum to minimum flow is about 2 or 3 to 1, compared to 30 to 1 for the Mississippi River and 35 to 1 for the Columbia River. A 67-year average of lake levels by month shows that high water occurs, on the average, in September for Lake Superior and in June for Lake Ontario. Lows occur in March and December–January, respectively. The mean range in seasonal levels, for this period, is about 30 centimetres (one foot) for Lake Superior and about 45 centimetres (1.5 feet) for Lake Ontario. The pattern varies considerably from year to year, however, and periods of exceptional precipitation and drought are shown in the records. These events ultimately affect the downstream lakes, but, because of their relatively small discharge volumes, it takes 3.5 years for 60 percent of the full effect of a supply change to Lake Huron–Michigan to appear in the outflow from Lake Ontario.

The seasonal changes in a lake's level may be superimposed on longer term trends, which in some cases dominate. Several of the large lakes of the world have lengthy water-level records that illustrate long-term periods of relative abundance of water and drought. In Central Africa, Lakes Victoria, Albert, Tanganyika, and Nyasa exhibit substantial long-term features, some of which are consistent, suggesting that a common climatological factor is responsible for their existence (see further CLIMATE AND WEATHER). Nevertheless, others of these features are not consistent within the lakes and have not been adequately explained.

The principal climatological factors that would most affect long-term lake-level variations have not been recorded for long periods at many locations. Regular precipitation observations were not made before about 1850. Some useful evidence is found in such natural records as tree rings and peat-bog stratigraphy.

On a worldwide basis, there is evidence of a period of low levels in the middle 19th century and near the end of the first quarter of the 20th century. Lake George, in Australia, the Caspian Sea, several lakes in western North America, and Pangong Lake, in Tibet, are examples that have exhibited these features.

LAKE EXTINCTION

Possible causes of lake extinction

The life history of a lake may take place over just a few days, in the case of one formed by a beaver dam, or, for the largest lakes, it may cover geological time periods. A lake may come to its end physically through loss of its water or through infilling by sediments and other materials. Reference has previously been made to the chemical-biological death of a lake, which is not necessarily the end of it as a physical entity but may in fact be its termination as a desirable body of water.

Geological processes involving the uplift and subsequent erosion of mountains and the advance and retreat of glaciers, establish lake basins and then proceed to destroy them through infilling. Lake basins may also lose their water through drought or through changes in the drainage pattern that result in depletion of water inflows or enhancement of outflows.

The chemical-biological changes within a lake's history offer a fine example of ecological succession. In the early stages a lake contains little organic material and has a poorly developed littoral zone. Particularly in temperate zones, such conditions favour a plentiful oxygen content, and the lake is said to be oligotrophic. As erosion progresses and as lake enrichment and organic content increase, the lake may become sufficiently productive to place an excessive demand upon the oxygen content. When periods of oxygen depletion occur, a lake is said to be eutrophic. An intermediate stage in this course of events is called mesotrophy. In the case of oligotrophy the vertical oxygen distribution is essentially uniform, or orthograde. Under eutrophic conditions, oxygen values decrease with depth, and the vertical distribution is called clinograde.

The limits of oligotrophic and eutrophic conditions have been set in terms of the rate at which oxygen is depleted from the hypolimnion. These limits are arbitrary but are approximately 0.03 and 0.05 milligrams per square centimetre per day as the upper limit of oligotrophy and the lower limit of eutrophy, respectively.

As eutrophic conditions develop, bottom sediments become enriched in organic material, and bottom plants spread throughout the littoral zone. As infilling proceeds, the plant-choked littoral zone spreads lakeward. Eventually, the littoral zone becomes a marsh, and the central part of the lake diminishes to a pond. When the lake finally ceases to exist, terrestrial vegetation may flourish, even to the extent of forestation.

(R.K.L.)

BIBLIOGRAPHY. G.E. HUTCHINSON, *A Treatise on Limnology,* vol. 1, *Geography, Physics, and Chemistry* (1957), contains a comprehensive treatment of the physical, and chemical, aspects of lakes and includes an excellent bibliography of previous work. Less fundamental but more recent summaries of scientific lake studies include GERALD A. COLE, *Textbook of Limnology,* 3rd ed. (1983); CHARLES GOLDMAN and ALEXANDER HORNE, *Limnology* (1982); and ROBERT G. WETZEL, *Limnology,* 2nd ed. (1983). For the hydrologic and geologic aspects of lakes, the interested reader should see: J. PROUDMAN, *Dynamical Oceanography* (1953); B. KINSMAN, *Wind Waves: Their Generation and Propagation on the Ocean Surface* (1965); R.L. WIEGEL, *Oceanographical Engineering* (1964); and S.N. DAVIS and R.J.M. DEWEIST, *Hydrogeology* (1966). The subject of water use and planning is covered in R.J. CHORLEY (ed.), *Water, Earth and Man* (1969); J.G. NELSON and M.J. CHAMBERS (eds.), *Water* (1969); W.R.D. LEWELL and B.T. BOWER, *Forecasting the Demands for Water* (1968); UNITED STATES WATER RESOURCES COUNCIL, *The Nation's Water Resources: The First National Assessment* (1968); and F.E. MOSS, *The Water Crisis* (1967). See also W.R. EDMONDSON, *Fresh-Water Biology,* 2nd ed. (1959). Journals of particular interest include: *Limnology and Oceanography* (bimonthly), *Proceedings* of the International Association for Great Lakes Research, and several oceanographic and meteorological journals.

(R.K.L./Ed.)

Lamp Shells: Phylum Brachiopoda

The Brachiopoda are a phylum of marine invertebrates that are covered by two valves, or shells; one valve covers the dorsal, or top, side; the other covers the ventral, or bottom, side. The valves, of unequal size, are bilaterally symmetrical; *i.e.,* the right and left sides are mirror images of one another. Brachiopods (from the Greek words meaning "arm" and "foot") are commonly known as lamp shells because they resemble early Roman oil lamps.

Brachiopods occur in all oceans. Although no longer numerous, they were once one of the most abundant forms of life.

Members of this phylum first appeared rather early in zoological history. It is possible, by means of fossil representatives, to survey their evolution from the Cambrian Period (about 570,000,000 years ago) to the present. Although some of the evolutionary development is revealed, it is still imperfectly understood. Other than their usefulness in dating geological periods, members of this phylum have no economic value, except as curios and museum pieces.

This article is divided into the following sections:

GENERAL FEATURES

Size range and diversity of structure. Most brachiopods are small, 2.5 centimetres (about one inch) or less in length or width; some are minute, measuring one millimetre (more than 1/30 of an inch) or slightly more; some fossil forms are relative giants—about 38 centimetres (15 inches) wide. The largest modern brachiopod is about 10 centimetres (four inches) in length.

Great diversity existed among brachiopods in the past; modern brachiopods, however, exhibit little variety. They are commonly tongue-shaped and oval lengthwise and in cross section. The surface may be smooth, spiny, covered with platelike structures, or ridged. Most modern brachiopods are yellowish or white, but some have red stripes or spots; others are pink, brown, or dark gray. The tongue-shaped shells (*Lingula*) are brown with dark-green splotches; rarely, they are cream yellow and green.

Distribution and abundance. Today, brachiopods, numbering about 300 species representing 80 genera, are abundant only locally. In parts of the Antarctic they outnumber all other large invertebrates. They are common in the waters around Japan, southern Australia, and New Zealand. Although rare in the Indian Ocean, some unusual types are common along the coast of South Africa. In Caribbean and West Indian waters, 12 species occur. The east and west coasts of the North Atlantic Ocean are sparsely occupied by brachiopods; the waters around the British Isles contain a few species, and a few genera live in the Mediterranean Sea. The West Coast of the United States and Hawaii have a number of brachiopod species, and the coasts of Chile and Argentina have a considerable variety, including the largest living species. Some live in the polar regions, and a few are abyssal; *i.e.,* they inhabit deep parts of the ocean.

NATURAL HISTORY

Reproduction. Not much is known about the reproduction of brachiopods. Except in three genera, the sexes are separate. Eggs and sperm are discharged into the mantle cavity through funnel-shaped nephridia, or excretory organs, on each side of the mouth. Fertilization takes place outside the shell. In a few genera the young develop inside the female in brood pouches formed by a fold of the mantle, a soft extension of the body wall. Some fossil forms had internal cavities that may have served as brood chambers. The egg develops into a free-swimming larva that settles to the bottom. The free-swimming stage of the articulate brachiopods (whose valves articulate by means of teeth and sockets) lasts only a few days, but that of the inarticulates may last a month or six weeks. In inarticulate larvae the pedicle, a stalklike organ, develops from a so-called mantle fold along the valve margin; in articulates it develops from the caudal, or hind, region.

Articulate and inarticulate forms

Behaviour and ecology. About 60 percent of brachiopods live in shallow water (less than 100 fathoms—about 180 metres [600 feet]) on the shelf areas around the continents. More than 35 percent occupy waters deeper than 100 fathoms, and a few live in the abyss down to more than 6,000 metres (about 20,000 feet). *Lingula* lives from the tidal zone to 23 fathoms (about 42 metres [138 feet]). Most modern branchiopods anchor by the pedicle to pebbles, to the undersides of stones, or to other hard objects. They prefer quiet water and protected surroundings. *Lingula* lives in mud or sand and is attached at the bottom of its burrow.

Brachiopods feed by opening the shell and bringing in food-bearing currents by lashing of the cilia (hairlike structures) attached to the filaments of the lophophore, a horseshoe-shaped organ that filters food particles from the seawater. Cilia in lophophore grooves bring food particles, often trapped in mucus, to the mouth. Brachiopods feed on minute organisms or organic particles. Articulate brachiopods, which have a blind intestine, may depend partly on dissolved nutrients.

Shells of some articulate brachiopods have a fold, which forms a trilobed anterior that helps keep lateral, incoming food-bearing currents separated from outgoing, waste-bearing currents. When feeding, *Lingula* protrudes its an-

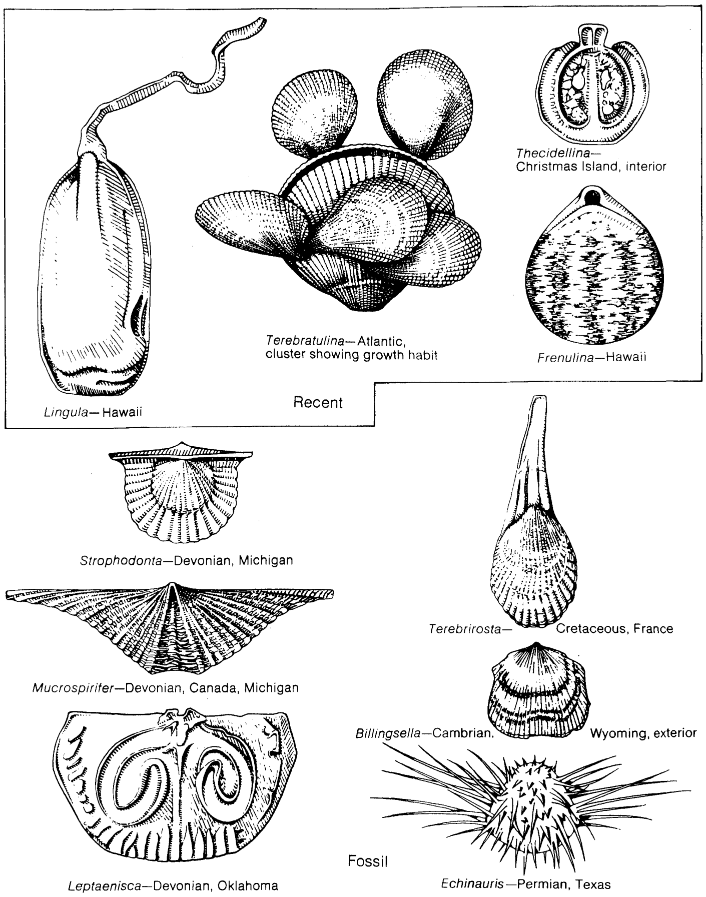

Lingula—Hawaii
Terebratulina—Atlantic, cluster showing growth habit
Thecidellina—Christmas Island, interior
Frenulina—Hawaii
Recent
Strophodonta—Devonian, Michigan
Terebrirosta—Cretaceous, France
Mucrospirifer—Devonian, Canada, Michigan
Billingsella—Cambrian, Wyoming, exterior
Fossil
Leptaenisca—Devonian, Oklahoma
Echinauris—Permian, Texas

Figure 1: Representative brachiopods, fossil and living.
By courtesy of the United States National Museum

terior (front) end above the mud and arranges its setae (bristle-like structures) into three tubes. These channel the water into lateral incoming and medial, or central, outgoing currents. Some coralliform brachiopods of the Permian Period (280,000,000 to 225,000,000 years ago) are thought to have fed by rapid beating of the dorsal valve, causing a sucking in and expulsion of food-bearing water. Some ostreiform (oyster-shaped) types of the same period are believed to have fed by gentle pulsation of the dorsal valve.

FORM AND FUNCTION

Two major groups of brachiopods are recognized, based on the presence or absence of articulation of the valves by teeth and sockets. The valves of inarticulate brachiopods are held together by muscles. *Lingula,* with its elongated, tonguelike shell, is an example. Its convex valves bulge outward at the middle and taper posteriorly, or away from the hinge. A long, fleshy pedicle protrudes between the valves at the tapered end. The pedicle of *Lingula* differs from that of most other brachiopods in being flexible and capable of movement—an aid in burrowing and in attaching the animal in its burrow. The shell interior is divided into posterior coelomic (internal-body) and anterior mantle cavities. The internal organs are located in the coelom. The digestive system consists of mouth, gullet, stomach, intestine, and anus, all surrounded by a liver, or digestive gland. A complex set of muscles opens the valves and slides them laterally, or sideways, when feeding. The mantle cavity is occupied by the lophophore. *Lingula* lives in a burrow in mud or sand with the tip of its pedicle attached in mucus at the bottom of the burrow. The contractile pedicle permits extension of the shell when feeding or retraction if the animal is startled.

The cavities within the shell

The articulate-brachiopod shell is typified by *Waltonia,* which is small (about two centimetres [3/4 inch]) and red in colour, with a smooth or slightly ridged shell. This type of shell is more highly specialized than that of most inarticulate species and is composed of three layers. The outer layer, called periostracum, is made of organic substance and is seldom seen in fossils. A middle layer consists of calcium carbonate (calcite). The inner layer is composed of calcite fibres and may be punctate—*i.e.,* perforated by minute pits—or it may be pseudopunctate, with rods (taleolae) of calcite vertical to the surface. Impunctate shells have neither pits nor taleolae.

Many hinged brachiopods attach to the substrate, or surface, by a tough, fibrous pedicle; but some specialized forms are cemented to the substrate by the beak of the ventral valve. Cemented forms are commonly distorted, scalelike, or oyster shaped or resemble a cup coral. The pedicle of some brachiopods is atrophied; their shells lie loose on the sea floor.

The shell of an articulate brachiopod tapers posteriorly to a beak. The ventral valve is usually the larger. The hinge may be narrow or wide. Many hinged genera have a flat or curved shelf, called the palintrope, between the beak and the hinge line. The ventral palintrope is divided at the middle by the delthyrium, a triangular opening for the pedicle. The delthyrium may remain open or be wholly or partly closed by small plates growing from its margins. In some families the delthyrium is closed completely or partly by one plate, the pseudodeltidium, anchored to the delthyrial margins. The articulating teeth occur at the angles of the delthyrium and may or may not be supported by vertical dental plates, which may be separate or united to form a so-called spondylium. Teeth are of two types, deltidiodont and cyrtomatodont. Deltidiodont teeth grow anteriorly with the palintrope and leave a growth path along the delthyrial edge; cyrtomatodont teeth are knoblike and occur in shells without a hinge line. They grow anteriorly but are kept knoblike by posterior resorption.

Muscles, sockets, and related structures

The dorsal valve contains structures called crura that diverge from the beak. In some fossil forms the crural bases (brachiophores) bound a triangular cavity, the notothyrium, in which the diductor, or opening, muscles are attached onto the floor or to a ridge, or boss, called the cardinal process at the apex. The notothyrium may be closed by a solid plate, the chilidium. In more highly developed genera a hinge plate bearing the pedicle or dorsal adjustor muscles occurs between the crural bases. The hinge plate is said to be divided when it is incomplete but undivided when it forms a flat or concave structure. The hinge plate is often supported by a median septum, or wall. The hinge sockets are located between the inside shell wall and a socket ridge to which the hinge plate is

From *Invertebrate Zoology* by Paul A. Meglitsch, Copyright © 1967 byOxford University Press, Inc. Reprinted by permission

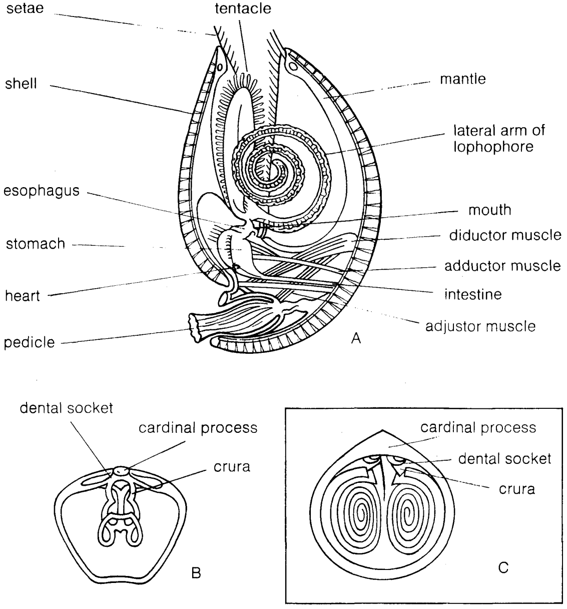


Figure 2: *Body plans of Brachiopoda.*
(A) *Magellania,* side view. (B) *Waltonia,* top view of brachial valve. (C) *Atrypa,* interior view of brachial valve.

attached. In many specialized genera the crura support calcareous loops or spires (brachidia), the inner skeleton of the lophophore. Structures corresponding in function but of different origin and with different names occur in the pedicle region of some inarticulate brachiopods.

The fleshy body of the articulate brachiopod is divided transversely by the body wall into a posterior visceral cavity filled with coelomic fluid and an anterior mantle cavity filled with seawater. The visceral cavity contains the U-shaped digestive canal, four reproductive glands, and a liver, or digestive gland, held in place by mesenteries (sheets of tissue). Extensions of the coelom into the mantle hold the eggs and sperm. The mouth leads into a saclike stomach that ends in an intestine; there is no anus. The liver surrounds the stomach. Waste is excreted through the mouth. The nervous system, which consists of two principal ganglia, or nerve centres, encircles the esophagus and sends branches to other parts of the body. One pair of excretory organs (nephridia) occurs in most brachiopods, but two pairs may be present.

The mantle cavity is lined by the thin, shell-secreting mantle that is fringed by setae at its edges. Within the mantle cavity is the lophophore, which may be a simple or complicated loop, often horseshoe-shaped. Ciliated filaments along the loop direct food-bearing currents to the mouth, which is located on the body wall between the branches of the lophophore and crura.

The shell opens by contraction of diductor muscles that extend from near the centre of the ventral valve to the process under the dorsal beak. These muscles pull the dorsal beak forward, rotating it on a line joining the hinge teeth. Contraction of the adductor muscles closes the valves; in the ventral valve these are located between the diductors. Pedicle muscles or adjustors extending from the pedicle to the hinge plate of the dorsal valve rotate the shell on the pedicle. Where the muscles are attached to the shell there are scars, which are helpful in the identification of genera.

PALEONTOLOGY

Early evolution

Brachiopods were among the first animals to appear at the beginning of the Cambrian Period (570,000,000 years ago). Their evolution and distribution was wide and rapid. More than 35,000 species in more than 2,500 genera are known, and the number of described species increases yearly. Articulate and inarticulate brachiopods appeared at the same time in a relatively advanced state of development, indicating a long evolution from forms without shells, an evolution apparently lost or unrecorded in Precambrian times.

The Inarticulata, the most abundant brachiopods of the Cambrian, soon gave way to the Articulata and declined greatly in number and variety toward the end of the Cambrian. They were represented in the Ordovician (500,000,000 to 430,000,000 years ago) but decreased thereafter. In the Cretaceous (136,000,000 to 65,000,000 years ago) the punctate calcareous Inarticulata proliferated, but this trend soon ended. The Inarticulata dwindled through the Cenozoic (65,000,000 years ago) to the Recent. Only nine genera are known during the Recent Epoch (last 10,000 years). Inarticulate genera represent about 6.5 percent of all brachiopod genera.

The Articulata, diverse and most numerous from Ordovician times to the present, were, in the Cambrian, represented by several specialized forms. Articulate evolution tended toward shell elaboration for bottom dwelling and perfection of feeding mechanisms from the simple looped lophophore to the elaborate lobate and spiral forms. The Orthida, the most common articulate brachiopods of the Cambrian and Ordovician, decreased in numbers after the Ordovician, and the impunctate Orthida became extinct in the Early Devonian (about 395,000,000 years ago); the punctate Orthida lingered into the Permian Period (280,000,000 to 225,000,000 years ago). The Strophomenida appeared in the Early Ordovician and increased rapidly. They were abundant and varied in the Devonian, becoming even more so by Permian times. This large order became greatly reduced at the end of the Permian Period. The Pentamerida, never prolific, flourished in the Ordovician; an evolutional burst of huge forms occurred in the Silurian (430,000,000 to 395,000,000 years ago), but after that the pentamerids decreased into the Devonian (395,000,000 to 345,000,000 years ago) and became extinct early in the late part of that period. The Spiriferida are conspicuous for the great elaboration of the spiral brachidium. They appeared in the Ordovician, were widely distributed into the Permian, and survived into the Jurassic, which began 190,000,000 years ago. The Rhynchonellida were abundant from mid-Ordovician throughout the Paleozoic. They survived into the Triassic (225,000,000 to 190,000,000 years ago) and had a rebirth in the Jurassic, after which they declined into the Cenozoic. They now number only 14 genera.

Development of modern dominant forms

The Terebratulida, now the dominant group, appeared in the early Devonian and rapidly expanded in the mid-Devonian to produce a number of gigantic forms; a few long-looped and short-looped genera persisted into the Permian. The Terebratulida survived the Permian and were widely distributed in the Triassic and evolved into a great variety of forms in the Jurassic, especially the short-looped types. Decline of the short-looped terebratulids began in the Late Cretaceous (75,000,000 years ago); they have continued to dwindle into the present and are now outnumbered by the long-looped terebratulids.

CLASSIFICATION

Distinguishing taxonomic features. Brachiopods possess a lophophore (a feeding structure that filters food from seawater), excretory organs (nephridia), and simple circulatory, nervous, and reproductive systems. Brachiopods have usually been divided into two classes, Articulata and Inarticulata.

Annotated classification. The classification below is based on that proposed by A. Williams and A.J. Rowell in 1965 in *Treatise on Invertebrate Paleontology.*

PHYLUM BRACHIOPODA (lamp shells)
Marine invertebrates with two valves, or shells; lophophore horseshoe-shaped; about 300 living species known; more than 30,000 extinct species described; occur in all oceans.

Class Inarticulata
Shell does not articulate, is usually composed of chitinophosphatic material; shell muscles complex; pedicle (stalk) develops from ventral mantle, a soft extension of the body wall; intestine with anal opening.

Order Lingulida
Shell usually contains phosphate, rarely calcareous, biconvex (*i.e.,* both valves convex), beak for attachment to surface apical, or located at the tip, in both valves; fleshy pedicle emerging between the valves at the tapered end; about 51 genera; Cambrian to Recent.

Order Acrotretida
Usually circular in outline; shell either contains phosphate or is punctate calcareous; pedicle opening confined to the ventral valve; 62 genera; Early Cambrian (550,000,000 years ago) to Recent.

Order Obolellida
Mostly calcareous, biconvex, shape nearly circular to elongated; position of pedicle opening variable; dorsal valve with marginal beak; 5 genera; Early to mid-Cambrian.

Order Paterinida
Shell with phosphate, rounded or elliptical; pedicle opening partly closed by cover called homeodeltidium; dorsal valve similar to the ventral but with a convex homeochilidium; 7 genera; Early Cambrian to mid-Ordovician (450,000,000 years ago).

Class Articulata
Shells articulate by means of teeth and sockets; shells always calcareous; musculature less complicated than in Inarticulata; larval pedicle develops from rear region; no outside opening from intestine.

Order Kutorginida
Calcareous, biconvex interarea (smooth surface in area between beak and hinge line) present; delthyrium (opening in the pedicle) closed by a plate, the pseudodeltidium; dorsal valve with interarea; muscle area narrow and elongated in both valves; 3 genera; Early to mid-Cambrian.

Order Orthida
Usually biconvex, wide-hinged, with interareas in both valves; teeth deltidiodont (leave a growth path along margin of pedicle opening); hinge structures consist of brachiophores (supporting

structures), shell substance punctate or impunctate—*i.e.*, with or without pits; more than 200 genera; Early Cambrian through Permian (225,000,000 years ago).

Order Strophomenida

Teeth deltidiodont when present; ventral muscles large; shell substance pseudopunctate (with rods of calcite), rarely impunctate; more than 400 genera; Early Ordovician (490,000,000 years ago) to Early Jurassic (180,000,000 years ago).

Order Pentamerida

Biconvex, ventral valve usually with a spondylium (united dental plates); delthyrium usually open; dorsal-valve brachiophores supported by bracing plates; impunctate; nearly 100 genera; mid-Cambrian to Late Devonian (350,000,000 years ago).

Order Rhynchonellida

Narrow-hinged with functional pedicle; dorsal valve with or without a median septum; lophophore (of Recent genera) dorsally spiral and attached to crura (supporting structures); spondylia rare; nearly 300 genera; Ordovician to Recent.

Order Spiriferida

Lophophore supported by a calcareous spiral structure (brachidium); punctate or impunctate, usually biconvex; delthyrium open or closed; more than 300 genera; mid-Ordovician to Jurassic (136,000,000 years ago).

Order Terebratulida

Pedicle functional, cyrtomatodont teeth; lophophore supported wholly or in part by a calcareous loop, short or long and free or attached to a median septum; more than 300 genera; Early Devonian to Recent.

Critical appraisal. The classes Articulata and Inarticulata were first proposed by T.H. Huxley in 1869. Before 1932 they were further subdivided into four orders based on the imperfectly known larval development and formation of the shell around the pedicle opening. In 1927 a fifth order was proposed, and it was suggested that a classification be based on the pedicle development of the larvae.

Most brachiopods are extinct, and larval development can only be conjectured. Because of this, the early classification schemes have been abandoned. Eleven orders distributed in Huxley's classes have been retained in the present classification, which is still being modified. On the basis of hinge and tooth types some systematists have divided the Articulata into two subclasses, Protremata and Telotremata. The Protremata are wide-hinged forms with deltidiodont teeth. The Telotremata are narrow-hinged brachiopods with cyrtomatodont teeth.

BIBLIOGRAPHY. M.J.S. RUDWICK, *Living and Fossil Brachiopods* (1970), a modern, readable, and comprehensive account of the brachiopods; R.C. MOORE (ed.) *Treatise on Invertebrate Paleontology,* pt. H, *Brachiopoda,* 2 vol. (1965), a technical, thorough account of brachiopods and their current classification, with extensive topical bibliographies.

(G.A.C.)

Land Reform and Tenure

The concept of land reform has varied over time according to the functions performed by land itself: as a factor of production, a store of value and wealth, a status symbol, or a source of social and political influence. Land value reflects its relative scarcity, which in a market economy usually depends on the ratio between the area of usable land and the size of the population. As the per capita land area declines, the relative value of land rises, and land becomes increasingly a source of conflict among economic and social groups in the community.

The patterns of wealth and income distribution and of social and political influence are partly determined by the laws governing land tenure. These laws specify the acceptable forms of tenure and the privileges and responsibilities that go with them. They define the land title and the extent to which the owner can freely dispose of it and of the income accruing from its use. In this sense, the form of tenure determines the wealth and income distribution based on the land: if private ownership is permitted, class differentiation is unavoidable; in contrast, public ownership eliminates such distinctions. The forms of tenure range from temporary, conditional holding to ownership in fee simple, which confers total unencumbered rights of control and disposal over the land.

This article is divided into the following sections:

The purposes of land reform

Historically, land reform meant reform of the tenure system or redistribution of the land ownership rights. In recent decades the concept has been broadened in recognition of the strategic role of land and agriculture in development. Land reform has therefore become synonymous with agrarian reform or a rapid improvement of the agrarian structure, which comprises the land tenure system, the pattern of cultivation and farm organization, the scale of farm operation, the terms of tenancy, and the institutions of rural credit, marketing, and education. It also deals with the state of technology, or with any combination of these factors, as shown by recent reform movements, regardless of the political or ideological orientation of the reformers.

OBJECTIVES OF REFORM

Reform is usually introduced by government initiative or in response to internal and external pressures, to resolve or prevent an economic, social, or political crisis. Reform, therefore, may be considered a problem-solving mechanism. The true motives for reform, however, may be quite different from those announced by the reformer.

The distinction between the real and proclaimed objectives may be especially significant if the proclaimed objectives have been forced upon reformers who do not lend their full support to those objectives. The reformers may proclaim certain objectives merely to appease the peasants, to undermine the opposition, to win international backing, or to safeguard their own positions. The proclaimed purposes of land reform, however, will be taken as a point of departure in this article.

Political and social objectives. The most common proclaimed objective of land reform is to abolish feudalism, which usually means overthrowing the landlord class and transferring its powers to the reforming elite or its surrogates. If "foreigners" happen to be among the landlord class, the objectives become the defeat of imperialism and the end of foreign exploitation.

Another common objective is to free the peasants from subjugation to and dependence on the exploiters and make them active citizens by restoring what assertedly had been taken away from them.

A third objective is to create democracy—a stated purpose of both capitalist and Communist reformers. Most capitalist reforms are based on the premise that individual private ownership in the form of independent family farms will promote and sustain democratic institutions.

Communist reformers, in contrast, usually aim at overthrowing both feudalism and capitalism on the premise that, as a means of production, private ownership of land inherently breeds exploitation. In practice, this means "returning land to the tillers" and creating a classless, democratic society. A more immediate and practical goal of Communist reformers has often been to rally the peasants in support of the new order and against the former regime.

Finally, reform may be introduced simply as the most expedient way to resolve a crisis or avoid a revolution. The reformer, in this case, will introduce and implement just enough reform to appease the peasants and contain the conflict. This happens especially when the reformers are still in sympathy with the landlord class and consciously prefer a moderate rather than a radical reform.

These political objectives tend to undergo change during the period of implementation and are, therefore, kept vague enough to permit flexibility and modification as conditions change.

All land reforms emphasize the need to improve the peasants' social conditions and status, to alleviate poverty, and to redistribute income and wealth in their favour. They try to create employment opportunities and education and health services and to redistribute the benefits to the community at large, the younger generation as the main target.

Economic objectives. Economic development has become a major objective of governments and political parties in recent decades. Efforts have been made to encourage agricultural progress by means of agrarian reform in favour of the peasant who does not own his land or whose share of the crop is relatively small, and who therefore has little incentive to invest capital or expend effort to improve the land and raise productivity. Another mechanism has been to encourage labour-intensive cultivation, on the assumption that traditional or feudal landowners often use their land extensively and wastefully.

An equally important economic objective is to promote optimum-scale farming operations. Excessively large farms (*latifundia*) and excessively small farms (*minifundia*) tend to be inefficient. Therefore, reform aims at creating farms of optimum size given the land quality, the crop, and the level of technology.

Finally, reform aims at coordinating agriculture with the rest of the economy. In their quest for economic development and industrialization, reformers attempt to make the rural sector more responsive to the needs of the industrial sector for labour, food, industrial raw materials, capital, and foreign currency. These functions are often expected to be performed simultaneously.

TYPES OF REFORM

Whether it is called land reform or agrarian reform, the operational concept covers five main types of reform, classified according to whether they deal with land title and terms of holding, land distribution, the scale of operation, the pattern of cultivation, or supplementary measures such as credit, marketing, or extension services.

Title settlement and terms of holding land

Reforms concerned with the title to land and the terms of holding reflect a transition from tradition-bound to formal and contractual systems of landholding. Their implementation involves property surveys, recording of titles, and provisions to free the landholder from restrictions or obligations imposed by tradition. Property surveys are conducted wherever land is held by a tribe or clan or where reallocation of cultivable land routinely follows tradition. In these situations the landholder may lack the incentive to improve the land because the right of disposal belongs to the tribe, clan, or feudal lord, as in medieval Europe and in parts of present-day Africa and the South Pacific islands. Such reform affects landholding in at least three ways: it may increase security of tenure and hence incentives; it may reorganize the system of inheritance in favour of offspring; and it may bring land onto the market so that land transactions become possible. This reform, however, has little immediate effect on the scale of operation, but it does facilitate future land concentration and fragmentation. In countries where the terms of holding and tenancy are regulated by tradition, reform may seek to convert tenancy into a contractual agreement that offers some protection to the tenant and more security and incentive to improve the land and advance technology, as in Japan, India, and Pakistan.

The redistribution of land

The most common type of reform involves the redistribution of land titles from one individual to another, from individuals to a group or community at large, or from a group to individuals. The land of one landlord may be redistributed to many individuals, as in Egypt, Iran, or Ireland. Or the land of individuals may be reallocated in favour of the community at large by abolishing private ownership, as in the Soviet Union and China. Or, again, public land may be distributed to individuals, as in various parts of Latin America.

The impact of redistribution on the scale of operations and on marketability of the land depends on the form it takes and the restrictions attached to it. If the redistributed farm was previously operated as a unit, its division means fragmentation and reduction of scale; however, if it was operated in fragments by tenants, transfer of title to the tenants would not affect the scale. The final results depend on the measures taken to prevent adverse effects.

Efforts to improve the scale of operation

Land-tenure reform, of course, can improve the scale of operations by enlarging the farm or by reducing it. Enlargement applies when the holding is increased in size, either by adding to it or by consolidating its fragmented parts. Farm consolidation involves reallocation of the total farmland within a region by land exchange, sale, or lease such that no one loses and all gain by increasing efficiency. The scale of operations may be increased by pooling resources, as in farm cooperatives and collectives that offer facilities otherwise inaccessible to a small farm.

An equally common approach is to divide large, extensively cultivated farms into smaller and more intensively cultivable units. Reduction of the scale, however, has potential problems since it may result in excessively small units or in the breakup of efficiently run farms. Operations below the optimum level may inhibit improvements in technology, capital investment, and diversification.

Changing the pattern of cultivation

Changes in the pattern of cultivation relate directly to cultivation, land yield, and labour productivity. While other types of reform may influence productivity indirectly by enhancing security of tenure and the scale of operation, improvements in the pattern of cultivation affect productivity directly, through advances in technology, improved irrigation, and the application of chemicals.

Technological advance usually implies mechanization, although it may be biological and organizational only, as in crop rotation, reconditioning of the soil, improved seeding, or better utilization of available technology. The state of technology determines the level of productivity or the ratio between outputs and inputs. More advanced technology permits the cultivation of more land per unit of labour, deeper plowing, better timing of farm operations, reclamation of areas previously inaccessible, and possibly wider diversification of the crops than previously attainable. By easing the physical burden of farm work, it helps to conserve human energy. Improved technology may also be the most direct way to modify tradition without an open confrontation or a political revolution. Mechanization and advanced technology may, of course, cause displacement of labour, unemployment, or the absorption of capital at the expense of other sectors, at least in the short run; in the long run, the positive effects tend to prevail.

Improvements in irrigation include increasing the water supply, draining swampy land, and regulating the quantity and quality of water flow. Irrigation is especially important in that it involves large investments and infringes on tenure rights, both matters that invite public responsibility and intervention. Irrigation and technology are closely related to the use of fertilizer and other chemicals. Chemicals may be difficult to apply without irrigation, and neither may be practical unless farming technology has advanced beyond relatively primitive methods. Improvement of the pattern of cultivation may be inhibited, however, by traditional attitudes, the lack of skills, or the scarcity of capital. Another difficulty is that changes in the pattern of cultivation are usually long-term investments that may be too slow to satisfy immediate pressures for reform.

Measures that supplement land reform

Many improvements and changes may have to be implemented in areas outside the immediate sphere of agriculture, such as credit, marketing, and education. Unless the farmer is able and ready to take advantage of new opportunities and his product can be marketed profitably, reform efforts may be futile. Costly or inaccessible credit and the excessive charges of middlemen increase the relative costs of farming. Therefore, supervised credit, subsidies, and low-interest loans that help to replace traditional sources of credit have been common, and credit and marketing cooperatives and market regulation have been used to protect farmers against exploitation by middlemen.

Finally, improvements in general education are essential in any reform that involves the modernization of agriculture. Extension services, literacy promotion programs, the teaching of home economics, and vocational training are of special importance in helping the young and unemployed and in providing skilled labour for industry.

EVALUATION AND CRITERIA OF SUCCESS

The difficulties of appraising agricultural reforms

Agrarian reform is a complex process of directed change, and its effects touch society in many ways. Therefore its evaluation may be difficult because the various social, political, and economic objectives may be inconsistent with each other. Even champions of reform and planners may have different ideas about it. Moreover, there are no generally accepted criteria for determining the success of such a program, nor adequate tools for measuring its progress.

Economic criteria. Economic indicators and criteria are basically the requisites of economic development. Economic development may be defined as a sustained increase in and achievement of a given level of per capita real income. To be sustained, the rise of per capita real income must be accompanied by changes in the economic and social structures of society, increase in total investment (capital formation), higher productivity, and full employment.

Capital formation in agriculture implies that more resources will be put at the disposal of the farmer in the form of machinery, fertilizers, and irrigation and drainage facilities, all of which contribute to the productive capacity and productivity of the land and the worker. Because capital formation partly depends on domestic saving, a higher rate of capital formation may thus be an indicator of the success of the reform in aiding economic development.

Another indicator is change in land yield or labour productivity. A rise in yield or productivity implies higher efficiency, better use of resources, and an advance in the state of technology. It also suggests an increase in the level of income and potential saving and investment. In fact, the increase in productivity may be the most important single indicator of the contribution made by reform to economic development. Change in the level of rural employment or unemployment provides another indicator, which should be reflected in the level and distribution of income.

Finally, an important indicator may be the change in agriculture's responsiveness to the demands of industry and manufacturing. The ability of agriculture to provide labour, food, industrial raw materials, and a market for industrial products is a significant measure of its contribution to industrialization and development.

Social criteria. Social and political accomplishments are more difficult to measure. One of the important indicators may be the degree of peasant participation in activities such as voting, representation, and decision making. Social and political stability, or the tendency to change governments by constitutional and nonviolent means and continuity of the social and political order without resort to force, are other indicators.

But these various indicators can only suggest that change has taken place. The results depend on the magnitude and relevance of the change. Assuming that measurement is feasible, three approaches to evaluation may be followed: the goal achievement approach, the perceived achievement approach, and the closing-the-gap (integrative) approach.

Goal achievement considers a program successful to the extent that it realizes the goals specified prior to the reform. Probably the most common economic goal is maximization, according to which efficiency dictates that reform should continue up to the point where the marginal benefit of reform is equal to its marginal cost. Land distribution in this case will continue up to the point where its net benefit is zero. Another common goal is to realize incremental gains as the reform proceeds. The reform would be successful to the extent that the effects have been positive. A 20-percent increase in capital formation in a capital-poor economy, however, may be more significant than a 20-percent increase in a capital-rich economy. Similarly, a 20-percent decline in the number of dissatisfied peasants in a peaceful or democratic society may be more conducive to stability and harmony than a 20-percent decline in a radical or violent society. Hence, this approach requires that a critical minimum achievement be specified as a criterion of success in each situation. Or, as a third alternative, the evaluation may compare the results with targets proclaimed in advance. The degree of success will be the extent to which those targets have been realized. Problems, however, will arise, especially when the goals happen to be contradictory or change over time.

Methods of assessment

Perceived achievement considers reform successful if the relevant parties perceive their goals as having been satisfied. One of the main objectives of reform has been to reduce conflict and promote harmony, both of which depend on whether a person or group perceives its expectations as fulfilled, whether it has hope that these expectations will be fulfilled, and whether it is able to express these expectations openly. The evaluation, therefore, is primarily subjective, and the net impact can be assessed only by synthesizing the views of the parties affected, including those who might be satisfied to have their losses minimized, just as much as others would want their gains maximized.

The closing-the-gap approach considers a reform successful to the extent that it closes the gap between the sector subject to reform and the more advanced sectors in society; in other words, reform would be expected to help integrate agriculture with the rest of the economy and the rural population with the urban community in terms of opportunities and levels of living. In this sense the reform would remove the dualism between the agrarian and the nonagrarian, between the technologically backward and the technologically advanced, and thus would increase labour mobility in response to the demands of the economy and development.

HISTORY OF LAND REFORM

The ideas and principles discussed so far may be illustrated by a selective survey of the history of land reform.

Greek system and reforms

Ancient reforms. The recorded history of reform begins with the Greeks and Romans of the 6th and 2nd centuries BC, respectively. Land in ancient Athens was held in perpetuity by the tribe or clan, with individual holdings periodically reallocated according to family size and soil fertility. Population increase, expansion of trade, growth of a money economy, and the opening up of business opportunities eventually made financial transactions in land an economic necessity. Land itself continued to be inalienable, but the right to use the land could be mortgaged. Thus, peasants could secure loans by surrendering their rights to the product of the land, as "sale with the option of redemption." Lacking other employment, the debtor continued to cultivate the land as *hektēmor,* or sixth partner, delivering five-sixths of the product to the creditor and retaining the rest for himself. Mortgaged land was marked by *horoi,* or mortgage stones, which served as symbols of land enserfment. When Solon was elected archon, or chief magistrate, *c.* 594 BC, his main objective was to free the land and destroy the *horoi.* His reform law, known as the *seisachtheia,* or "shaking-off the burdens," cancelled all debts, freed the *hektēmoroi,* destroyed the *horoi,* and restored land to its constitutional holders. Solon also prohibited the mortgaging of land or of personal freedom on account of debt.

The impact of the reform was extensive but of short duration. The *hektēmoroi* were freed, but since no alternative sources of support or credit were provided and creditors were uncompensated, dissatisfaction and instability persisted. Two decades of anarchy were followed by a revolution, *c.* 561 BC, that brought Peisistratus to power. He enforced the reform and distributed lands of his adversaries (who were killed or exiled) among the small holders. He also extended loans to aid cultivation and prevent migration to the city and expanded silver mining to create employment. Although the amount of land redistributed is unknown, Peisistratus was apparently able to satisfy the peasantry, secure their loyalty, and stay in power for life, but the economic effects are too vague to evaluate.

The reforms of the Gracchi

The Roman reform by Tiberius and Gaius Gracchus came between 133 and 121 BC. The land reform law, or *lex agraria,* of Tiberius was passed by popular support against serious resistance by the nobility. It applied only to former public land, *ager publicus,* which had been usurped and concentrated in the hands of large landholders. Land concentration reduced the number of owners and hence the number of citizens and those eligible to serve in the army. In addition, such concentration was accompanied by a shift from cultivation to grazing, which reduced employment and increased the poverty of the peasants, producing a crisis. The motives of the reformers continue to be debated, but it would appear that concern for the poor and political stability were major factors.

The *lex agraria* specified minimum and maximum individual landholdings, with an allowance for male children of the family. Excess land would be expropriated and compensation paid for improvements. A standing *collegium,* or commission, was to enforce the law, but implementation was delayed because Tiberius was killed in the year of its passage. When Gaius was elected tribune about a decade later, he revived the reform and went even further. He colonized new land and abolished rent on small holdings since rent on large holdings had been suspended as compensation for expropriation. Gaius was killed in 121 BC, however, and within a decade the reform was reversed: private acquisition of public land was legalized, the land commission was dissolved, rent on public land was abolished, all holdings were declared private property, and squatting on public land was prohibited. Even colonization was ended, and colonies established by Gaius were broken up. Another period of land concentration was inaugurated.

Modern European reforms. The French Revolution brought a new era in the history of land reform. Reform meant dealing with survivals of the medieval tenures that had left a common heritage in most European countries and, through them, in the colonies. The measures and approaches varied from place to place and period to period.

The French Revolution and after

On the eve of the Revolution, French society was polarized, with the nobility and clergy on one side and the rising business class on the other. The middle class was relatively small, especially in the rural areas. The majority of the peasants were hereditary tenants, either *censiers,* who paid a fixed money rent, or *mainmortables,* or serfs, who paid rent in the form of labour services, *corvée,* of about three days a week. The peasants paid various other feudal dues and taxes, from which the nobility and clergy were exempted. The Revolution overthrew the ancien régime and the feudal order and introduced land reform.

The reform repealed feudal tenures, freed all persons from serfdom, abolished feudal courts, and cancelled all payments not based on real property, including tithes. Rents based on real property were redeemable. Once the law had been passed, however, the peasants seized the land and refused to pay any rents or redemption fees; in 1792 all payments were finally cancelled. Land of the clergy and political emigrants was confiscated and sold at auction, together with common land. The terms of sale, however, often favoured the wealthy, which may explain the rise of a new class of large landowners among the supporters of Napoleon.

The social and political objectives of the reformers were fully realized. The *censiers* and serfs became owners. Feudalism was destroyed, and the new regime won peasant support. The economic effects, however, were limited. Incentives could not be increased substantially since the peasants already had full security of tenure prior to the reform. The scale of operations was not changed; and no facilities for credit, marketing, or capital formation were created. The major achievements were the reinforcement of private, individual ownership and perpetuation of the small family farm as a basis of democracy. The small family farm has characterized French agriculture ever since.

There were other reforms in most European countries. England resolved its land problems by the enclosure movement, which drove the small peasants into the towns, consolidated landholdings, and promoted large-scale operation and private ownership. Sweden and Denmark pioneered between 1827 and 1830 by peacefully abolishing village compulsion, or imposed labour service, and the strip system of cultivation, by consolidating the land, and by dividing the commons among the peasants. Though influenced by the French Revolution, only after the 1848 revolutions did Germany, Italy, and Spain free the peasants and redistribute the land. Reform in Ireland took a whole century before substantive results were achieved, in the mid-1930s, after Ireland was divided into Northern Ireland and the Irish Free State. The tenants were converted into owners by subsidized purchase of the land.

Russian reforms

The first major Russian reform was the emancipation of the serfs in 1861. At the time of emancipation about 45 percent of the land was private property and the remainder was held as allotment land, cultivated in units averaging 9.5 acres (3.8 hectares) by the peasant serfs against rent in kind and labour, payable to feudal lords. In contrast, fewer than 1,000 noble families owned about 175,000,000 acres (70,000,000 hectares) and received rent therefrom. Conflict between such extremes of poverty and wealth caused restlessness among the peasants and rendered reform inevitable. As Tsar Alexander II put it: "It is better to abolish serfdom from above than to await the day when it will begin to abolish itself from below."

The Emancipation Act of 1861 abolished serfdom and distributed allotment land among the peasants. The homestead became hereditary property of the individual, but the field land was vested in the village *mir* as a whole. The peasant paid redemption through the village authority, while the landlord received state bonds as compensation equal to 75 to 80 percent of the land market value. Though legally freed, the private serf had to ransom his freedom by surrendering a part of the allotment land. In contrast, serfs belonging to the imperial family were emancipated in 1863 and received the maximum amount of land fixed by law. Serfs of the state were emancipated in 1866 and allowed to keep the land they occupied against money rent. The Cossacks received two-thirds of the land, to be held in common, but in lieu of redemption payments they had to serve 20 years in the army. The serfs in mines and households were freed but received no economic assets.

Redemption payments, however, soon proved too burdensome, village restrictions were tight, and the allotment land area declined, all of which led to renewed restlessness and disturbances. Following the revolt of 1905, the government, under Pyotr Stolypin, tried to create middle-class, independent farmers by replacing the village tenure with private ownership, consolidating holdings, and encouraging land purchase by individuals; but the time was too short for effective implementation. The Soviet Revolution overthrew the tsarist regime and introduced the concepts of public ownership and collectivization.

By decree in 1918, the Soviets abolished private ownership of land, made farming the sole basis of landholding, and declared collectivization a major objective of policy. Marketing of agricultural products became a state monopoly. In 1929 Stalin embarked on a full course of collectivization, and by 1938 collective farms occupied 85.6 percent of the land and state farms 9.1 percent. Credit facilities and tractor stations supplemented collectivization, while agricultural production was integrated in the national plan for industrialization and development.

The costs of Soviet reform included the destruction of capital and the death of large numbers of kulaks, or rich peasants. Total output and productivity increased, however, and capital formation was made possible through forced saving, taxes, and regulated prices. The peasant received extensive social services such as health care, and education and better working conditions. The objectives of the decree of 1918 have been fully realized.

Reform in eastern Europe

Reform in eastern Europe was complicated by the fact that most of the eastern European countries remained under foreign rule until the middle of the 19th century or later. In Hungary, the Decree of 1853 abolished the *robot,* or forced labour and feudal dues, freed the serfs, liberalized land transaction, and encouraged consolidation. The Romanian reform of 1864 freed the serfs and distributed both the land and the redemption payments in proportion to the number of cows or oxen each peasant had. Formal emancipation in Bulgaria was introduced by the Turkish government in the 1850s, but actual reform came in 1880, after independence. Each peasant, including sharecroppers and wage workers, who had worked the land for 10 years without interruption, was entitled to the land he had cultivated. With the exception of Bulgaria, the distribution of ownership throughout most of eastern Europe remained highly uneven. Political instability reached a dangerous point between the two world wars. Following World War II, the eastern European countries established Communist governments with a strong tendency toward collective, cooperative, and mechanized agriculture.

Land reform in Mexico

Mexico. The Mexican reform of 1915 followed a revolution and dealt mainly with lands of Indian villages that had been illegally absorbed by neighbouring haciendas (plantations). Legally there was no serfdom; but the Indian wage workers, or peons, were reduced to virtual serfdom through indebtedness. Thus, the landlords were masters of the land and of the peons. The immediate aim of reform was to restore the land to its legal owners, settle the title, and use public land to reconstruct Indian villages. The motives were mainly to reduce poverty and inequality and to secure political stability, which was then in the balance. A decree of 1915 voided all land alienations that had taken place illegally since 1856 and provided for extracting land from haciendas to reestablish the collective Indian villages, or ejidos. The 1917 constitution reaffirmed those provisions but also guaranteed protection of private property, including haciendas. Nevertheless, a combination of loopholes, litigation, and reactionary forces slowed implementation, and effective reform came only after passage of the Agrarian Code of 1934 and the sympathetic efforts of Pres. Lázaro Cárdenas.

The reform restored many villages and freed the peons, but land concentration and poverty continued. In 1950, more than 31 percent of the private cropland was owned by fewer than 0.5 percent of the owners. Small-scale operation was retained or encouraged, a fact explaining the decline of output in the early years. More recently, efficiently run farms have been exempted from distribution.

The social and political impact was more positive. The peasants acquired more land and liberty, and control by landlords was reduced, although it was replaced by village restrictions. At least legally, farming became the basis of landholding. Some have seen in land reform the reason for Mexico's political stability, although there have been sporadic peasant uprisings and other violent encounters.

REFORMS SINCE WORLD WAR II

Communist and capitalist reforms

Recent decades have witnessed widespread, comprehensive reform programs, but the concept has undergone major changes. The eastern European countries and China originally followed the Soviet model, with different modifications in the individual countries. A few other countries have continued to follow that model, with major emphasis on "land to the tiller," cooperation, collective ownership, large-scale operation, and mechanization, and with economic development as the common denominator. In capitalist-oriented reforms, private ownership, family farming, and dual tenures have remained basic objectives with the aim of promoting democracy, equality, stability, and development. Under the influence and with the guidance of the United Nations, nonsocialist reforms of the 1950s were equated with community development and emphasized institutional and rural self-help in addition to land redistribution. In the 1960s the emphasis shifted to agricultural productivity and economic development by means of large-scale operation, new technology, and cooperation. The 1970s witnessed the advent of "integrated rural development" as the focus of reform and as a way of combining productive activities with improvements in the social and physical infrastructure. The integrated approach, however, soon proved to be unmanageable, and the emphasis shifted to the "target" group as the focus of reform. The most recent conception of reform has been to satisfy "basic needs," with or without land distribution, although no policymaker in the capitalist countries would openly question the idea of land redistribution or the creation of small family farms. These experiments with the concept of reform have been accompanied by attempts to broaden the concept to incorporate women as equal beneficiaries of reform in their own right. The results have been mixed.

Japan. The Japanese reform came immediately after World War II at the insistence of the Allied Occupation Army. The reform was designed to fit the uniquely high literacy rate and advanced industrial level of the country. Although the Meiji government had formally abolished feudalism and declared the land to be the property of the peasants, usurpation of land by the rich and by moneylenders had created classes of perpetual tenants and absentee landlords. In 1943, 66 percent of the land was operated by tenants against rent in kind that averaged 48 percent of the farmers' product, while population pressure resulted in fragmentation of holdings. The social class structure was closely tied to tenure, the owners in each village being at the top of the structure. Conflict between landlords and peasants was widespread.

After the war, the crisis was revived by food shortages, the breakdown of the urban economy, and the return of absentee landlords to the land. The Occupation Army insisted on reform, presumably to democratize the society and rehabilitate the economy. The reform law of 1946 established a ceiling on individual holdings and provided for expropriation and resale of excess land to the tenants against long-term payments. The government compensated the landlords in cash and bonds redeemable in 30 years. Tenants were protected by contract, and rents were reduced to a maximum of 25 percent of the product. The redistributed land was made inalienable, though this restriction was relaxed four years later. The program also provided for marketing and credit cooperatives. An important supplementary measure was the Local Autonomy Law of 1947, which decentralized the power structure and put village affairs in the hands of the villagers.

Within two years tenancy declined by more than 80 percent. Rent control and land distribution helped to equalize incomes in the villages and rehabilitate the sociopolitical status of the peasants. Crop yields per unit of land increased, but despite improved techniques the output per worker declined. In general the reform seemed to realize the objectives of the reformers and the peasants, although smallness of scale, low per capita incomes, underemployment, and insufficient mechanization have persisted. Even black market rents developed. These problems were tolerable because their effects were mitigated by the upsurge of the urban economy and the ability of the Japanese farmer to supplement the family income from nonagricultural employment. Even so, the farmers continue to depend on government subsidy to stay in farming.

Egypt. The Egyptian reform of 1952 followed the revolution that overthrew the monarchy and brought young middle-class leaders to the helm. Though affecting only about 12 percent of the arable land, it was applied thoroughly and touched all aspects of rural life. Egypt had two main forms of tenure: private ownership and *waqf,* or land held in trust and dedicated to charitable or educational purposes. *Waqf* land was inalienable, but private land was subject to speculation and concentration. In 1950, 1 percent of the owners had more than 20 percent of the private land, and 7 percent had more than two-thirds. The operating unit was small, with 77 percent of all the holdings occupying less than one acre each. Tenancy was widespread and rents were exorbitant. The peasants were exploited by middlemen who sublet the land to tenants, mediated between them and the market, and extended credit at high rates of interest.

The Reform Law of 1952

The revolutionary reformers aimed at abolishing feudalism, recruiting peasant support, promoting economic development, and bringing the villagers back into the stream of national life. The Agrarian Reform Law of 1952 put a ceiling on individual holdings at 200 *faddān*s (one *faddān* = 1.038 acres), later reduced to 100 *faddān*s, with special allowance for male children. The excess land was expropriated and distributed to the peasants in parcels not exceeding five *faddān*s. Compensation was given in bonds, while land recipients had to repay in annual installments. The new owners were obligated to join cooperatives for production, marketing, and credit. Tenancy conditions were also regulated, with contract replacing traditional terms; rent could not exceed 50 percent of the product, nor could a tenant hold more than 50 acres, to avoid subletting. An interesting feature of the reform was the special attention given to college graduates by allowing them up to 20-*faddān* parcels.

The reform was enforced quickly and had a great impact on the morale of the peasants. The economic effects, however, were minor since agriculture was intensive and land yield high. Producer cooperatives served only to offset the impact of distribution on the scale of operation. Some increases in yield have been claimed, but the evidence is still inconclusive. Furthermore, little capital was redirected into productive investment since the compensation bonds were not negotiable. Peasant savings remained limited, income increments being spent mostly on consumption. Finally, underemployment in agriculture has remained widespread. The defects of the agrarian structure continue to prevail, and relatively large ownerships exist, while certain groups in Egypt are calling for reversal of the reform.

The social and political effects, however, were far reaching. Redistribution and regulation of rent raised the incomes of small owners and tenants. Cooperatives replaced the middleman and captured his share for the farmer. The peasant gained social status and enjoyed a higher level of political participation, mostly in support of the revolutionary regime. These effects, however, can be easily exaggerated. The peasants became dependent on the cooperatives whether they liked them or not. Great differences in landholding continued to exist, and peasant incomes remained low. Black market rents appeared. The example of Egypt suggests that successful reform in densely populated countries requires an upsurge in the industrial sector to relieve population pressure and permit technical advance and higher productivity in agriculture.

Southeast Asia. The model of Japan's reform has been attempted in Southeast Asia, especially in Taiwan, South Korea, and South Vietnam, all influenced by American experts and by the anti-Communism of the respective governments. The objectives were to sustain the political order, raise living standards, and promote some degree of economic development. The reforms began with regulation of tenancy, restriction of rent, and the institution of written contract for leases, following which tenants were to be transformed into owners. Taiwan's reform was implemented between 1949 and 1953, in three stages. First, rents, which had sometimes reached 70 percent of the product, were reduced to 37.5 percent. Next, tenant-farmed public land was sold to the tenants. Finally, tenant-farmed private land was bought by the government and resold to the tenants.

The Vietnamese reform was introduced in 1955. Rents were reduced to a maximum of 25 percent of the product. A ceiling of 247 acres (100 hectares) was put on individual holdings, however, and only the excess land was subject to redistribution in parcels of 7.4 to 12.4 acres (three to five hectares) to the tenants. The collapse of the South Vietnamese regime and the unification of South and North Vietnam ended that reform and replaced it with the socialist model of North Vietnam.

The reform in Taiwan, as in South Vietnam prior to unification, was supplemented by other measures described as community development, such as adult education, credit facilities, improved technology, and other social services. Though land consolidation was attempted, the scale of operation was little affected. The main effect seems to have been the regulation of tenancy and the redistribution of rent incomes. An innovation of Taiwan's reform was the partial compensation of landlords with industrial shares in public enterprises, which helped them and helped industry.

Taiwan's reform has been hailed as a major success, in both economic and political terms. Some observers, however, are unwilling to reach such a conclusion until restrictions are removed and the peasants have a free choice of tenure and farm organization.

South Korea's land reform (under the Land Reform Law of June 1949) roughly followed the Japanese model by removing tenancy, creating small ownerships, implementing the law thoroughly and promptly, and depending heavily on nonagricultural (basically industrial) employment to absorb labour and supplement rural income. Like the Japanese and Taiwanese reforms, Korea's successful reform was generously supported by foreign aid.

The Philippines introduced a reform program in 1963, which aimed primarily at replacing share tenancy with lease contracts and eventually with ownership, and at revitalizing agriculture through extension services. By the mid-1980s the program had given titles to about 400,000 tenants and secure leases to another 600,000, but the economic viability of the new units has been uncertain because of their small scale and the lack of supplementary facilities. The main effects initially were seed improvement, greater use of fertilizers, and an increase in contractual tenancy. To combat the negative effects of small-scale farming, the Philippine government has resorted to what it calls the "compact farm," which is a voluntary grouping of small farms to be operated under one management as one consolidated farm. The problem of surplus labour, however, remains to be solved.

Reform in Malaysia

Various other reforms have been introduced in Southeast Asia, but the only innovative program has been that of Malaysia. The program in Malaysia has been highly organized and development oriented. It tries to promote social and economic objectives by emphasizing the production of rubber and palm oil for export and gradually transforming the landless into hereditary tenants on newly reclaimed and settled plantations. A typical plantation covers 4,500 to 5,000 acres (1,800 to 2,000 hectares) of jungle land and absorbs about 400 families. The land is cleared and planted by contract, and a village is constructed, with all the necessary services, before the settlers arrive. Each house has a quarter of an acre for a household garden. Cropland is divided in blocks of 120 to 200 acres (48 to 80 hectares), to be worked by a team of 15 to 25 people until the plants have matured. Upon maturation, each settler receives a share by lottery and a lease title for 99 years. This tenure arrangement precludes alienation, subdivision, or subleasing; it thus protects the tenant farmer and sidesteps the Islāmic laws of inheritance, which tend toward fragmentation of the land.

The settler is responsible for the cost of clearing and planting, but the government pays the administrative costs. The settler is guaranteed supplementary employment to earn subsistence income pending maturity of the plants, and cultivation is guided by experts. The rate of settlement is determined by the overall economic plan. It is clear that landholding has become tied to cultivation; fragmentation and diseconomies of scale have been avoided, and cultivation has become a rational economic operation. The Malaysian program has much in common with the cooperative settlements of Israel and the Gezira Scheme in The Sudan.

Latin America. Except for the early example of Mexico, reform in Latin America has been recent and appears to have come only in response to the threat of social and political instability and mounting international pressures. Reform in Latin America after World War II must be seen against a background of rapidly increasing population and of extreme contrasts between plantation economies and small units; high concentration of land ownership, income, and power and dire poverty; modern farming and relatively backward cultivation methods; and nationalism and extensive foreign ownership of land. In addition, Latin-American society is complicated by its ethnic mixtures and by dependence on staple trade items such as sugar, tobacco, cocoa, coffee, and beef cattle.

Reform in Latin America has reflected the ideologies and objectives of the regime in power. Brazil has had several attempts at reform. The measures have been indirect and relatively mild, the most important being taxation of idle land and large plantations and reclamation and settlement of the Amazon region, with provisions for credit and tenancy protection. The results have been modest, however, largely because of the physical and biological hardships faced by settlers in the tropical Amazon environment. Peru has deviated by creating collective administrations of the nationalized feudal estates. The title resides in the nation, and the estates are run by the Agricultural Societies of Social Interest (SAIS), a mechanism devised to avoid breaking up economically efficient enterprises rather than to modify the tenure institutions.

Reform in Cuba

At the other end of the Latin-American spectrum is the Cuban reform that followed the revolution of 1958. Cuba retained private ownership but reduced it substantially in favour of the public sector. As proclaimed a few months before the overthrow of the old regime, the reform aimed at the elimination of *latifundia* tenure, expropriation of land owned by foreign companies, higher standards of living for the peasantry, and national economic development. It began by setting a ceiling of 30 *caballerías* (one *caballería* = 33 acres, or 13.4 hectares) on individual holdings, with a maximum of 100 *caballerías* if economic operations required such a scale. All foreign-owned land was nationalized. Public land on which rice and cattle were raised was converted into state farms, and the peasants became permanent wage workers on these farms. Sugar plantations were converted into cooperatives to avoid their subdivision into small uneconomic units. Before long the ceiling on individual holdings was lowered to five *caballerías,* and all such holdings became private family farms. The rest were nationalized, and the expropriated owners were compensated with a pension for life. The reform was supplemented by the organization of national farmer associations; people's stores; credit, housing, and educational facilities; and the production of machinery and fertilizers. In 1963 a major reorganization of state farms took place; they were subdivided on the basis of crop specialization into smaller operational units of about 469 *caballerías.*

Effects of the reform were comprehensive and immediate. The tenure institutions were radically changed in favour of public ownership, while *minifundia* and tenancies were abolished. Socially and politically, the reform realized the objectives of the reformers. Economically, the government claimed higher yields of sugarcane, vegetables, and fruit, but this claim has been disputed by foreign observers.

Other Latin-American reforms fall between those of Brazil and Cuba, though closer to the former than to the latter in comprehensiveness and thoroughness. For example, the reform in Costa Rica has overlooked land concentration and income inequality and concentrated on the squatters, or *parásitos,* who in 1961 numbered between 12,000 and 16,000 people. The reform aimed at legalizing existing squatter holdings, preventing further squatting, and conserving virgin land. Even this modest program was implemented very slowly. As late as 1973, 7.3 percent of the landholdings comprised 67 percent of the total agricultural land. Colombia has had reform programs for at least 30 years, but concentration of ownership, fragmented holdings, backward methods of cultivation, inequality of income distribution, and widespread poverty have remained characteristic; in 1970, 4.3 percent of the holdings contained 67.4 percent of the total area.

Chile

Chile undertook various reform programs before achieving concrete results. In 1962 a program was enacted to encourage settlement of new land, but only about 1,000 families were settled. A comprehensive reform was introduced in 1965 with three main objectives: to make the agricultural workers owners of the land they had cultivated previously, to increase agricultural and livestock production, and to facilitate social mobility and peasant participation in political life. The Chilean reform was unique in its method of implementation. Once the plantation had been designated for expropriation and the prospective owners selected, they were organized into *asentamientos,* or settlement groups. The group elected a committee to take charge of settlement. The members cultivated the land as a team for three to five years. Meanwhile they received training and guidance in social participation, decision making, and modern farming. Upon completion of the transition period, the land was divided among those who had shown promise, to be held outright and without restriction. All new owners were obligated to join cooperatives, the form of these being determined by the members. The socialist regime that came into office in 1970 expedited the expropriation process and the creation of settlement groups or cooperative farms under peasant committees. By 1972 all the potential land, which had been in farms larger than 200 acres (80 hectares), had been expropriated and reallocated. The new regime that took over in 1973 decided, however, to privatize the land and reverse much of the reform by returning large areas to the former owners, dissolving the cooperatives, and creating private ownerships in their place. Most of the reverse changes had been completed by 1979. Nevertheless, most of the excess land in farms of more than 200 acres remained in the hands of the reform beneficiaries. Owners of less than 12 acres (five hectares) were hardly affected; those who owned between 12 and 50 acres (five and 20 hectares) benefitted most. In the final analysis, less than 15 percent of the agricultural land was affected by the reform between 1965 and 1979 under three regimes.

Observers of the Latin-American scene have been pessimistic regarding the adequacy of these land reform programs. With the exception of Cuba, capital formation in agriculture has not increased substantially; the pattern of land distribution has undergone little change; social and political stability have remained in question; and the agrarian structure is still considered defective.

Other recent reforms. Attempts to reform the agrarian structure have been made in most other countries, with varying degrees of seriousness. India and Pakistan have concentrated on abolishing intermediaries who prevailed as survivals of traditional and feudal tenures. In India the tenants have become hereditary holders, with the title vested in the state. India has left reform to the states and emphasized peaceful and compensatory methods; hence the results have varied from one state to another. Pakistan, following the revolution of 1958, enacted a reform that made most of the tenants owners. In both countries, however, small-scale farming has persisted, while Pakistan has continued to tolerate and protect owners of up to 500 acres (200 hectares). In neither country has fragmentation been effectively reduced or have capital formation and cultivation methods significantly advanced.

India and Pakistan

In contrast, after the Communists came to power in China, private ownership was eliminated and the peasants were organized in village communes. Extensive supplementary measures have been tried, and the role and organization of the commune have varied according to the pressures on the economy. The most recent innovation in China's agriculture has been the "production responsibility system," which allows the commune to contract with its members for quotas of output; the members are free to sell the surplus on the open market. The change is seen as an incentive generator, but land cannot be rented, bought, sold, or used except as authorized by the commune. The effects of China's agrarian policy on peasant living conditions and the Chinese economy have been generally accepted as positive, genuine, and impressive.

In 1962 Iran made owners of most of the former sharecroppers, in the classic tradition of Western-type reform, mainly to create political stability. Given Iran's revolution of 1979, however, the reform evidently was not sufficient to sustain the old social order. Reform was also introduced in Syria, Iraq, Algeria, Libya, and other countries of the Middle East and North Africa following independence or revolution. Most of these reforms were influenced by the Egyptian example, with the state playing a major role. In all cases emphasis has been placed on farm cooperatives, although they have been largely ineffective.

In contrast, tropical Africa has witnessed a wave of innovative reform in recent years. Reform has sometimes come in "packages," which combine tenure reform and other measures affecting cultivation and productivity. Among the innovations is the "villagization," or *ujamaa,* program of Tanzania, according to which a group of families lives, works, and makes decisions together and shares the costs and benefits of farming the land. The program began as a voluntary movement in 1967, but by 1977 it had become almost mandatory. At the same time, "block farming" and individual holdings had become acceptable forms of cooperation. The Ujamaa Villages Act of 1975 made the village the main rural administration and development unit. The most radical reforms in Africa, however, have been those of Ethiopia in 1975 and of Mozambique in 1979. Both vested the land title in the nation and abolished rent, sale, and absentee control of the land. The land was placed in the hands of the tillers, who have guaranteed right of use for themselves and for their descendants. Except in the public sector, farming is a small, family operation with a high degree of equality of landholding but of uncertain efficiency.

CONCLUSIONS

Land reform and agrarian reforms have become synonymous, indicating that reform programs have become more comprehensive and encompass much more than the reform of land tenure or land distribution. Reform movements have recurred throughout history, as have the crises they are intended to deal with, because reform has rarely dealt with the roots of the crises. Reform has served as a problem-solving mechanism and therefore has only been extensive enough to cope with the immediate crisis. Reformers have often faced hard choices: to promote and sustain private ownership with inequality or to institute public or collective ownership with equality but with restrictions on the individuals' private interests; to spread employment by supporting labour-intensive, low-productivity techniques or to promote high productivity through capital-intensive, efficient methods; to pursue gradual "repair and maintenance" reform that is basically ineffective or to promote revolutionary, comprehensive, effective but disruptive reform. In capitalist reforms these contradictions have usually been resolved in favour of the first set of options; in socialist reforms, in favour of the second. Land tenure reform seems to have been of little significance in creating substantive economic change, although it has been important for improving the status of peasants and maintaining social and political stability. Most reforms have narrowed the gap between reform beneficiaries and other farmers through land redistribution and tenancy control, but only the comprehensive socialist reforms have narrowed the gap between agriculture and other sectors of the economy.

Land redistribution programs have had limited success for several reasons. They often have deprived the farm of the former landlord's contributions without providing a substitute. They have inhibited mobility of labour by giving the peasant a stake in the land, though only in the form of an inefficient minifarm. They frequently have threatened large, efficiently run farms and therefore have had to be compromised. They have provided compensation for the expropriated land and hence left wealth and income distribution largely unaffected. They have been conditional upon peasant participation in social and political activity and cooperative organization, even though the peasant was unprepared for these activities. Moreover, the redistribution of land has rarely been fortified by protective measures that could prevent reconcentration of ownership and the recurrence of crises. Nevertheless, major efforts have been expended by the Food and Agriculture Organization of the United Nations and other international bodies and by governments to devise viable frameworks for solving agricultural and rural problems emanating from defective agrarian structures. (E.H.T.)

BIBLIOGRAPHY. The following references form a point of departure for further study; each of these, in turn, provides additional bibliographies.

Philosophy and logic: A. WHITNEY GRISWOLD, *Farming and Democracy* (1952), a lucid scholarly analysis of the political significance of various farm structures and a discussion of the family farm as a cornerstone of democracy; DOREEN WARRINER, *Land Reform in Principle and Practice* (1969), a provocative discussion of the evolution of reform and conflict between theory and practice; ERICH H. JACOBY, *Evaluation of Agrarian Structures and Agrarian Reform Programs* (1966), a checklist of administrative and organizational changes that contribute to reform success; WORLD CONFERENCE ON AGRARIAN REFORM AND RURAL DEVELOPMENT, *The Peasants' Charter* (1981), a declaration of principles and program of action by most of the nonsocialist developing countries; FOLKE DOVRING, *Land and Labor in Europe in the Twentieth Century,* 3rd rev. ed. (1965), an examination of the political and ideological bases of land policy since the French Revolution; DAVID A. PRESTON (ed.), *Environment, Society, and Rural Change in Latin America: The Past, Present, and Future in the Countryside* (1980), a collection of essays, pessimistic as to the efficacy of most reform; ELIAS H. TUMA, *Twenty-Six Centuries of Agrarian Reform: A Comparative Analysis* (1965), a synthesis of theory and history, with an attempt to formulate a general theory of agrarian reform, and "Agrarian Reform in Historical Perspective Revisited," *Comparative Studies in Society and History,* 21:3–29 (1979), which updates and retests the theory of the previous reference.

Ancient reforms: IVAN M. LINFORTH, *Solon the Athenian* (1919, reprinted 1971); W.J. WOODHOUSE, *Solon the Liberator: A Study of the Agrarian Problems in Attica in the Seventh Century* (1938, reissued 1968); W. WARDE-FOWLER, "Notes on Gaius Gracchus," *English Historical Review,* 20:209–227 and 417–433 (1905); and E.G. HARDY, "Were the Lex Thoria of 118 B.C. and the Lex Agraria of 111 B.C. Reactionary Laws?" *Journal of Philosophy,* 31:268–286 (1910).

Medieval and modern reforms: ELIAS H. TUMA, *European Economic History: Tenth Century to the Present* (1971, reprinted 1979), includes summaries of the agrarian reform history of most countries of Europe since the Middle Ages. On early French reforms, see G. LEFEBVRE, *The Coming of the French Revolution: 1789* (1947, reissued 1967; originally published in French, 1939); and M.I.B. BLOCH, *French Rural History: An Essay on Its Basic Characteristics* (1966; originally published in French, 1952–56). Reform in tsarist Russia is addressed in G.T. ROBINSON, *Rural Russia Under the Old Regime: A History of the Landlord-Peasant World and a Prologue to the Peasant Revolution of 1917* (1932, reissued 1960).

Any study of recent decades must begin with the UNITED NATIONS series, *Progress in Land Reform: Sixth Report* (1979); and with FAO, *Review and Analysis of Agrarian Reform and Rural Development in the Developing Countries Since the Mid-1960s* (n.d.), and *Land Reform: Land Settlement and Cooperatives* (semiannual), which offers feature articles, summaries of recent reform legislation, and a comprehensive bibliography of new literature. See also M. RIAD EL GHONEMY (ed.), *How Development Strategies Benefit the Rural Poor* (1984), an FAO summary of reform programs. For other country studies, reform summaries, and analyses, see V.E. STANIS, *Socialist Transformation of Agriculture: Theory and Practice* (1976), covering reform in eastern Europe after World War II; RUSSELL KING, *Land Reform: A World Survey* (1977), a good summary for the general reader, with references; and AJIT KUMAR GHOSE (ed.), *Agrarian Reform in Contemporary Developing Countries* (1983), covering developments in agrarian reform through the early 1980s, with illustrations from the field. JEAN LE COZ, *Les Reformes agraires: De Zapata à Mao Tsé Tung et La FAO* (1974), which compares the Soviet model with the Chinese, and with reform in Latin America and the Middle East, and highlights the role of the Food and Agriculture Organization in reform implementation.

Studies of individual countries and regions include: MAURICE H. DOBB, *Soviet Economic Development Since 1917* (1948); GEORGE L. YANEY, *The Urge to Modernize: Agrarian Reform in Russia: 1861–1930* (1982): ERIC J. HOOGLUND, *Land and Revolution in Iran: 1960–1980* (1982); VIVIENNE SHUE, *Peasant China in Transition: The Dynamics of Development Toward Socialism, 1949–1956* (1980), a close study of the initial stages of agricultural development and state control of the rural economy; ANTHONY Y.C. KOO, *Land Market Distortion and Tenure Reform* (1982), a study concentrating on Taiwan and Southeast Asia; and STEVEN E. SANDERSON, *Agrarian Populism and the Mexican State: The Struggle for Land in Sonora* (1981), covering 1917–76, with much discussion of the national situation.

Language

Language interacts with every other aspect of human life in society, and it can be understood only if it is considered in relation to society. This article attempts to survey language (both spoken and written) in this light and to consider its various functions and the purposes it can and has been made to serve. Because each language is both a working system of communication in the period and in the community wherein it is used and also the product of its past history and the source of its future development, any account of language must consider it from both these points of view.

The science of language is known as linguistics. It includes what are generally distinguished as descriptive linguistics and historical linguistics. Linguistics is now a highly technical subject; it embraces, both descriptively and historically, such major divisions as phonetics, grammar, and semantics, dealing in detail with these various aspects of language. For a full account of the theory and methods of linguistic science, see the article LINGUISTICS. For a general survey of known living and dead languages, consult the article LANGUAGES OF THE WORLD.

This article is divided into the following sections:

Characteristics of language

DEFINITIONS OF LANGUAGE

Many definitions of language have been proposed. Henry Sweet, an English phonetician and language scholar, stated: "Language is the expression of ideas by means of speech-sounds combined into words. Words are combined into sentences, this combination answering to that of ideas into thoughts." The U.S. linguists Bernard Bloch and George L. Trager formulated the following definition in their *Outline of Linguistic Analysis* (1942): "A language is a system of arbitrary vocal symbols by means of which a social group cooperates." Definitions like these and, indeed, any succinct definition make a number of presuppositions and beg a number of questions. The first, for example, puts excessive weight on "thought," and the second uses "arbitrary" in a specialized, though legitimate, way (see below).

A number of considerations enter into a proper understanding of language as a subject:

1. Every physiologically and mentally normal person acquires in childhood the ability to make use, as both speaker and hearer, of a system of vocal communication that comprises a circumscribed set of noises resulting from movements of certain organs within his throat and mouth. By means of these he is able to impart information, to express feelings and emotions, to influence the activities of others, and to comport himself with varying degrees of friendliness or hostility toward persons who make use of substantially the same set of noises.

2. Different systems of vocal communication constitute different languages; the degree of difference needed to establish a different language cannot be stated exactly. No two people speak exactly alike; hence, one is able to recognize the voices of friends over the telephone and to keep distinct a number of different unseen speakers in a radio broadcast. Yet, clearly, no one would say that, for that reason, they speak different languages. Generally, systems of vocal communication are recognized as different languages if they cannot be understood without specific learning by both parties, though the precise limits of mutual intelligibility are hard to draw and belong on a scale rather than on either side of a definite dividing line. Substantially different systems of communication that may impede but do not prevent mutual comprehension are referred to as dialects of a language. In order to describe in detail the actual different speech patterns of individuals, the term idiolect, meaning the speech habits of a single person, has been coined.

Languages, dialects, and idiolects

3. Normally, people acquire a single language initially—their first language, or mother tongue, the language spoken by their parents or by those with whom they are brought up from infancy. Subsequent "second" languages are learned to different degrees of competence under various conditions, but the majority of the world's population remains largely monolingual. Complete mastery of two languages is designated as bilingualism; in a few special cases—such as upbringing by parents speaking different languages at home—speakers grow up as bilinguals, but ordinarily the learning, to any extent, of a second or other language is an activity superimposed on the prior

mastery of one's first language and is a different process intellectually.

4. Language, as described above, is species-specific to man. Other members of the animal kingdom have the ability to communicate, through vocal noises or by other means, but the most important single feature characterizing human language (that is, every individual language), against every known mode of animal communication, is its infinite productivity and creativity. Human beings are unrestricted in what they can talk about; no area of experience is accepted as necessarily incommunicable, though it may be necessary to adapt one's language in order to cope with new discoveries or new modes of thought.

Animal communication systems are by contrast very tightly circumscribed in what may be communicated. Indeed, displaced reference, the ability to communicate about things outside immediate temporal and spatial contiguity, which is fundamental to speech, is found elsewhere only in the so-called language of bees. Bees are able, by carrying out various conventionalized movements (referred to as bee dances) in or near the hive, to indicate to others the locations and strengths of nectar sources. But nectar sources are the only known theme of this communication system. Surprisingly, however, this system, nearest to human language in function, belongs to a species remote from man in the animal kingdom and is achieved by very different physiological activities from those involved in speech. On the other hand, the animal performance superficially most like human speech, the mimicry of parrots and of some other birds that have been kept in the company of humans, is wholly derivative and serves no independent communicative function. Man's nearest relatives among the primates, though possessing a vocal physiology very similar to that of humans, have not developed anything like a spoken language.

HISTORICAL ATTITUDES TOWARD LANGUAGE

As is evident from the introduction above, human life in its present form would be impossible and inconceivable without the use of language. People have long recognized the force and significance of language. Naming—applying a word to pick out and refer to a fellow human being, an animal, an object, or a class of such beings or objects—is only one part of the use of language, but it is an essential and prominent part. In many cultures men have seen in the ability to name an ability to control or to possess; this explains the reluctance, in several primitive and other communities, with which names are revealed to strangers and the taboo restrictions found in several parts of the world on using the names of persons recently dead. Lest it be thought that attitudes like this have died out in modern civilized communities, it is instructive to consider the widespread and perhaps universal taboos on naming directly things considered obscene, blasphemous, or very fearful. Indeed, use of euphemistic substitutes for words referring to death and to certain diseases actually seems to be increasing in some civilized areas.

The practice of naming

Not surprisingly, therefore, several independent traditions ascribe a divine or at least a supernatural origin to language or to the language of a particular community. The biblical account, representing ancient Jewish beliefs, of Adam's naming the creatures of the Earth under God's guidance is well known:

> So out of the ground the Lord God formed every beast of the field and every bird of the air, and brought them to the man to see what he would call them; and whatever the man called every living creature, that was its name (Gen. 2:19).

Norse mythology preserves a similar story of divine participation in the creation of language, and in India the god Indra is said to have invented articulate speech. In the much more sophisticated debate on the nature and origin of language given in Plato's Socratic dialogue *Cratylus,* Socrates is made to speak of the gods as those responsible for first fixing the names of things in the proper way.

A similar divine aura pervades early accounts of the origin of writing. The Norse god Odin was held responsible for the invention of the runic alphabet. The inspired stroke of genius whereby the ancient Greeks adapted a variety of the Phoenician consonantal script so as to represent the distinctive consonant and vowel sounds of Greek, thus producing the first alphabet such as is known today, was linked with the mythological figure Cadmus, who, coming from Phoenicia, was said to have founded Thebes and introduced writing into Greece. The Arabs had a traditional account of their script, together with the language itself, being given to Adam by God.

Biblical traditions concerning language

The later biblical tradition of the Tower of Babel (Gen. 11:1–9) exemplifies three aspects of early thought about language: (1) divine interest in and control over its use and development, (2) a recognition of the power it gives to man in relation to his environment, and (3) an explanation of linguistic diversity, of the fact that people in adjacent communities speak different and mutually unintelligible languages, together with a survey of the various speech communities of the world known at the time to the Hebrews.

The origin of language has never failed to provide a subject for speculation, and its inaccessibility adds to its fascination. Informed investigations of the probable conditions under which language might have originated and developed are seen in the late-18th-century essay of the German philosopher Johann Gottfried von Herder, "Abhandlung über den Ursprung der Sprache" ("Essay on the Origin of Language"), and in numerous other treatments. But people have tried to go further, to discover or to reconstruct something like the actual forms and structure of man's first language. This lies forever beyond the reach of science, in that spoken language in some form is almost certainly coeval with *Homo sapiens.* The earliest records of written language, the only linguistic fossils man can hope to have, go back no more than about 4,000 or 5,000 years. Attempts to derive human speech from imitations of the cries of animals and birds or from mere ejaculations of joy and grief, as if onomatopoeia were the essence of language, were ridiculed for their inadequacy by the Oxford philologist F. Max Müller in the 19th century and have been dubbed the bowwow and pooh-pooh theories.

On several occasions attempts have been made to identify one particular existing language as representing the original or oldest tongue of mankind, but, in fact, the universal process of linguistic change rules out any such hopes from the start. The Greek historian Herodotus told a story that King Psammetichus of Egypt caused a child to be brought up without ever hearing a word spoken in its presence. On one occasion it ran up to its guardian as he brought it some bread, calling out "bekos, bekos"; this, being said to be the Phrygian word for bread, proved that Phrygian was the oldest language of mankind. The naïvetè and absurdity of such an account have not prevented its repetition elsewhere and at other times.

In Christian Europe the position of Hebrew as the language of the Old Testament gave valid grounds through many centuries for regarding Hebrew, the language in which God addressed Adam, as the parent language of all mankind. Such a view continued to be expressed even well into the 19th century. Only since the mid-1800s has linguistic science made sufficient progress finally to clarify the impracticability of speculation along these lines.

Language and thinking

When people have begun to reflect on language, its relation to thinking becomes a central concern. Several cultures have independently viewed the main function of language as the expression of thought. Ancient Indian grammarians speak of the soul apprehending things with the intellect and inspiring the mind with a desire to speak; and in the Greek intellectual tradition Aristotle declared, "Speech is the representation of the experiences of the mind" (*On Interpretation*). Such an attitude passed into Latin theory and thence into medieval doctrine. Medieval grammarians envisaged three stages in the speaking process: things in the world exhibit properties; these properties are understood by the mind of man; and, in the manner in which they have been understood, so they are communicated to others by the resources of language.

Rationalist writers on language in the 17th century gave essentially a similar account: speaking is expressing thoughts by signs invented for the purpose, and words of different classes (the different parts of speech) came into being to correspond to the different aspects of thinking.

Such a view of language continued to be accepted as generally adequate and gave rise to the sort of definition proposed by Henry Sweet and quoted above. The main objection to it is that it either gives so wide an interpretation to thought as virtually to empty the word of any specific content or gives such a narrow interpretation of language as to exclude a great deal of normal usage. A recognition of the part played by speaking and writing in social cooperation in everyday life has highlighted the many and varied functions of language in all cultures, apart from the functions strictly involved in the communication of thought, which had been the main focus of attention for those who approached language from the standpoint of the philosopher. To allow for the full range of language used by speakers, more comprehensive definitions of language have been proposed in recent years on the lines of the second one quoted above (*i.e.,* "A language is a system of arbitrary vocal symbols by means of which a social group cooperates").

A rather different criticism of accepted views on language began to be made in the 18th century, most notably by the French philosopher Étienne Bonnot de Condillac in "Essai sur l'origine des connaissances humaines" (1746; "Essay on the Origin of Human Knowledge") and by Johann Gottfried von Herder. These men were concerned with the origin and development of language in relation to thought in a way that earlier students had not been. The medieval and rationalist views implied that man as a rational, thinking creature invented language to express his thoughts, fitting words to an already developed structure of intellectual competence. With the examination of the actual and the probable historical relations between thinking and speaking, it became more plausible to say that language emerged not as the means of expressing already formulated judgments, questions, and the like but as the means of thought itself, and that man's rationality developed together with the development of his capacity for speaking.

Language as a means of thought itself

The relations between thought and speech are certainly not fully explained today, and it is clear that it is a great oversimplification to define thought as subvocal speech, in the manner of some behaviourists. But it is no less clear that propositions and other alleged logical structures cannot be wholly separated from the language structures said to express them. Even the symbolizations of modern formal logic are ultimately derived from statements made in some natural language and are interpreted in that light.

The intimate connection between language and thought, as opposed to the earlier assumed unilateral dependence of language on thought, opened the way to a recognition of the possibility that different language structures might in part favour or even determine different ways of understanding and thinking about the world. Obviously, all people inhabit a broadly similar world, or they would be unable to translate from one language to another; but, equally obviously, they do not all inhabit a world exactly the same in all particulars, and translation is not merely a matter of substituting different but equivalent labels for the contents of the same inventory. From this stem the notorious difficulties in translation, especially when the systematizations of science, law, morals, social structure, and so on are involved. The extent of the interdependence of language and thought—linguistic relativity, as it has been termed—is still a matter of debate, but the fact of such interdependence can hardly fail to be acknowledged.

WAYS OF STUDYING LANGUAGE

Languages are immensely complicated structures. One soon realizes how complicated any language is when trying to learn it as a second language. If one tries to frame an exhaustive description of all the rules embodied in one's language—the rules by means of which a native speaker is able to produce and to understand an infinite number of correct, well-formed sentences—one can easily appreciate the complexity of the knowledge acquired by a child in mastering his mother tongue. The descriptions of languages written so far are in most cases excellent as far as they go, but they still omit more than they contain of an explicit account of a native speaker's competence in his language, by virtue of which one calls him a speaker of English, French, Swedish, or Swahili. The most recent developments in the study of language have served to reveal just how much more there is to do to bring palpable fact within systematic statement.

A detailed treatment of the science of linguistics is found elsewhere (see LINGUISTICS). Here it is proposed simply to give a brief outline of the way language or languages can be considered and described from different points of view, or at different levels, each contributing something essential and unique to a full understanding of the subject.

Phonetics and phonology. The most obvious aspect of language is speech. Speech is not essential to the definition of an infinitely productive communication system, such as is constituted by a language. But, in fact, speech is the universal material of human language, and the conditions of speaking and hearing have, throughout human history, shaped and determined its development. The study of speech sounds and of the physiology of speaking is called phonetics; this subject is dealt with further below, as well as in the article SPEECH: *The phonetics of speech.* Articulatory phonetics relates to the physiology of speech and acoustic phonetics to the physics of sound waves, their transmission and reception.

Speech, the universal material of human language

Phonetics covers much of the ground loosely referred to in language study as pronunciation. But, from a rather different point of view, speech sounds are also studied in phonology. Every language makes use of a very wide range of the articulations and resultant sounds that are available within the human vocal and auditory resources. Each language uses a somewhat different range, and this is partly responsible for the difficulty of learning to speak a foreign language and for speaking it "with an accent." But mere repertoires of sounds are not all that is involved. Far fewer general classes of sounds are distinctive (carry meaning differences) in any language than the number of sounds that are actually phonetically different. The English *t* sounds at the beginning and end of "tot" and in the two places in "stouter" are all different, though these differences are not readily noticed by English speakers; and, rightly, the same letter is used for them all. Similar statements could be made about most or all of the other consonant and vowel sounds in English.

What is distinctive in one language may not be distinctive in another or may be used in a different way; this is an additional difficulty to be overcome in learning to speak and understand a foreign language. In Chinese and in several other languages loosely called tone languages, the pitch, or tone, on which a syllable is said helps to distinguish one word from another: *ma* in northern Chinese on a level tone means "mother," on a rising tone means "hemp," and on a falling tone means "to curse." In English and in most of the languages of Europe (though not all—Swedish and Norwegian are exceptions) pitch differences do not distinguish one word from another, but form part of the intonation tunes that contribute to the structure and structural meaning of spoken sentences.

Languages differ in the ways in which consonant and vowel sounds can be grouped into syllables in words. English and German tolerate several consonants before and after a single vowel: "strengths" has three consonant sounds before and three after a single vowel sound (*ng* and *th* stand for one sound each). Italian does not have such complex syllables, and in Japanese and Swahili, for example, the ratio of consonant and vowel sounds in syllables and in words is much more even. Speakers of such languages find English words of the sort just mentioned very hard to pronounce, though to an Englishman they are perfectly "natural," "natural" in this context meaning "within the sounds and sound sequences whose mastery is acquired in early childhood as part of one's mother tongue."

All these considerations relating to the use of speech sounds in particular languages fall under the general heading of phonology; phonology is often regarded as one component of language structure.

Grammar. The other component is grammar. There is more to language than sounds, and words are not to be regarded as merely sequences of syllables. The concept

Relations between words in sentences

of the word is a grammatical concept; in speech, words are not separated by pauses, but they are recognized as recurrent units that make up sentences. Very generally, grammar is concerned with the relations between words in sentences. Classes of words, or parts of speech, as they are often called, are distinguished because they occupy different places in sentence structure, and in most languages some of them appear in different forms according to their function (English "man," "men"; "walk," "walked"; "I," "me"; and so on). Languages differ in the extent to which word-form variation is used in their grammar; Classical Chinese had almost none, English does not have much, and Latin and Greek had quite a lot. Conversely, English makes much more use of word order in grammar than did Latin or Greek.

Traditionally, grammar has been divided into syntax and morphology, syntax dealing with the relations between words in sentence structure, morphology with the internal grammatical structure of words. The relation between "boy" and "boys" and the relationship (irregular) between "man" and "men" would be part of morphology; the relation of concord between "the boy [or "man"] is here" and "the boys [or "men"] are here" would be part of syntax. It must, however, be emphasized that the distinction between the two is not as clear-cut as this brief illustration might suggest. This is a matter for debate among linguists of different persuasions; some would deny the relevance of distinguishing morphology from syntax at all, referring to grammatical structure as a whole under the term syntax.

Grammar is different from phonology, though the word grammar is often used comprehensively to cover both aspects of language structure. Categories such as plural, past tense, and genitive case are not phonological categories. In spoken language they are, like everything else, expressed in speech sounds, but within a language these may be very different for one and the same category. In English noun plurals, the added *-s* in "cats," the vowel changes in "man, men" and in "goose, geese," and the *-en* in "oxen" are quite different phonologically; so are the past-tense formatives such as *-ed* in "guarded," *-t* in "burnt," vowel change in "take, took," and vowel and consonant change in "bring, brought." In Latin the genitive case can be represented in singular nouns by *-ī, -is, -ae, -ūs,* and *-eī.* The phonological difference does not matter, provided only that the category distinction is somehow expressed.

The same is true of the orthographic representation of grammatical differences, and the examples just given illustrate both cases. This is why the grammar of written language can be dealt with separately. In the case of dead languages, known with certainty only in their written forms, this must necessarily be done; insofar as the somewhat different grammar of their spoken forms made use of sound features not represented in writing (*e.g.,* stress differences), this can, at best, only be inferred or reconstructed.

Meaning expressed by grammatical forms, vocabulary, and intonation

Grammatical forms and grammatical structures are part of the communicative apparatus of languages, and along with vocabulary, or lexicon (the stock of individual words in a language), they serve to express all the meanings required. Spoken language has, in addition, resources such as emphatic stressing and intonation (see below). This is not to say, however, that grammatical categories can be everywhere directly related to specific meanings. Plural and past tense are fairly clear as regards meaning in English, but even here there are difficulties; in "if I knew his address I would tell you," the past-tense form "knew" refers not to the past but to an unfulfilled condition in the present. In some other languages greater problems arise. The gender distinctions of French, German, and Latin are very much part of the grammar of these languages, but only in a small number of words do masculine, feminine, and neuter genders correspond with differences of sex, or with any other category of meaning in relation to the external world (see also LINGUISTICS).

Semantics. Language exists to be meaningful; the study of meaning, both in general theoretical terms and in reference to a specific language, is known as semantics. It embraces the meaningful functions of phonological features, such as intonation, and of grammatical structures and the meanings of individual words. Once again, it must be stressed that questions arising from the relations between grammar and meaning and between grammar and phonology are the subjects of continuing controversy today (see also LINGUISTICS: *Semantics*).

Language variants

The word language contains a multiplicity of different designations. Two senses have already been distinguished: language as a universal species-specific capability of mankind, and languages as the various manifestations of that capability, as with English, French, Latin, Swahili, Malay, and so on. There is, of course, no observable universal language over and above the various languages that have been or are spoken or written; but one may choose to concentrate on the general and even the universal features, characteristics, and components of different languages and on the ways in which the same sets of descriptive procedures and explanatory theories may be applied to different languages. In so doing one may refer to language (in general) as one's object of study. This is what is done by linguists, or linguistic scientists, persons devoting themselves to the scientific study of languages (as opposed to the popular sense of polyglots, persons having a command of several different languages).

DIALECTS

Intercomprehensibility of dialects

It has already been pointed out that no two persons speak exactly alike, and within the area of all but the smallest speech communities (groups of people speaking the same language) there are subdivisions of recognizably different types of language, called dialects, that do not, however, render intercommunication impossible nor markedly difficult. Because intercomprehensibility lies along a scale, the degree required for two or more forms of speech to qualify as dialects of a single language, instead of being regarded as separate languages, is not easy to quantify or to lay down in advance, and the actual cutoff point must in the last resort be arbitrary. In practice, however, the terms dialect and language can be used with reasonable agreement. One speaks of different dialects of English (Southern British English, Northern British English, Scottish English, Midwest American English, New England American English, Australian English, and so on, with, of course, many more delicately distinguished subdialects within these very general categories), but no one would speak of Welsh and English or of Irish and English as dialects of a single language, although they are spoken within the same areas and often by people living in the same villages as each other. (Ro.H.R./Ed.)

Varieties of dialects. *Geographic dialects.* The most widespread type of dialectal differentiation is geographic. As a rule, the speech of one locality differs at least slightly from that of any other place. Differences between neighbouring local dialects are usually small, but, in travelling farther in the same direction, differences accumulate. Every dialectal feature has its own boundary line, called an isogloss (or sometimes heterogloss). Isoglosses of various linguistic phenomena rarely coincide completely, and by crossing and interweaving they constitute intricate patterns on dialect maps. Frequently, however, several isoglosses are grouped approximately together into a bundle of isoglosses. This grouping is caused either by geographic obstacles that arrest the diffusion of a number of innovations along the same line or by historical circumstances, such as political borders of long standing, or by migrations that have brought into contact two populations whose dialects were developed in noncontiguous areas.

Local and regional dialects

Geographic dialects include local ones (*e.g.,* the Yankee English of Cape Cod or of Boston, the Russian of Moscow or of Smolensk) or regional ones, such as Delaware Valley English, Australian English, or Tuscan Italian. Such entities are of unequal rank; South Carolina English, for instance, is included in Southern American English. Regional dialects do have some internal variation, but the differences within a regional dialect are supposedly smaller than differences between two regional dialects of the same rank. In a number of areas ("linguistic landscapes") where the dialectal differentiation is essentially even, it is hardly justified to speak of regional dialects. This uniformity has led many linguists to deny the meaningfulness of such a notion altogether; very frequently, however, bundles of isoglosses—or even a single isogloss of major importance—permit the division of a territory into regional dialects (see Figure 1 for the dialectal division of American English in the Atlantic states). The public is often aware of such divisions, usually associating them with names of geographic regions or provinces, or with some feature of pronunciation; *e.g.,* Southern English or Russian *o*-dialects and *a*-dialects. Especially clear-cut cases of division are those in which geographic isolation has played the principal role; *e.g.,* Australian English or Louisiana French.

Adapted from H. Kurath, *Word Geography of the Eastern United States;* the University of Michigan Press

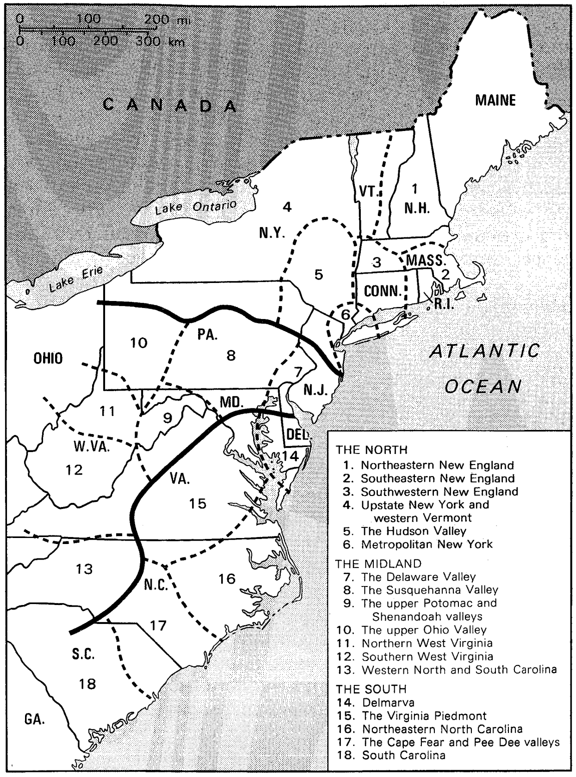

0 100 200 mi
0 100 200 300 km
CANADA
MAINE
Lake Ontario
Lake Erie
VT.
N.H.
N.Y.
MASS.
CONN.
R.I.
PA.
OHIO
N.J.
MD.
DEL.
W.VA.
VA.
N.C.
S.C.
GA.
ATLANTIC OCEAN
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
THE NORTH
1. Northeastern New England
2. Southeastern New England
3. Southwestern New England
4. Upstate New York and western Vermont
5. The Hudson Valley
6. Metropolitan New York
THE MIDLAND
7. The Delaware Valley
8. The Susquehanna Valley
9. The upper Potomac and Shenandoah valleys
10. The upper Ohio Valley
11. Northern West Virginia
12. Southern West Virginia
13. Western North and South Carolina
THE SOUTH
14. Delmarva
15. The Virginia Piedmont
16. Northeastern North Carolina
17. The Cape Fear and Pee Dee valleys
18. South Carolina

Figure 1: Dialect areas of the eastern United States.

Social dialects. Another important axis of differentiation is that of social strata. In many localities, dialectal differences are connected with social classes, educational levels, or both. More highly educated speakers and, often, those belonging to a higher social class tend to use more features belonging to the standard language, whereas the original dialect of the region is better preserved in the speech of the lower and less educated classes. In large urban centres, innovations unknown in the former dialect of the region frequently develop. Thus, in cities the social stratification of dialects is especially relevant and far-reaching, whereas in rural areas, with a conservative way of life, the traditional geographic dialectal differentiation prevails.

Technical terminology and slang

Educational differences among speakers strongly affect the extent of their vocabulary. In addition, practically every profession has its own expressions, which include the technical terminology and sometimes also the casual words or idioms peculiar to the group. Slang, too, is characterized mainly by a specific vocabulary and is much more flexible than an ordinary dialect, as it is subject to fashion and depends strongly on the speaker's age group. Slang—just as a professional dialect—is used mainly by persons who are in a sense bidialectal; *i.e.,* they speak some other dialect or the standard language, in addition to slang. Dialectal differences also often run parallel with the religious or racial division of the population.

Dialectal change and diffusion. The basic cause of dialectal differentiation is linguistic change. Every living language constantly undergoes changes in its various elements. Because languages are extremely complex systems of signs, it is almost inconceivable that linguistic evolution could affect the same elements and even transform them in the same way in all localities where one language is spoken and for all speakers in the same locality. At first glance, differences caused by linguistic change seem to be slight, but they inevitably accumulate with time (*e.g.,* compare Chaucer's English with modern English or Latin with modern Italian, French, Spanish, or Romanian). Related languages usually begin as dialects of the same language.

When a change (an innovation) appears among only one section of the speakers of a language, this automatically creates a dialectal difference. Sometimes an innovation in dialect A contrasts with the unchanged usage (archaism) in dialect B. Sometimes a separate innovation occurs in each of the two dialects. Of course, different innovations will appear in different dialects, so that, in comparison with its contemporaries, no one dialect as a whole can be considered archaic in any absolute sense. A dialect may be characterized as relatively archaic, because it shows fewer innovations than the others; or it may be archaic in one feature only.

After the appearance of a new dialectal feature, interaction between speakers who have adopted this feature and those who have not leads to the expansion or the curtailment of its area or even to its disappearance. In a single social milieu (generally the inhabitants of the same locality, generation, and social class), the chance of the complete adoption or rejection of a new dialectal feature is very great; the intense contact and consciousness of membership within the social group fosters such uniformity. When several age groups or social strata live within the same locality and especially when people speaking the same language live in separate communities, dialectal differences are easily maintained.

Natural barriers and language change

The element of mutual contact plays a large role in the maintenance of speech patterns; that is why differences between geographically distant dialects are normally greater than those between dialects of neighbouring settlements. This also explains why bundles of isoglosses so often form along major natural barriers—impassable mountain ranges, deserts, uninhabited marshes or forests, or wide rivers—or along political borders. Similarly, racial or religious differences contribute to linguistic differentiation because contact between members of one faith or race and those of another within the same area is very often much more superficial and less frequent than contact between members of the same racial or religious group. An especially powerful influence is the relatively infrequent occurrence of intermarriages, thus preventing dialectal mixture at the point where it is most effective; namely, in the mother tongue learned by the child at home.

Unifying influences on dialects. Communication lines such as roads (if they are at least several centuries old), river valleys, or seacoasts often have a unifying influence. Also, important urban centres, such as Paris, Utrecht, or Cologne, often form the hub of a circular region in which approximately the same dialect is spoken. In such areas, the prestige dialect of the city has obviously expanded. As a general rule, those dialects, or at least certain dialectal features, with greater social prestige tend to replace those that are valued lower on the social scale.

In times of less frequent contact between populations, dialectal differences increase; in periods of greater contact, they diminish. The general trend in modern times is for dialectal differences to diminish, above all through the replacement of dialectal traits by those of the standard language. Mass literacy, schools, increased mobility of populations, and, in the last few decades, the evergrowing role of mass communications all contribute to this tendency. Naturally, the extent of such unifying action varies greatly in different linguistic domains, depending on the level of civilization. Nevertheless, the most thorough example of linguistic force exerted by a single dominating civilization belongs to ancient times: in the Hellenistic era, almost all ancient Greek dialects were replaced by the so-called koine, based on the dialect of Athens.

Mass migrations may also contribute to the formation

Dialect differentiation and population movements

of a more or less uniform dialect over broad geographic areas. Either the resulting dialect is that of the original homeland of a particular migrating population or it is a dialect mixture formed by the levelling of differences among migrants from more than one homeland. The degree of dialectal differentiation depends to a great extent on the length of time a certain population has remained in a certain place. Thus, it is understandable that the diversification of the English language is far greater in the British Isles than, for example, in North America (especially if the number of dialectal differences is considered on a comparable area basis, such as how many per 1,000 square miles). In the U.S. itself much greater diversity is evident among dialects in old colonial America—along the Atlantic coast—than among dialects west of the Appalachians. It is also typical that phonological differences are more far-reaching in Switzerland among Swiss-German dialects than throughout the vast territory where the Russian language is spoken, extending from Leningrad to eastern Siberia. Such a situation results not only from migrations of the Russian population, (as compared to the centuries of Swiss stability) but also from the contrasting geographical configurations: in the U.S.S.R., there is unobstructed communication in all directions; in mountainous Switzerland, the territory is carved into small, isolated units.

Migrations and, more rarely, geographical phenomena may in some areas cause a much stronger dialectal differentiation in one direction than in others. Isoglosses in the U.S., for example, run predominantly in an east-west direction, reflecting the westward stream of migration during the colonization of areas west of the Appalachians. Similarly, the majority of isoglosses in Russia follow latitude, but in the opposite (west–east) direction.

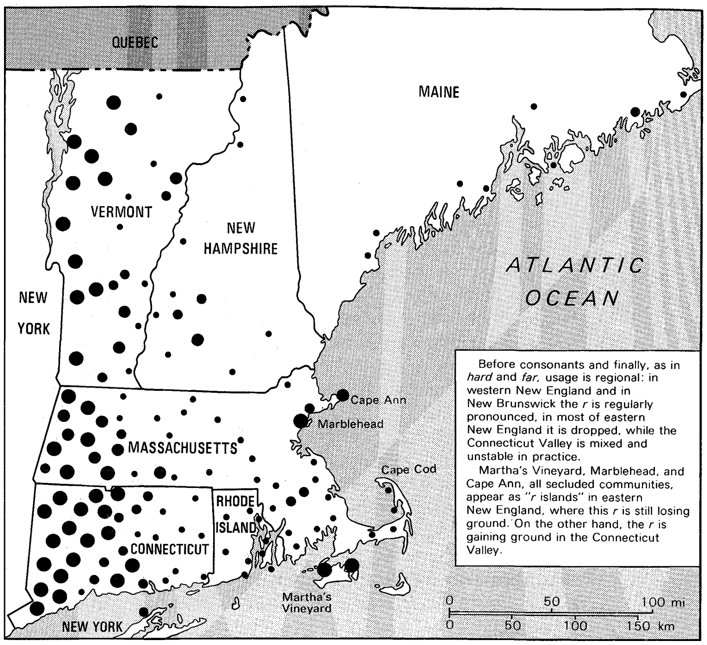
Focal, relic, and transitional areas. Dialectologists often distinguish between focal areas—which provide sources of numerous important innovations and usually coincide with centres of lively economic or cultural activity—and relic areas—places toward which such innovations are spreading but have not usually arrived. (Relic areas also have their own innovations, which, however, usually extend over a smaller geographical area.) Relic areas or relic phenomena are particularly common in out-of-the-way regional pockets or along the periphery of a particular language's geographical territory. An example of a focal area in the U.S. would be the Boston region, while rural Maine and New Hampshire and Cape Cod and Nantucket Island would be typical relic areas (see Figure 2).

The borders of regional dialects often contain transitional areas that share some features with one neighbour and some with the other. Such mixtures result from unequal diffusion of innovations from both sides. Similar unequal diffusion in mixed dialects in any region also may be a consequence of population mixture created by migrations.

In regions with many bilingual speakers (*e.g.*, along the border between two languages) dialects of both languages will often undergo changes influenced by the other tongue. This is manifested not only in numerous loanwords but often also in the adoption of phonological or grammatical features. Such phenomena are particularly frequent in a population that once spoke one language and only later adopted the second language. In extreme cases, a so-called creolized language develops. (Creoles are pidgin languages that have become the only or major language of a speech community. See LANGUAGES OF THE WORLD: *Pidgin*.)

Written form of dialects

Standard languages. Standard languages arise when a certain dialect begins to be used in written form, normally throughout a broader area than that of the dialect itself. The ways in which this language is used—in administrative matters, literature, economic life—lead to the minimization of linguistic variation. The social prestige attached to the speech of the richest, most powerful, and most highly educated members of a society transforms their language into a model for others; it also contributes to the elimination of deviating linguistic forms. Dictionaries and grammars help to stabilize linguistic norms, as do the activity of scholarly institutions and, sometimes, governmental intervention. The base dialect for a country's standard language is very often the original dialect of its capital—in France, Paris; in England, London; in Russia, Moscow. Or the base may be a strong economic and cultural centre—in Italy, Florence. Or the language may be a combination of several regional dialects; *e.g.*, German or Polish.



Before consonants and finally, as in *hard* and *far,* usage is regional: in western New England and in New Brunswick the *r* is regularly pronounced, in most of eastern New England it is dropped, while the Connecticut Valley is mixed and unstable in practice.

Martha's Vineyard, Marblehead, and Cape Ann, all secluded communities, appear as "*r* islands" in eastern New England, where this *r* is still losing ground. On the other hand, the *r* is gaining ground in the Connecticut Valley.

Figure 2: New England pronunciation of preconsonantal and final *r*. The largest circles indicate regular use of this *r*, the smallest ones sporadic use, and the two intermediate sizes rather evenly divided usage.

Adapted from A.J. Bronstein, *The Pronunciation of American English—An Introduction to Phonetics* (1960), Appleton-Century-Crofts; with permission from the American Council of Learned Societies

Even a standard language that was originally based on one local dialect changes, however, as elements of other dialects infiltrate into it over the years. The actual development in any one linguistic area depends on historical events. Sometimes even the distribution of standard languages may not correspond to the dialectal situation. Dutch and Flemish dialects are a part of the Low German dialectal area, which embraces all of northern Germany, as well as the Netherlands and part of Belgium. In one part of the dialectal area, however, the standard language is based on High German, and, in the other part, the standard language is Dutch or Flemish, depending on the nationality of the respective populations. In the U.S., where there is no clearly dominant political or cultural centre—such as London or Paris—and where the territory is enormous, the so-called standard language shows perceptible regional variations in pronunciation.

In most developed countries, the majority of the population has an active (speaking, writing) or at least passive (understanding) command of the standard language. Very often the rural population, and not uncommonly the lower social strata of the urban population as well, are in reality bidialectal. They speak their maternal dialect at home and with friends and acquaintances in casual contacts, and they use the standard language in more formal situations. Even the educated urban population in some regions uses the so-called colloquial language informally. In the German-, Czech-, and Slovene-speaking areas of middle Europe, for example, a basically regional dialect from which the most striking local features have been eliminated is spoken. The use of this type of language is supported by psychological factors, such as feelings of solidarity with a certain region and pride in its traditions or the relaxed mood connected with informal behaviour. (P.I./Ed.)

SLANG

The above-mentioned social and regional dialect variations within languages are more or less natural; that is, they arise largely from the conditions of language use and language transmission, without deliberate intent on the part of speakers. There are, however, some deliberately created variations within languages. The socially unifying force of a single language or dialect and the divisive force of language and dialect differences have always been apparent. Language was a main inspiration in 19th-century European nationalism and is one of the factors in nationalist movements in the world today. Not surprisingly, groups within a society that set a special value on group identity and group consciousness deliberately develop and foster private dialects that are known to insiders but are mysterious and baffling to those not belonging to the group. The various underworld jargons and special trade argots can be cited, such as Loucherbème in Paris (from *boucher* "butcher," by systematic deformation), the private slang vocabulary of many schools and colleges, and the systematic and regular alteration of certain words in some private languages of schoolboys and students—American "pig Latin" and the former "Oxford" and British public-school "-agger" talk ("nogger," meaning agnostic; "wagger pagger bagger," meaning wastepaper basket), developed from universally recognized forms such as "rugger," Rugby football. From all over the world the transient but zealously cherished speech styles of teenagers can be noted, whereby, along with particular fashions in clothes, they maintain and assert their identity over against the childhood they have left and the adult world they do not yet recognize. (Ed.)

Group identity and private dialects

Development of slang. Slang emanates from conflicts in values, sometimes superficial, often fundamental. When an individual applies language in a new way to express hostility, ridicule, or contempt, often with sharp wit, he may be creating slang, but the new expression will perish unless it is picked up by others. If the speaker is a member of a group that finds that his creation projects the emotional reaction of its members toward an idea, person, or social institution, the expression will gain currency according to the unanimity of attitude within the group. A new slang term is usually widely used in a subculture before it appears in the dominant culture. Thus slang—*e.g.,* "sucker," "honkey," "shave-tail," "jerk"—expresses the attitudes, not always derogatory, of one group or class toward the values of another. Slang sometimes stems from within the group, satirizing or burlesquing its own values, behaviour, and attitudes; *e.g.,* "shotgun wedding," "cake eater," "greasy spoon." Slang, then, is produced largely by social forces rather than by an individual speaker or writer who, single-handed (like Horace Walpole, who coined "serendipity" more than 200 years ago), creates and establishes a word in the language. This is one reason why it is difficult to determine the origin of slang terms.

The specialized languages of subcultures

Creators of slang. Civilized society tends to divide into a dominant culture and various subcultures that flourish within the dominant framework. The subcultures show specialized linguistic phenomena, varying widely in form and content, that depend on the nature of the groups and their relation to each other and to the dominant culture. The shock value of slang stems largely from the verbal transfer of the values of a subculture to diametrically opposed values in the dominant culture. Names such as fuzz, pig, fink, bull, and dick for policemen were not created by officers of the law. (The humorous "dickless tracy," however, meaning a policewoman, *was* coined by male policemen.)

Occupational groups are legion, and while in most respects they identify with the dominant culture, there is just enough social and linguistic hostility to maintain group solidarity. Terms such as scab, strike-breaker, company-man, and goon were highly charged words in the era in which labour began to organize in the United States; they are not used lightly even today, though they have been taken into the standard language.

In addition to occupational and professional groups, there are many other types of subcultures that supply slang. These include sexual deviants, narcotic addicts, ghetto groups, institutional populations, agricultural subsocieties, political organizations, the armed forces, Gypsies, and sports groups of many varieties. Some of the most fruitful sources of slang are the subcultures of professional criminals who have migrated to the New World since the 16th century. Old-time thieves still humorously refer to themselves as FFV—First Families of Virginia.

In criminal subcultures, pressure applied by the dominant culture intensifies the internal forces already at work, and the argot forming there emphasizes the values, attitudes, and techniques of the subculture. Criminal groups seem to evolve about this specialized argot, and both the subculture and its slang expressions proliferate in response to internal and external pressures.

Sources. Most subcultures tend to draw words and phrases from the contiguous language (rather than creating many new words) and to give these established terms new and special meanings; some borrowings from foreign languages, including the American Indian tongues, are traditional. The more learned occupations or professions like medicine, law, psychology, sociology, engineering, and electronics tend to create true neologisms, often based on Greek or Latin roots, but these are not major sources for slang, though nurses and medical students adapt some medical terminology to their slang, and air force personnel and some other branches of the armed services borrow freely from engineering and electronics.

Linguistic processes forming slang. The processes by which words become slang are the same as those by which other words in the language change their form or meaning or both. Some of these are the employment of metaphor, simile, folk etymology, distortion of sounds in words, generalization, specialization, clipping, the use of acronyms, elevation and degeneration, metonymy, synecdoche, hyperbole, borrowings from foreign languages, and the play of euphemism against taboo. The English word trip is an example of a term that has undergone both specialization and generalization. It first became specialized to mean a psychedelic experience resulting from the drug LSD. Subsequently, it generalized again to mean any experience on any drug, and beyond that to any type of "kicks" from anything. Clipping is exemplified by the use of "grass" from "laughing grass," a term for marijuana. "Funky," once a very low term for body odour, has undergone elevation among jazz buffs to signify "the best"; "fanny," on the other hand, once simply a girl's name, is currently a degenerated term that refers to the buttocks (in England, it has further degenerated into a taboo word for the female genitalia). There is also some actual coinage of slang terms.

Characteristics of slang. Psychologically, most good slang harks back to the stage in human culture when animism was a worldwide religion. At that time, it was believed that all objects had two aspects, one external and objective that could be perceived by the senses, the other imperceptible (except to gifted individuals) but identical with what we today would call the "real" object. Human survival depended upon the manipulation of all "real" aspects of life—hunting, reproduction, warfare, weapons, design of habitations, nature of clothing or decoration, etc.—through control or influence upon the *animus,* or imperceptible phase of reality. This influence was exerted through many aspects of sympathetic magic, one of the most potent being the use of language. Words, therefore, had great power, because they evoked the things to which they referred.

Primitive origins of slang

Civilized cultures and their languages retain many remnants of animism, largely on the unconscious level. In Western languages, the metaphor owes its power to echoes of sympathetic magic, and slang utilizes certain attributes of the metaphor to evoke images too close for comfort to "reality." For example, to refer to a woman as a "broad" is automatically to increase her girth in an area in which she may fancy herself as being thin. Her reaction may, thus, be one of anger and resentment, if she happens to live in a society in which slim hips are considered essential to feminine beauty. Slang, then, owes much of its power to shock to the superimposition of images that are incongruous with images (or values) of others, usually members of the dominant culture. Slang is most popular when its imagery develops incongruity bordering on social satire. Every slang word, however, has its own history and reasons for popularity. When conditions change, the term may change in meaning, be adopted into the standard language, or continue to be used as slang within certain enclaves of the population. Nothing is flatter than dead slang. In 1910, for instance, "Oh you kid" and "23-skiddoo" were quite stylish phrases in the U.S. but they have gone with the hobble skirt. Children, however, unaware of anachronisms, often revive old slang under a barrage of older movies rerun on television.

Some slang becomes respectable when it loses its edge; "spunk," "fizzle," "spent," "hit the spot," "jazz," "funky," and "p.o.'d," once thought to be too indecent for feminine ears, are now family words. Other slang survives for centuries, like "bones" for dice (Chaucer), "beat it" for run away (Shakespeare), "duds" for clothes, and "booze" for liquor (Dekker). These words must have been uttered as slang long before appearing in print, and they have remained slang ever since. Normally, slang has both a high birth and death rate in the dominant culture, and excessive use tends to dull the lustre of even the most colourful and descriptive words and phrases. The rate of turnover in slang words is undoubtedly encouraged by the mass media, and a term must be increasingly effective to survive.

Transience and persistence

While many slang words introduce new concepts, some of the most effective slang provides new expressions—fresh, satirical, shocking—for established concepts, often very respectable ones. Sound is sometimes used as a basis for this type of slang, as, for example, in various phonetic distortions (*e.g.,* pig Latin terms). It is also used in rhyming slang, which employs a fortunate combination of both sound and imagery. Thus, gloves are "turtledoves" (the gloved hands suggesting a pair of billing doves), a girl is a "twist and twirl" (the movement suggesting a girl walking), and an insulting imitation of flatus, produced by blowing air between the tip of the protruded tongue and the upper lip, is the "raspberry," cut back from "raspberry tart." Most slang, however, depends upon incongruity of imagery, conveyed by the lively connotations of a novel term applied to an established concept. Slang is not all of equal quality, a considerable body of it reflecting a simple need to find new terms for common ones, such as the hands, feet, head, and other parts of the body. Food, drink, and sex also involve extensive slang vocabulary. Strained or synthetically invented slang lacks verve, as can be seen in the desperate efforts of some sportswriters to avoid mentioning the word baseball—*e.g.,* a batter does not hit a baseball but rather "swats the horsehide," "plasters the pill," "hefts the old apple over the fence," and so on.

The most effective slang operates on a more sophisticated level and often tells something about the thing named, the person using the term, and the social matrix against which it is used. Pungency may increase when full understanding of the term depends on a little inside information or knowledge of a term already in use, often on the slang side itself. For example, the term Vatican roulette (for the rhythm system of birth control) would have little impact if the expression Russian roulette were not already in wide usage.

Diffusion of slang. Slang invades the dominant culture as it seeps out of various subcultures. Some words fall dead or lie dormant in the dominant culture for long periods. Others vividly express an idea already latent in the dominant culture and these are immediately picked up and used. Before the advent of mass media, such terms invaded the dominant culture slowly and were transmitted largely by word of mouth. Thus a term like snafu, its shocking power softened with the explanation "situation normal, all fouled up," worked its way gradually from the military in World War II by word of mouth (because the media largely shunned it) into respectable circles. Today, however, a sportscaster, news reporter, or comedian may introduce a lively new word already used by an in-group into millions of homes simultaneously, giving it almost instant currency. For example, the term uptight was first used largely by criminal narcotic addicts to indicate the onset of withdrawal distress when drugs are denied. Later, because of intense journalistic interest in the drug scene, it became widely used in the dominant culture to mean anxiety or tension unrelated to drug use. It kept its form but changed its meaning slightly.

Slang and the mass media

Other terms may change their form or both form and meaning, like "one for the book" (anything unusual or unbelievable). Sportswriters in the U.S. borrowed this term around 1920 from the occupational language of then legal bookmakers, who lined up at racetracks in the morning ("the morning line" is still figuratively used on every sports page) to take bets on the afternoon races. Newly arrived bookmakers went to the end of the line, and any bettor requesting unusually long odds was motioned down the line with the phrase, "That's one for the end book." The general public dropped the "end" as meaningless, but old-time gamblers still retain it. Slang spreads through many other channels, such as popular songs, which, for the initiate, are often rich in double entendre.

When subcultures are structurally tight, little of their language leaks out. Thus the Mafia, in more than a half-century of powerful criminal activity in America, has contributed little slang. When subcultures weaken, contacts with the dominant culture multiply, diffusion occurs, and their language appears widely as slang. Criminal narcotic addicts, for example, had a tight subculture and a highly secret argot in the 1940s; now their terms are used freely by middle-class teenagers, even those with no real knowledge of drugs.

Uses of slang. Slang is used for many purposes, but generally it expresses a certain emotional attitude; the same term may express diametrically opposed attitudes when used by different people. Many slang terms are primarily derogatory, though they may also be ambivalent when used in intimacy or affection. Some crystallize or bolster the self-image or promote identification with a class or in-group. Others flatter objects, institutions, or persons but may be used by different people for the opposite effect. "Jesus freak," originally used as ridicule, was adopted as a title by certain street evangelists. Slang sometimes insults or shocks when used directly; some terms euphemize a sensitive concept, though obvious or excessive euphemism may break the taboo more effectively than a less decorous term. Some slang words are essential because there are no words in the standard language expressing exactly the same meaning; *e.g.,* "freak-out," "barn-storm," "rubberneck," and the noun "creep." At the other extreme, multitudes of words, vague in meaning, are used simply as fads.

Slang expressing emotional attitudes

There are many other uses to which slang is put, according to the individual and his place in society. Since most slang is used on the spoken level, by persons who probably are unaware that it is slang, the choice of terms naturally follows a multiplicity of unconscious thought patterns. When used by writers, slang is much more consciously and carefully chosen to achieve a specific effect. Writers, however, seldom invent slang.

It has been claimed that slang is created by ingenious individuals to freshen the language, to vitalize it, to make the language more pungent and picturesque, to increase the store of terse and striking words, or to provide a vocabulary for new shades of meaning. Most of the originators and purveyors of slang, however, are probably not conscious of these noble purposes and do not seem overly concerned about what happens to their language.

Attitudes toward slang. With the rise of naturalistic writing demanding realism, slang began to creep into English literature even though the schools waged warfare against it, the pulpit thundered against it, and many women who aspired to gentility and refinement banished it from the home. It flourished underground, however, in such male sanctuaries as lodges, poolrooms, barbershops, and saloons.

By 1925 a whole new generation of U.S. and European naturalistic writers was in revolt against the Victorian restraints that had caused even Mark Twain to complain, and today any writer may use slang freely, especially in fiction and drama. It has become an indispensable tool in the hands of master satirists, humorists, and journalists. Slang is now socially acceptable, not just because it is slang but because, when used with skill and discrimination, it adds a new and exciting dimension to language. At the same time, it is being seriously studied by linguists and other social scientists as a revealing index to the culture that produces and uses it. (D.W.M./Ed.)

Slang in literature

OTHER SPECIALIZED LANGUAGES

Jargon. Sometimes, as in the case of criminal argots, part of the function of special languages is deliberately to mislead and obstruct the rest of society and the authorities in particular; they may even become wholly impenetrable to outsiders. But this is not the sole or main purpose of most specialized varieties of language. Professions whose members value their standing in society and are eager to render their services to the public foster their own vocabulary and usage, partly to enhance the dignity of their profession and the skills they represent but partly also to increase their efficiency. An example of this is the language of the law and of lawyers.

The cultivation and maintenance of specialized types of language by certain professions should not be regarded as trivially or superficially motivated. In general usage, languages are necessarily imprecise, or they would lack the flexibility and infinite extensibility demanded of them. But for certain purposes in restricted situations much greater precision is required, and part of the function of the particular style and vocabulary of legal language is the avoidance, so far as may be possible, of all ambiguity and the explicit statement of all necessary distinctions. This is why legal texts, when read out of their context, seem so absurdly pedantic and are an easy target for ridicule. Similar provision for detail and clarity characterizes the specialist jargons of medicine and of the sciences in general and also of philosophy. Indeed, one might regard the formulas of modern symbolic logic as the result of a consciously developed and specialized written language for making precise the relations of implication and inference between statements that, when couched in everyday language, are inexact and open to misinterpretation. Some would go as far as to say that traditional metaphysics is no more than the result of misunderstanding everyday discourse and that the main purpose of philosophy is to resolve the puzzles that arise from such misunderstandings.

Stereotyped language of games

The use of specialized types of language in fostering unity is also evidenced in the stereotyped forms of vocabulary employed in the playing of certain games. Tennis scores use the sequence "love, 15, 30, 40, and game"; cricketers verbally appeal to the umpire when a batsman may be out by calling "How's that?" and the ways of being out are designated by stereotypes, "run out," "leg before wicket," "stumped," and so forth. The esoteric language of horse racing and its associated wagering of money is well known, though not readily understood by outsiders.

The ancient but persistent recognition of the power of language is apparent in the respect for correctness in the use of language in any sphere of life having supernatural connections. Those credited with such connections employ special formulas and rigidly prescribed modes of diction; examples of the language of magic and of magicians are widespread, ranging from the usages of shamans and witch doctors to the ritual "abracadabra" of the mock magic displayed by conjurors at children's parties.

The efficacy of religious worship and of prayers is frequently associated with the strict maintenance of correct forms of language, taught by priests to their successors, lest the ritual become invalid. In ancient India the preservation in all its supposed purity of the language used in the performance of certain religious rituals (Sanskrit) gave rise to one of the world's most important schools of linguistics and phonetics. In the Christian churches one can observe the value placed by Church of England and Episcopalian churchmen on the formal English of the Authorized Version of the Bible and of *The Book of Common Prayer,* despite recent attempts at replacing these ritual forms of language by forms taken from modern spoken vernaculars.

Pidgins and creoles. Some specialized languages were developed to keep the outsider at bay. In other circumstances, languages have been deliberately created to facilitate communication with outsiders. This happens when people speaking two different languages have to work together, usually in some form of trade relation or administrative routine. In such situations the so-called pidgins arise, more or less purposively made up of vocabulary items from each language, with mutual abandonment of grammatical complexities that would cause confusion to either party. Pidgins have been particularly associated with areas settled by European traders; examples have been Chinook Jargon, a lingua franca based on an American Indian language and English and formerly used in Washington and Oregon, and Beach-la-mar, an English-based pidgin of parts of the South Seas.

Creolization of a pidgin language

Sometimes, as the result of relatively permanent settlement and the intermixture of two speech communities, a pidgin becomes the first language, or mother tongue, of later generations, ultimately displacing both the original languages. First languages arising in this way from artificially created pidgins are called creoles. Notable among creoles is the language of Haiti, Haitian Creole, built up from the French of the settlers and the African language of the former slaves; it shows lexical and grammatical features of both sources.

Creoles differ from pidgins in that, as first languages, they are subject to the natural processes of change like any other language (see below *Linguistic change*); and, despite the deliberately simplified form of the original pidgin, in the course of generations creoles develop their own complexities. The reason is plain to see. The restricted uses to which pidgins were first put and for which they were devised did not require any great flexibility. Once such a language becomes the first or only language of many people, it must perforce acquire the resources (*i.e.,* the complexity) to respond adequately to all the requirements of a natural language (see also LANGUAGES OF THE WORLD).

Nonverbal language. Speech and writing are, indeed, the fundamental faculties and activities referred to by the term language. There are, however, areas of human behaviour for which the term is used in a peripheral and derivative sense.

When individuals speak, they do not normally confine themselves to the mere emission of speech sounds. Because speaking usually involves at least two parties in sight of each other, a great deal of meaning is conveyed by facial expression, tone of voice, and movements and postures of the whole body but especially of the hands; these are collectively known as gestures. The contribution of bodily gestures to the total meaning of a conversation is in part culturally determined and differs in different communities. Just how important these visual symbols are may be seen when one considers how much less effective telephone conversation is as compared with conversation face to face; the experience of involuntarily smiling at the telephone receiver and immediately realizing that this will convey nothing to the hearer at the other end of the line is common. Again, the part played in emotional contact and in the expression of feelings by facial expressions and tone of voice, quite independently of the words used, has been shown in tests in which subjects have been asked to react to sentences that appear as friendly and inviting when read but are spoken angrily and, conversely, to sentences that appear as hostile but are spoken with friendly facial expressions. It is found that it is the visual accompaniments and tone of voice that elicit the main emotional response. A good deal of what goes under the heading of sarcasm exploits these contrasts.

Expression by vocal, nonlinguistic utterances

Just as there are paralinguistic activities such as facial expressions and bodily gestures integrated with and assisting the communicative function of spoken language, so there are vocally produced noises that cannot be regarded as part of any language, though they help in communication and in the expression of feeling. These include laughter, shouts and screams of joy, fear, pain, and so forth, and conventional expressions of disgust, triumph, and so on, traditionally spelled "ugh!," "ha ha!," etc., in English. Such nonlexical ejaculations differ in important respects from language: they are much more similar in form and meaning throughout mankind as a whole, in contrast to the great diversity of languages; they are far less arbitrary than most of the lexical components of language; and they are much nearer the cries of animals produced under similar circumstances and, as far as is known, serve similar expressive and communicative purposes. As noted above, some people have tried to trace the origin of language itself to them.

A language is a symbol system. It may be regarded, because of its infinite flexibility and productivity, as the symbol system *par excellence.* But there are other symbol systems recognized and institutionalized in the different cultures of mankind. Examples of these exist on maps and blueprints and in the conventions of representational art (*e.g.,* the golden halos around the heads of saints in religious paintings). Other symbol systems are musical notation and dance notation, wherein graphic symbols designate musical pitches and other features of musical performance and the movements of formalized dances. More loosely, because music itself can convey and arouse emotions and certain musical forms and structures are often associated with certain types of feeling, one frequently reads of the "language of music" or even of "the grammar of music." The terms language and grammar are here being used metaphorically, however, if only because no symbol system other than language has the same potential of infinite productivity, extension, and precision.

Languages are used by human beings to talk and write to other human beings. Derivatively, bits of languages may be used by humans to control machinery, as when different buttons and switches are marked with words or phrases designating their functions. A recent and specialized development of man-machine language is seen in the various "computer languages" now in use; *e.g.,* Cobol, Algol, and Fortran. These are referred to as programming languages, and they provide the means whereby sets of "instructions" and data of various kinds can be supplied to computers in forms acceptable to these machines. Various types of such languages are in use for different purposes. The development and use of computer languages must now be regarded as a distinct science in itself (for more information, see INFORMATION PROCESSING AND INFORMATION SYSTEMS: *Programming systems and programming language categories*).

Physiological and physical basis of speech

Primacy of speech in language

For an adequate understanding of human language it is necessary to keep in mind the absolute primacy of speech. In societies in which literacy is all but universal and language teaching at school begins with reading and writing in the mother tongue, one is apt to think of language as a writing system that may be pronounced. In point of fact, language is a system of spoken communication that may be represented in various ways in writing.

Man has almost certainly been in some sense a speaking animal from early in the emergence of *Homo sapiens* as a recognizably distinct species. The earliest known systems of writing go back perhaps some 5,000 years. This means that for many hundreds of thousands of years human languages were transmitted from generation to generation and were developed entirely as spoken means of communication. Moreover, in the world as it is today, literacy is still the privilege of a minority in many language communities. Even when literacy is widespread, some languages remain unwritten if they are not economically or culturally important enough to justify creating an alphabet for them and teaching them; then literacy is acquired in a second language learned at school. Such is the case with many speakers of South American Indian languages, who become literate in Spanish or Portuguese. A similar situation prevails in some parts of Africa, where reading and writing are taught in languages spoken over relatively wide areas. In all communities, speaking is learned by children before writing, and all people act as speakers and hearers much more than as writers and readers.

It is, moreover, a total fallacy to suppose that the languages of illiterate or so-called primitive peoples are less structured, less rich in vocabulary, and less efficient than the languages of literate civilizations. The lexical content of languages varies, of course, according to the culture and the needs of their speakers, but observation bears out the statement of the U.S. anthropological linguist Edward Sapir made in 1921: "When it comes to linguistic form, Plato walks with the Macedonian swineherd, Confucius with the head-hunting savage of Assam."

All this means that the structure and composition of language and of all languages have been conditioned by the requirements of speech, not those of writing. Languages are what they are by virtue of their spoken, not their written, manifestations. The study of language must be based on a knowledge of the physiological and physical nature of speaking and hearing. The details of these aspects of language are covered in SPEECH: *The phonetics of speech* and *The physiology of speech*; only the essentials are given here.

SPEECH PRODUCTION

Speaking is in essence the by-product of a necessary bodily process, the expulsion from the lungs of air charged with carbon dioxide after it has fulfilled its function in respiration. Most of the time one breathes out silently; but it is possible, by adopting various postures and by making various movements within the vocal tract, to interfere with the egressive airstream so as to generate noises of different sorts. This is what speech is made of.

Vocal tract and speech organs

The vocal tract comprises the passage from the trachea (windpipe) to the orifices of the mouth and nose; all the organs used in speaking lie in this passage. Conventionally, these are called the organs of speech, and the use in several languages of the same word for the tongue as a part of the body and for language shows the awareness people have of the role played by this part of the mouth in speaking. But few if any of the major organs of speech are exclusively or even mainly concerned with speaking. The lips, the tongue, and the teeth all have essential functions in man's bodily economy, quite apart from talking; to think, for example, of the tongue as an organ of speech in the same way that the stomach is regarded as the organ of digestion is fallacious. Speaking is a function superimposed on these organs, and the material of speech is a waste product, spent air, exploited to produce perhaps the most wonderful by-product ever created.

Relatively few types of speech sounds are produced by other sources of air movement; the clicks in some South African languages are examples, and so is the fringe linguistic sound used in English to express disapproval, conventionally spelled "tut." In all languages, however, the great majority of speech sounds have their origin in air expelled through the contraction of the lungs. Air forced through a narrow passage or momentarily blocked and then released creates noise, and characteristic components of speech sounds are types of noise produced by blockage or narrowing of the passage at different places.

If the vocal cords (really more like two curtains) are held taut as the air passes through them, the resultant regular vibrations in the larynx produce what is technically called voice, or voicing. These vibrations can be readily observed by contrasting the sounds of *f* and *v* or of *s* and *z* as usually pronounced; "five" and "size" each begin and end with voiceless and voiced sounds, respectively, which are otherwise formed alike, with the tongue and the lips in the same position. Most consonant sounds and all vowel sounds in English and in the majority of languages are voiced, and voice, in this sense, is the basis of singing and of the rise and fall in speaking that is called intonation, as well as of the tone distinctions in tone languages. The vocal cords may be drawn together more or less tightly, and the vibrations will be correspondingly more or less frequent. A rise in frequency causes a rise in perceived vocal pitch. Speech in which voice is completely excluded is called whispering.

Above the larynx, places of articulation in frequent use are between the back of the tongue and the soft palate, between the blade of the tongue and the ridge just behind the upper front teeth, and between the lips. Stoppage and release (technically, plosion) at these places form the *k* (often written as *c,* "cat"), *t,* and *p* sounds in English and, when voicing is also present, the *g* (as in "gay"), *d,* and *b* sounds. Obstruction at these and other places sufficient to cause noise gives rise to what are called fricative sounds; in English these include the normal pronunciations of *s, z, f,* and *v* and the *th* sounds in "thin" and "then." A vowel is characterized as the product of the shape of the entire tract between the lips and larynx, without local obstruction though usually with voicing from the vocal

Formation of vowels

cords. It is contrasted with a consonant, though the exact division between these two categories of speech sound is not always easy to draw. Different shaping of the tract produces the different vowel sounds of languages.

The soft plate may be raised or lowered. It is lowered in normal breathing and allows air to pass in and out through the nose. In the utterance of most speech sounds it is raised, so that air passing through the mouth alone forms the sound; if it is lowered, air passes additionally or alternatively through the nose, thus producing the nasal sounds. All but a very few languages have nasal consonants (the English sounds *m, n,* and *ng* as in "sing") and some, such as French, have nasalized vowels as well. A few speakers regularly allow air to pass through their nasal passages all the time while they are speaking; such persons are said to "speak through the nose."

All articulatory movements, including the initial expulsion of air from the lungs, may be made with greater or less vigour, giving rise to louder or softer speech as a whole or to greater loudness on one part of what is said for emphasis or contrast.

Every different configuration and movement of the vocal tract creates corresponding differences in the air vibrations that comprise and transmit sound. These vibrations, like those of all noises, extend outward in all directions from the source, gradually decreasing to zero or to below the threshold of audibility. They are called sound waves, and they consist of rapid rises and falls in air pressure. The speed at which pressure rises and falls is the frequency. Speech sounds involve complex waves containing vibrations at a number of different frequencies, the lowest being the voice pitch of singing and intonation, produced by the vocal cords in voiced sounds.

Articulatory and acoustic descriptions of speech

The eardrum responds to the different frequencies of speech, provided they retain enough energy, or amplitude (*i.e.,* are still audible). The perceptibly different speech sounds that comprise the spoken utterances of any language are the result of the different impacts on one's ears made by the different complexes of frequencies in the waves produced by different articulatory processes. As the result of careful and detailed observation of the movements of the vocal organs in speaking, aided by various instruments to supplement the naked eye, a great deal is now known about the processes of articulation. In recent years an array of other instruments has provided much information about the nature of the sound waves produced by articulation. Speech sounds may be and have been described and classified both from an articulatory viewpoint, in terms of how they are produced, and from an acoustic viewpoint, by reference to the resulting sound waves (their frequencies, amplitudes, and so forth). Articulatory descriptions are more readily understood, being couched in terms such as nasal, bilabial lip-rounded, and so on. Acoustic terminology requires a knowledge of the technicalities involved for its comprehension. In that almost every person is a speaker and a hearer, it is clear that both sorts of description and classification are important, and each has its particular value for certain parts of the scientific study of language.

LANGUAGE ACQUISITION

As far as the production of speech sounds is concerned, all human beings are physiologically alike. People have differently shaped faces, as much as they differ in other aspects of bodily build, but it has been shown time and again that a child learns to speak the language of those who bring him up from infancy. In most cases these are his biological parents, especially his mother, but one's first language is acquired from environment and learning, not from physiological inheritance. Adopted infants, whatever their race or physical type and whatever the language of their actual parents, acquire the language of the adoptive parents who raise them just as if they had been their own children.

Different shapes of lips, throat, and other parts of the vocal tract have an effect on the voice quality of people's speech, and this is part of the individuality of each person's voice referred to above. Physiological differences, including size of throat and larynx, both overall and in relation to the rest of the vocal tract, are largely responsible for the different pitch ranges characteristic of men's, women's, and children's speech. But these individual differences do not affect one's ability or aptitude to acquire and speak any particular language.

Speech is species-specific to mankind. Physiologically, animal communications systems are of all sorts. The type that is in some ways functionally the nearest to human language, the system of bee dances, is entirely remote from human speech and does not make use of sound at all. The animal sounds superficially most resembling speech, the imitative cries of parrots and some other birds, are produced by very different physiological means: birds have no teeth or lips but vocalize by means of the syrinx, a modification of the windpipe above the lungs. Almost all mammals and many other animal species make vocal noises and evince feelings thereby and keep in contact with each other through a rudimentary sort of communication, but those members of the animal kingdom nearest to man genetically, the primates, have proved highly resistant to the acquisition of speech.

Advantages of speech as a means of communication

Man's development of speech has been linked to his upright posture and the freeing of his vocal cords from the frequent need to "hold one's breath" in using the arms for locomotion. Certainly, speaking and hearing—as man's primary means of communication—have a number of striking advantages: speech does not depend on daylight or on mutual visibility, it can operate in all directions over reasonably wide areas, and it can be adjusted in loudness to cope with distance. As is seen in crowded rooms, it is possible to pick out some one person's voice despite a good deal of other noise and in the midst of other voices speaking the same language. In addition, the physical energy required in speaking is extremely small in relation to the immense power wielded by speech in man's life, and scarcely any other activity, such as running, walking, or tool using, interferes seriously with the process (even eating and drinking can be carried on simultaneously with talking, and the reluctance one may feel for "speaking with one's mouth full" is more a matter of cultural convention and good manners than of physiological difficulty).

The characteristics just outlined pertain to all of the world's languages, including those of the allegedly primitive peoples. What is more a matter of controversy is the extent to which man's biological inheritance is involved in language acquisition and language use. The fact that language is species-specific to man argues an essential cerebral or mental component, and in the last century certain aspects of speech control and use were located in a particular part of the human brain (Broca's convolution).

No one inherits the ability to speak a particular language, but every normal human child is born with the ability and the drive to acquire a language, namely, the one to which he is predominantly exposed from infancy. He brings to this task a very considerable innate ability, because his exposure is largely to a random selection of utterances (apart from any attempts at systematic teaching that he may encounter) occuring within his earshot or addressed to him. Yet by late childhood he has, through progressive stages, acquired the central or basic vocabulary of the language, together with its phonological and grammatical structure. Observation shows that this is substantially the same situation the world over, among literate and illiterate communities, and that much the same number of years of childhood is taken up by the process. Thus, it would appear that, objectively considered, all languages are roughly equal in complexity and in difficulty of mastery.

Innate human capacity for language acquisition

It is, therefore, clear that all normal humans bring into the world an innate faculty for language acquisition, language use, and grammar construction. The last phrase refers to the internalization of the rules of the grammar of one's first language from a more or less random exposure to utterances in it. The human child is very soon able to construct new, grammatically acceptable sentences from material he has already heard; unlike the parrot in human society, he is not limited to the mere repetition of whole utterances.

What is currently under debate is the part played by this innate ability and its exact nature. Until the 1950s scholars considered language acquisition to be carried out largely by analogical creation from observed patterns of sentences occurring in utterances heard and understood by the child. Such a view, much favoured by persons inclined to a behaviourist interpretation of human learning processes (*e.g.,* the U.S. linguist Leonard Bloomfield), stressed the very evident differences between the structures of different languages, particularly on the surface. Since the late 1950s, a number of linguists have been placing much more emphasis on the inherent grammar-building disposition and competence of the human brain, which is activated by exposure to utterances in a language, especially during childhood, in such a way that it fits the utterances into predetermined general categories and structures. Such linguists, inheritors of the 17th- and 18th-century interest in "universal grammar," put their stress on the underlying similarities of all languages, more especially in the deeper areas of grammatical analysis (for the distinction between deep structure and surface structure in grammar, see the article LINGUISTICS).

Meaning and style in language

The whole object and purpose of language is to be meaningful. Languages have developed and are constituted in their present forms in order to meet the needs of communication in all its aspects.

It is because the needs of human communication are so various and so multifarious that the study of meaning is probably the most difficult and baffling part of the serious study of language. Traditionally, language has been defined, as in the definition quoted above, as the expression of thought, but, as was seen, this involves far too narrow an interpretation of language or far too wide a view of thought to be serviceable. The expression of thought is just one among the many functions performed by language in certain contexts.

TYPES OF MEANING

Distinguishing meaning by word order or grammatical form

Structural, or grammatical, meaning. First, one must recognize that the meaning of any sentence comprises two parts, the meanings of the words it contains and the structural or grammatical meaning carried by the sentence itself. In English "the dog chased the cat" and "the boy chased the cat" differ in meaning because "dog" and "boy" are different words with different word meanings; the same applies to equivalent sentences in other languages. The two sentences "the dog chased the cat" and "the cat chased the dog," though containing exactly the same words, are different in meaning because the different word orders distinguish what are conventionally called subject and object. In Latin the two corresponding sentences would be distinguished not by word order, which is grammatically indifferent and largely a matter of style, but by different shapes in the lexical equivalents of "dog" and "cat." In Japanese the grammatical distinction of subject and object, normally marked by the word order subject-object-verb, can be reinforced by a subject particle after the first word and an object particle after the second.

The formal resources of any language for making distinctions in the structural meanings of sentences are limited by two things: the linear (time) dimension of speaking and the limited memory span of the human brain. Writing copies the time stream of speech with the linear flow of scripts. Diagrams and pictures employ two dimensions, and models employ three; but writing is partially relieved of memory-span restrictions by the permanence of visual marks. Because written texts are almost entirely divorced from oral pronunciation, sentence length and sentence complexity can be carried to extremes, as may be observed in some legal and legislative documents that are virtually unintelligible if read aloud.

Within these linear restrictions, distinctions corresponding to the main uses of language can be made. All languages can employ different sentence structures to state facts (declarative), to ask questions (interrogative), and to enjoin or forbid some course of action (imperative). More delicate means exist to soften or modify these basic distinctions: *e.g.,* "It's cold today, isn't it?"; "Isn't it still raining?"; "Shut the door, would you mind"; "Don't be long, will you?" Languages use their resources differently for these purposes, but, generally speaking, each seems to be equally flexible structurally. The principal resources are word order, word form, syntactic structure, and, in speech, pitch and stress placement. In English, as an example, a word or phrase can be highlighted by being placed first in the sentence when it would not normally occur there: compare "he can't bear loud noises" with "loud noises he can't bear" or "loud noises, he can't bear them." The object noun or noun phrase can also be put first by making the sentence passive; this allows the original subject to be omitted if one does not know or does not want to refer to an agent: "the town was destroyed (by the revolutionaries)." Within and together with all these possibilities, almost any word can be made contrastively prominent by being stressed (spoken more loudly) or by being uttered on a higher pitch, and very often these two are combined: "I asked you for *red* roses (not yellow)"; "I meant it for *you* (not her)"; "*I* know nothing about it (someone else may)." Prominence is especially associated with intonation, itself an important carrier of structural meaning in speech. One may state facts, ask questions, and give instructions with a variety of intonations indicating, along with visible gestures, different attitudes, feelings, and social and personal relations between speaker and hearer.

The possibilities of expressing structural meanings are a most important part of any language. They are acquired along with the rest of one's first language in childhood and are learned more slowly and with more difficulty in mastering a second or later language. Scholars are still only at the beginning of a full formal analysis of these resources, as far as most languages are concerned, and are still further from an adequate understanding of all the semantic functions performed by means of these resources.

Meanings of words

Lexical meaning. The other component of sentence meaning is word meaning, the individual meanings of the words in a sentence, as lexical items. The concept of word meaning is a familiar one. Dictionaries list words and in one way or another state their meanings. It is regarded as a sensible question to ask of any word in a language, "What does it mean?" This question, like many others about language, is easier to ask than to answer.

It is through lexical resources that languages maintain the flexibility their open-ended commitments demand. Every language has a vocabulary of many thousands of words, though not all are in active use, and some are known only to relatively few speakers. Perhaps the commonest delusion in considering vocabularies is the assumption that the words of different languages, or at least their nouns, verbs, and adjectives, label the same inventory of things, processes, and qualities in the world but unfortunately label them with different labels from language to language. If this were so, translation would be easier than it is; but the fact that translation, though often difficult, is possible indicates that people are talking about similar worlds of experience in their various languages.

Languages in part create the world in which men live. Of course, many words do name existing bits and pieces of earth and heaven: "stone," "tree," "dog," "woman," "star," "cloud," and so on. Others, however, do not so much pick out what is there as classify it and organize one's relations with it and with each other with regard to it. A range of living creatures are mammals or are vertebrates, because people classify them in these ways, among others, by applying selected criteria and so determining the denotation of the words mammal and vertebrate. Plants are vegetables or weeds according as groups of people classify them, and different plants are included and excluded by such classifications in different languages and different cultures.

Time and its associated vocabulary ("year," "month," "day," "hour," "minute," "yesterday," "tomorrow," and so on) do not refer to discrete sections of reality but enable people to impose some sort of order, in agreement with others, on the processes of change observed in the world. Personal pronouns pick out the persons speaking, spoken to, and spoken about; but some languages make different distinctions in their pronouns from those made in English. For example, in Malay, *kita,* which means "we," including the person addressed, is distinct from *kami,* a form for "we" that includes the speaker and a third person or persons but excludes the person addressed. In Japanese and in several other languages, a variety of words denoting the 1st and 2nd persons indicate additionally the observed or intended social relationship of those involved.

Other word meanings are even more language and culture bound, and in consequence harder to translate. "Right" and "wrong," "theft," "inheritance," "property," "debt," "sin," and "crime" (as different sorts of wrongdoing) are just a few of the words regulating one's conduct and relations with one's fellows in a particular culture. Translation becomes progressively harder as one moves to languages of more remote cultures, and it has been said that it requires "a unification of cultural context." Insofar as a person's understanding of the universe and of the relations between himself and other people is closely linked with the language he speaks, it must be assumed, and the evidence confirms this assumption, that the child progressively acquires such understanding along with his language.

Onomatopoeic words

The great majority of word shapes bear no direct relation to their lexical meanings. If they did, languages would be more alike. What are called onomatopoeic words are rather similar in shape through different languages: French *coucou,* English "cuckoo," and German *Kuckuck* directly mimic the call of the bird. English "dingdong" and German *bim-bam* share several sound features in common that partially resemble the clanging of bells. More abstractly, some direct "sound symbolism" has been seen between certain sound types and visual or tactile shapes. Most people agree that the made-up word "oomboolu" would better designate a round, bulbous object than a spiky one. In addition, the appropriateness of the vowel sound represented by *ee* in English "wee" and *i* in French *petit* "small" and Italian *piccolo* "small" for expressing things of small size has been traced in several languages.

All this, however, is a very small part of the vocabulary of any language. For by far the largest number of words in a language there is no direct association between sound and meaning. English "horse," German *Pferd,* French *cheval,* Latin *equus,* and Greek *hippos* are all unrelated to the animal so named, except that these words are so used in the languages concerned. This is what is meant by the term arbitrary in the second definition of language quoted at the beginning of this article. Vocabulary has to be largely arbitrary, because the greater part of the world and of man's experience is not directly associated with any kind of noise, and it is a contingent, though universal, fact of history and biology that sound and not the material of some other sense is the basis of human language.

The relations between sentence structure and structural meanings are also largely arbitrary and tacitly conventional. Though loudness and stress for emphasis and certain linguistic indications of anger, excitement, and the like are more closely akin to nonlinguistic ejaculations and are somewhat similar across language divisions, actual intonations and features such as word order, word inflection, and grammatical particles, used in maintaining distinctions in structural meaning, differ markedly in different languages.

SEMANTIC FLEXIBILITY

Not only are word meanings somewhat different in different languages; they are not fixed for all time in any one language. Semantic changes take place all along (see below), and at any moment the semantic area covered by a word is indeterminately bordered and differs from context to context. This is a further aspect and condition of the inherent and necessary flexibility of language.

General and specific designations. A person can be as precise or as imprecise as he needs or wishes to be. In general, words are fairly imprecise; yet for particular purposes their meanings can be tightened up, usually by bringing in more words or phrases to divide up a given field in more detail. "Good" contrasts generally with "bad"; but one can, for example, grade students as "first-class," "excellent," "very good," "good," "fair," "poor," and "failed" (or "bad"). In this case, "good" now covers a restricted and relatively low place in a field of associated terms. Colour words get their meanings from their mutual contrasts. The field of visually discriminable hues is very large and goes far beyond the resources of any vocabulary as it is normally used. Children learn the central or basic colour words of their language fairly early and at the same time; such terms as red and green are normally learned before subdivisions such as crimson and scarlet or chartreuse. It is well known that languages make their primary divisions of the spectrum of colours in different places; Japanese *aoi* covers many of the hues referred to in English by "green" and "blue," while "blue" covers much of the range of the two Russian words *goluboy* and *siny.* While the actual colour vocabularies of languages differ, however, recent research by Brent Berlin and Paul Kay has tried to show that "there exist universally for humans eleven basic perceptual color categories" that serve as reference points for the colour words of a language, whatever number may be regularly employed at any time.

Precision and imprecision of words

Ordinarily, considerable areas of indeterminate designation in colour vocabulary and in other fields are tolerated; between "red" and "purple" and between "purple" and "blue" there are hues that one would hesitate to assign firmly to one or the other and on which there would be considerable personal disagreement. When greater precision than normal is required—as, for example, in listing paint or textile colours—all kinds of additional terms can be brought into service to supplement the usual vocabulary: "off-white," "light cream," "lemon," "blush pink," and so on.

The vocabulary of kinship terms varies from language to language, reflecting cultural differences. English distinguishes the nearer kinsfolk by sex: "mother, father"; "sister, brother"; "aunt, uncle"; and others. Other languages, such as Malay, make a lexical distinction of age the primary one, with separate words for elder brother or sister and younger brother or sister. Still other languages—for example, some American Indian ones—use different words for the sister of a man and for the sister of a woman. But beyond this any language can be as precise as the situation demands in kin designation. When it is necessary, English speakers can specify "elder sister" and "female cousin," and within the overall category it is possible to distinguish "first and second cousins" and "cousins once removed," distinctions that it is ordinarily pedantic to make.

Vocabulary of mathematics

The best example of infinite precision available from a strictly limited lexical stock is in the field of arithmetic. Between any two whole numbers a further fractional or decimal number may always be inserted, and this may go on indefinitely: between 10 and 11, 10½ (10.5), 10¼ (10.25), 10⅛ (10.125), and so on. Thus, the mathematician or the physical scientist is able to achieve any desired degree of quantitative precision appropriate to his purposes; hence the importance of quantitative statements in the sciences—any thermometric scale contains far more distinctions of temperature than are reasonably available in the vocabulary of a language ("hot," "warm," "cool," "tepid," "cold," and so on). For this reason mathematics has been described as the ideal use of language, but for many purposes in everyday life the very imprecision of natural languages is the source of their strength and adaptability.

Neologisms. Every living language can readily be adapted to meet changes occurring in the life and culture of its speakers, and the main weight of such changes falls on vocabulary. Grammatical and phonological structures are relatively stable and change noticeably over centuries rather than decades (see below); but vocabularies can change very quickly both in word stock and in word meanings. Consider as an example the changes wrought by modern technology in the vocabularies of all European languages since 1945. Before that date "transistor" and "cosmonaut" did not exist, and "nuclear disarmament" would scarcely have had any clear meaning.

Every language can alter its vocabulary very easily, which means that every speaker can without effort adopt new words, accept or invent new meanings for existing words, and of course, cease to use some words or cease to use them in certain meanings. Dictionaries list some words and some meanings as "obsolete" or "obsolescent" to indicate this process. No two speakers share precisely the same vocabulary of words readily used and readily understood, though they may speak the same dialect. They will, however, naturally have the great majority of words in their vocabularies in common.

Languages have various resources for effecting changes in vocabulary. Meanings of existing words may change. With the virtual disappearance of falconry as a sport in England, "lure" has lost its original meaning of a bunch of feathers on a string by which hawks were recalled to their handler and is used now mainly in its metaphorical sense of enticement. The additional meaning of "nuclear" has already been mentioned; one may list it with words such as computer and jet, which acquired new ranges of meaning in the mid-20th century.

Creation of new words

All languages have the means of creating new words to bear new meanings. These can be new creations; "Kodak" is one such, invented at the end of the 19th century by George Eastman; "chortle," now in general use, was a jocular creation of the English writer and mathematician Lewis Carroll (creator of *Alice in Wonderland*); and "gas" was formed in the 17th century by the Belgian chemist and physician Jan Baptist van Helmont as a technical term in chemistry, loosely modelled on the Greek *chaos* ("formless void"). But mostly languages follow definite patterns in their innovations. Words can be made up without limit from existing words or from parts of words; the sources of "railroad," "railway," and "aircraft" are obvious, and so are the sources of "disestablishment," first cited in 1806 and thereafter used with particular reference to the status of the Church of England. The controversy over the relations between church and state in the 19th and early 20th centuries gave rise to a chain of new words as the debate proceeded: "disestablishmentarian," "antidisestablishmentarian," "antidisestablishmentarianism." Usually, the bits and pieces of words used in this way are those found in other such combinations, but this is not always so. The technical term permafrost (terrain that never thaws, as in the Arctic) contains a bit of "permanent" probably not hitherto found in any other word.

A particular source of technical neologisms in European languages has been the words and word elements of Latin and Greek. This is part of the cultural history of western Europe, in so many ways the continuation of Greco-Roman civilization. "Microbiology" and "dolichocephalic" are words well formed according to the rules of Greek as they would be taken over into English, but no records survive of *mikrobiologia* and *dolichokephalikos* ever having been used in Ancient Greek. The same is true of Latinate creations such as "reinvestment" and "longiverbosity." The long tradition of looking to Latin and, since the Renaissance, to Greek also as the languages of European civilization, keeps alive the continuing formation of learned and scientific vocabulary in English and other European languages from these sources. The dependence on the classical languages in Europe is matched by a similar use of Sanskrit words for certain parts of learned vocabulary in some modern Indian languages (Sanskrit being the classical language of India). Such phenomena are examples of loanwords, one of the readiest sources for vocabulary extension.

Transmission of loanwords

Loanwords are words taken into a language from another language (the term borrowing is used for the process). Most obviously, this occurs when new things come into speakers' experiences as the result of contacts with speakers of other languages. This is part of the history of every language, except for one spoken by an impossibly isolated community. "Tea" from Chinese, "coffee" from Arabic, and "tomato," "potato," and "tobacco" from American Indian languages are familiar examples of loanwords designating new products that have been added to the vocabulary of English. In more abstract areas, several modern languages of India and Pakistan contain many words that relate to government, industry, and current technology taken in from English. This is the result of British rule in these countries up to indpendence and the worldwide use of English as a language of international science since then.

In general, loanwords are rapidly and completely assimilated to the prevailing grammatical and phonological patterns of the borrowing language. The German word *Kindergarten,* literally "children's garden," was borrowed into English in the middle of the 19th century to designate an informal school for young children. It is now regularly pronounced as an English word, and the plural is kindergartens (not *Kindergärten,* as in German). Occasionally, however, some loanwords retain marks of their foreign origin: examples include Latin plurals such as cacti and narcissi (as contrasted with native patterns such as cactuses and narcissuses).

Languages differ in their acceptance of loanwords. An alternative way of extending vocabulary to cope with new products is to create a descriptive compound from within one's own language. English "aircraft" and "aeroplane" are, respectively, examples of a native compound and a Greek loan creation for the same thing. English "potato" is a loan; French *pomme de terre* (literally, "apple of the earth") is a descriptive compound. Chinese is particularly resistant to loans; "aircraft," "railway," and "telephone" are translated by newly formed compounds meaning literally "fly machine," "fire vehicle," and "lightning (electricity) language."

LANGUAGE AND CONCEPTUALIZATION

The ability to speak and the ability to conceptualize are very closely linked, and the child learns both these skills together at the same time. This is not to say that thinking is no more than subvocal speech, as some behaviourists have proposed; most people can think pictorially and in simple diagrams, some to a greater degree than others, and one has the experience of responding rationally to external stimuli without intervening verbalization. But, as 18th-century thinkers saw, man's rationality developed and still goes hand in hand with his use of language, and a good deal of the flexibility of languages has been exploited in man's progressive understanding and conceptualizing of the world he lives in and of his relations with other men. Different cultures and different periods have seen this process differently developed. The anthropological linguist Edward Sapir put it well: "The 'real world' is to a large extent unconsciously built up on the language habits of the group."

Language and thought patterns

Much of this lies in the irrecoverable prehistory of languages. The idea that there are still some primitive, almost "fossil" languages, embodying a very low level of conceptualization, is a vain one. All that can be said is that languages are different and that, in part, the world is seen differently through the eyes of speakers of different languages. But, in some cases, part of the lexical adaptation of a language to developing thought patterns can be followed through. Ancient Greece saw a wholly unique growth and flowering of civilization in the 1st millennium BC, which has put virtually the entire civilized world in its debt ever since. In Greek, along with the emergence of certain abstract concepts and ways of thinking, one can follow some of the changes of word meanings and the coining of new words that accompanied this. As an example, the word *dikē* originally meant "way" or "manner"; thereafter, it acquired the meaning of the right way of doing something, the right way of behaving, and finally abstract right. Its derivative *dikaiosynē,* traditionally translated "justice," became the subject of philosophical debate and analysis by the Greek philosophers and covered almost the whole range of moral obligation involved in the relations of one person with others in society. Similar debate and refinement of key terms in the various branches of thought covered by Greek philosophy can be followed through; indeed, the term philosophy is directly taken from Greek *philosophia,* a compound formed not later than the 5th century BC from *philo-* (compare *philein* "to love") and *sophia* "wisdom" to refer to abstract speculation and debate of a fundamental nature about the world and man's place in it.

More recently, the development of the lexical resources of the languages of civilization can be observed, in one way or another, as they keep up with the scientific progress that dominates contemporary life.

An examination of the lexical structure of languages throws some light on the relations among various aspects of man's conceptualization. Spatial relations and their expression seem to lie very deep in the content of vocabulary. Words referring to time are drawn metaphorically from spatial words with great frequency: "a long/short time," "the near future," "far ahead/separated in time." Although time is a continuum, people readily divide it up into bits and record it rather as they do materials extended in space: "five years," "three months," "six seconds." This last use of vocabulary may be a particular trait of European languages and some others. An American Indian language is reported not to do this nearly so readily; it uses cardinal numbers only for discrete, countable objects. A separate class of words aligns the vocabulary of sequential time with that of intensity, so that repetition of the same activity again and again (to a European) is rather the intensification of a single activity. Certain differences in cultural attitudes and world outlook are said to accompany this kind of linguistic difference.

Expression of abstract relationships

Spatial terms are also freely used in the expression of other, more abstract relationships: "higher temperature," "higher quality," "lower expectations," "summit of a career," "far removed from any sensible course of action," "a distant relationship," "close friends," "over and above what had been said." It has been theorized that the linguistic forms most closely associated semantically with the expression of relations—case inflections in languages exhibiting this category—are originally and basically spatial in meaning. This "localist" theory, as it has been called, has been debated since the beginning of the 19th century and probably cannot be accepted as it stands, but the fact that it can be proposed and argued shows the dominant position that spatial relations hold in the conceptualization and verbalization of relations in other realms of thought.

It has been maintained that the human brain has a preference for binary oppositions, or polarities. If this is so, it will help explain the numerous pairs of related antonyms that are found: "good, bad"; "hot, cold"; "high, low"; "right, wrong"; "dark, light"; and so on. For finer discriminations, these terms can be put into more narrowly specified fields containing more than two terms taken together, but their most general use is in binary contrasts. Here, however, one term seems to represent the fundamental semantic category in question. In asking about size, one asks "How big is it?"; about weight, "How heavy is it?"; and about evaluation, "How good is it?" It is possible to ask how small, how light, or how bad something is, but such questions presuppose that the thing in mind has already been graded on the small side, on the light side, or on the bad side.

STYLE

The capacity for conceptualization possessed and developed by languages is by no means the only purpose language serves. A person's speech, supplemented by facial expression and gesture when speaker and hearer are mutually in sight, indicates and is intended to indicate a great deal more than factual information, inquiries, and requests. The fact that some of these other functions are performed by parts of a language usually mastered later by foreign learners gives rise to misinterpretation and often makes foreign speakers appear rude or insensitive when they are, in actuality, simply deploying fewer resources in the language.

Expression of emotional attitudes

Within the range of the structural and lexical possibilities of a language, speakers are able to convey their emotional attitudes and feelings toward the person or persons they are addressing and toward the subject matter of what they are saying. They are also able to conceal such feelings as one form of linguistic deception, though this is usually a harder task. These same resources are also exploited to arouse appropriate feelings and responses in others, again independently of any factual content. This is the chosen field of the propagandist, the preacher, the orator, the barrister, and the advertiser. All languages make use of intonation and voice qualities in these different ways; a person can produce and recognize the intonation and type of voice employed in coaxing, in pleading, in browbeating, and in threatening, in pleasure, and in anger, as well as those appropriate for matter-of-fact statements and the exposition of details about which the speaker has little or no emotional involvement. To describe exactly which phonetic features are brought into play is quite another matter, involving advanced competence in phonetic discrimination and analysis. This is one of the areas of speech about which all too little is currently known. Grammar and vocabulary are equally involved, though differently in each language. English speakers know the difference between "Come and give me a hand!" and "Could you possibly come and help me?"; "He's got the gift of gab" and "He is undoubtedly a fluent and persuasive speaker" are each appropriate for different occasions. By greetings and leave-takings a great deal of intended interpretation of the social relations between individuals can be expressed. Much of this is the "good manners" taught to children and expected of adults; these aspects of language behaviour vary from culture to culture, but in none are they wholly absent. It is, of course, equally possible to be deliberately bad mannered or deliberately to flout a linguistic convention or expectation, but this can be done only by knowing what is expected in the situation. The refinements of rudeness, like the refinements of politeness, insofar as the use of language is involved, require a very good knowledge of a language if it is other than one's mother tongue.

Written language is no less adapted to conveying more than just factual information, asking factual questions, and giving instructions. Intonation and tone of voice are clearly not reproducible in existing orthographic systems, but part of the skill of a novelist or a reporter is to convey these features of speech in his descriptions. Additionally, grammatical and lexical choices are available to the writer, as reading the examples above will show, and everyone knows the special artistry and techniques involved in composing written memorandums or letters if they are to achieve precisely the purpose for which they are intended.

Styles: variations within a language

These variations, written and spoken, within a language or within any dialect of a language, may be referred to as styles. Each time a person speaks or writes he does so in one or another style, deliberately chosen with the sort of considerations in mind that have just been mentioned, even though in speech the choice may often be routine. Sometimes style, especially in literature, is contrasted with "plain, everyday language." In using such plain, unmarked types of speaking or writing, however, one is no less choosing a particular style, even though it is the most commonly used one and the most neutral in that it conveys and arouses the least emotional involvement or personal feelings.

Stylistic differences are available to all mature native speakers and in literate communities to all writers, as well as to foreigners who know a second language really well. But there is undoubtedly a considerable range of skills in exploiting all the resources of a language, and, whereas all normal adults are expected to speak correctly and, if literate, to write correctly, communities have always recognized and usually respected certain individuals as preeminently skilled in particular styles, as orators, storytellers, preachers, poets, scribes, belletrists, and so forth. This is the material of literature. Once it is realized that oral literature is just as much literature as the more familiar written literature, it can be understood that there is no language devoid of its own literature.

In all languages certain forms of utterance have been considered worthy of preservation, study, and cultivation. In writing, the nature of written surfaces makes this fairly easy, though not all written material is deliberately preserved; much of it is deliberately destroyed, and, although the chance survival of inscriptions on stone or clay is of the greatest value to the archaeologist and historian, a good deal of such material was never intended to survive. Literature, on the other hand, is essentially regarded as of permanent worth. Printing and, in earlier days, the copying of manuscripts are the means of preserving written literature. In illiterate communities certain persons memorize narratives, poems, songs, prayers, ritual texts, and the like, and these are passed on, with new creations in such styles, to succeeding generations. Such skills, preservative as well as creative, are likely to be lost along with much of the surrounding culture under the impact of literacy. Here, modern technology in the guise of the tape recorder has come to the rescue, and many workers in the field of unwritten languages are recording specimens of oral literatures with transcriptions and translations while speakers having the requisite knowledge and skills are still available. A great amount of such material, however, must have been irretrievably lost from illiterate cultures before the 20th century.

Culture and literature

All languages have a literature, but different types of literature flourish in different languages and in different cultures. A warrior caste or a general respect for martial prowess fosters heroic verse or prose tales; strongly developed magical and mystery cults favour ritualistic types of oral or written literature; urban yearnings for the supposed joys of country life encourage the development of pastoral poetry, itself an outgrowth of the songs of shepherds and rural workers; and the same sense of the jadedness of city life is the best ground for the cultivation of satirical verse and prose, a form of literature probably confined largely to urban civilizations. Every language has the resources to meet these and other cultural requirements in its literature as the occasions arise, but some literary forms are more deeply involved in the structure of the language itself; this is made clear by the relative difficulty of translating certain types of literature and literary styles from one language to another. Poetry, in particular, is closely bound to the structure of the language in which it is composed, and poetry is notoriously difficult to translate from one language into another.

The special vocabularies and linguistic forms used in several games have already been mentioned. Here one may point to the widespread existence of verbal games themselves, based on the accidental features of a particular language. English-speaking children are accustomed to riddles, puns, and spelling games: "I spy with my little eye something beginning with *p*" (notice the regular formula with which this opens). These and similar word games have been found all over the world. Homer records the punning use by Odysseus of No-man (Greek *Outis*) as his name when he was about to attack Cyclops, who then roared out "No-man is killing me!" and so failed to attract any help (*Odyssey* 9:366–408). In some languages that make use of lexically distinctive tones, tone puns (words alike but for having different tones) are a form of word play.

As an intellectual challenge, the crossword puzzle in all its varieties, originally an American development early in the 20th century, has maintained and indeed greatly increased its popularity over much of the literate world that employs the Latin (Roman) alphabet. Crossword-puzzle solvers rely heavily on the relative probabilities of letter sequences in written words to suggest an answer to a partly filled line; and, depending on the particular style of the originator, crossword clues make use of many sorts of formal features in the language, among them spelling puns, spoken puns, and accidental letter sequences in words and phrases. To be able to solve a crossword puzzle in a second language shows a high degree of skill and knowledge therein.

Language and culture

It has been seen that language is much more than the external expression and communication of internal thoughts formulated independently of their verbalization. In demonstrating the inadequacy and inappropriateness of such a view of language, attention has already been drawn to the ways in which one's mother tongue is intimately and in all sorts of details related to the rest of one's life in a community and to smaller groups within that community. This is true of all peoples and all languages; it is a universal fact about language.

Language as a part of culture

Anthropologists speak of the relations between language and culture. It is, indeed, more in accordance with reality to consider language as a part of culture. "Culture" is here being used, as it is throughout this article, in the anthropological sense, to refer to all aspects of human life insofar as they are determined or conditioned by membership in a society. The fact that a man eats or drinks is not in itself cultural; it is a biological necessity that he does so for the preservation of life. That he eats particular foods and refrains from eating other substances, though they may be perfectly edible and nourishing, and that he eats and drinks at particular times of day and in certain places are matters of culture, something "acquired by man as a member of society," according to the now-classic definition of culture by the English anthropologist Sir Edward Burnett Tylor. As thus defined and envisaged, culture covers a very wide area of human life and behaviour; and language is manifestly a part, probably the most important part, of it.

Although the faculty of language acquisition and language use is innate and inherited, and there is legitimate debate over the extent of this innateness, every individual's language is "acquired by man as a member of society," along with and at the same time as other aspects of that society's culture in which he is brought up. Society and language are mutually indispensable. Language can have developed only in a social setting, however this may have been structured, and human society in any form even remotely resembling what is known today or is recorded in history could be maintained only among people speaking and understanding a language in common use.

TRANSMISSION OF LANGUAGE AND CULTURE

Language is transmitted culturally; that is, it is learned. To a lesser extent it is taught, when parents deliberately encourage their children to talk and to respond to talk, correct their mistakes, and enlarge their vocabulary. But it must be emphasized that children very largely acquire their mother tongue (*i.e.,* their first language) by "grammar construction" from exposure to a random collection of utterances that they encounter. What is classed as language teaching in school either relates to second-language acquisition or, insofar as it concerns the pupils' first language, is in the main directed at reading and writing, the study of literature, formal grammar, and alleged standards of correctness, which may not be those of all the pupils' regional or social dialects. All of what goes under the title of language teaching at school presupposes and relies on the prior knowledge of a first language in its basic vocabulary and essential structure, acquired before school age.

Transmission of culture through language

If language is transmitted as part of culture, it is no less true that culture as a whole is transmitted very largely through language, insofar as it is explicitly taught. The fact that mankind has a history in the sense that animals do not is entirely the result of language. So far as researchers can tell, animals learn through spontaneous imitation or through imitation taught by other animals. This does not exclude the performance of quite complex and substantial pieces of cooperative physical work, such as a beaver's dam or an ants' nest, nor does it preclude the intricate social organization of some species, such as bees. But it does mean that changes in organization and work will be the gradual result of mutation cumulatively reinforced by survival value; those groups whose behaviour altered in any way that increased their security from predators or from famine would survive in greater numbers than others. This would be an extremely slow process, comparable to the evolution of the different species themselves.

There is no reason to believe that animal behaviour has materially altered during the period available for the study of human history, say the last 5,000 years or so, except, of course, when man's intervention by domestication or other forms of interference has itself brought about such alterations. Nor do members of the same species differ markedly in behaviour over widely scattered areas, again apart from differences resulting from human interference. Bird songs are reported to differ somewhat from place to place within species, but there is little other evidence for areal divergence. By contrast with this unity of animal behaviour, human cultures are as divergent as are human languages over the world, and they can and do change all the time, sometimes with great rapidity, as among the industrialized nations of the 20th century.

The processes of linguistic change and its consequences will be treated below. Here, cultural change in general and its relation to language will be considered. By far the greatest part of learned behaviour, which is what culture involves, is transmitted by vocal instruction, not by imitation. Some imitation is clearly involved, especially in infancy, in the learning process, but proportionately this is hardly significant.

Through the use of language, any skills, techniques, products, modes of social control, and so on can be explained, and the end results of anyone's inventiveness can be made available to anyone else with the intellectual ability to grasp what is being said. Spoken language alone would thus vastly extend the amount of usable information in any human community and speed up the acquisition of new skills and the adaptation of techniques to changed circumstances or new environments. With the invention and diffusion of writing, this process widened immediately, and the relative permanence of writing made the diffusion of information still easier. Printing and the increase in literacy only further intensified this process. Modern techniques for almost instantaneous transmission of the written and spoken word all over the globe, together with the rapid translation services now available between the major languages of the world, have made it possible for usable knowledge of all sorts to be made accessible to people almost anywhere in the world in a very short time. This accounts for the great rapidity of scientific, technological, political, and social change in the contemporary world. All of this, whether ultimately for the good or ill of mankind, must be attributed to the dominant role of language in the transmission of culture.

Diffusion of knowledge through writing and printing

LANGUAGE AND SOCIAL DIFFERENTIATION AND ASSIMILATION

The part played by varations within a language in differentiating social and occupational groups in a society has already been referred to above. In language transmission this tends to be self-perpetuating unless deliberately interfered with. Children are in general brought up within the social group to which their parents and immediate family circle belong, and they learn the dialect and speaking styles of that group along with the rest of the subculture and behavioral traits and attitudes that are characteristic of it. This is a largely unconscious and involuntary process of acculturation, but the importance of the linguistic manifestations of social status and of social hierarchies is not lost on aspirants for personal advancement in stratified societies. The deliberate cultivation of an appropriate dialect, in its lexical, grammatical, and phonetic features, has been the self-imposed task of many persons wishing "to better themselves" and the butt of unkind ridicule on the part of persons already feeling themselves secure in their social status or unwilling to attempt any change in it. Much of the comedy in George Bernard Shaw's *Pygmalion* turns on Eliza's need to unlearn her native Cockney if she is to rise in the social scale. Conversely, it is readily apparent today that middle class people, mostly adolescents, who for some reason want to "opt out" of the social group of their parents make every effort to abandon the distinctive aspects of the social dialect that would mark them, along with dress and general behaviour, as members of a group whose mores they are, at least temporarily, affecting to reject. Culturally and subculturally determined taboos play a part in all this, and persons desirous of moving up or down in the social scale have to learn what words to use and what words to avoid if they are to be accepted and to "belong" in their new position. All through the ages, a good part of the material for "comedies of manners" has come from the social role of language variation within a society.

The same considerations apply to changing one's language as to changing one's dialect. Language changing is harder for the individual and is generally a rarer occurrence, but it is likely to be widespread in any mass immigration movement. In the 19th and early 20th centuries, the eagerness with which immigrants and the children of immigrants from continental Europe living in the United States learned and insisted on speaking English is an illustration of their realization that English was the linguistic badge of full membership in their new homeland at the time when the country was proud to consider itself as the melting pot in which people of diverse linguistic and cultural origins would become citizens of a unified community.

Special vocabularies of games, trades, professions

The same sort of self-perpetuation, in the absence of deliberate rejection, operates in the special languages of games and of trades and professions (these are in the main concerned with special vocabularies). Game learners, apprentices, and professional students learn the locutions together with the rest of the game or the job. The specific words and phrases occur in the teaching process and are observed in use, and the novice is only too eager to display an easy competence with such phraseology as a mark of his full membership of the group; *e.g.,* golfers are keen to talk of birdies, fairways, and slicing.

Languages and variations within languages play both a unifying and a diversifying role in human society as a whole. Language is a part of culture, but culture is a complex totality containing many different features, and the boundaries between cultural features are not clear-cut, nor do they all coincide. Physical barriers such as oceans, high mountains, and wide rivers constitute impediments to human intercourse and to culture contacts, though modern technology in the fields of travel and communications make such geographical factors of less and less account. More potent today are political restrictions on the movement of people and of ideas, such as divide western Europe from Communist eastern Europe; the frontiers between these two political blocs represent much more of a cultural dividing line than any other European frontiers.

The distribution of the various components of cultures differs, and the distribution of languages may differ from that of nonlinguistic cultural features. This results from the varying ease and rapidity with which changes may be acquired or enforced and from the historical circumstances responsible for these changes. In contemporary Europe, as the result of World War II, a major political and cultural division cuts across an area of relative linguistic unity in East and West Germany. It is significant, however, that differences of vocabulary and usage are already noticeable in the German speech from each side, overlying earlier differences attributed to regional dialects; one may surmise that, if the present political situation endures for several more generations, the East–West frontier will come to mark a definite dialect boundary within the German language as well.

THE CONTROL OF LANGUAGE FOR CULTURAL ENDS

Second-language learning. Language, no less than other aspects of human behaviour, is subject to purposive interference. When people with different languages need to communicate, various expedients are open to them, the most obvious being second-language learning and teaching. This takes time, effort, and organization, and, when more than two languages are involved, the time and effort are that much greater. Most people are monolingual, and those with a working knowledge of three or four languages are much fewer than those with a competence in just one second language. Other expedients may also be applied. Ad hoc pidgins for the restricted purposes of trade and administration were mentioned above. Tacit or deliberate agreements have been reached whereby one language is chosen for international purposes when speakers of several different languages are involved. In the Roman Empire, broadly, the western half used Latin as a lingua franca, and the eastern half used Greek. In western Europe during the Middle Ages, Latin continued as the international language of educated people, and Latin was the second language taught in schools. Later, the cultural, diplomatic, and military reputation of France made French the language of European diplomacy. This use of French as the language of international relations persisted until the present century. At important conferences among representatives of different nations, it is usually agreed which languages shall be officially recognized for registering the decisions reached; and the provisions of treaties are interpreted in the light of texts in a limited number of languages, those of the major participants.

International and diplomatic languages

Since World War II the dominance of the English-speaking peoples in science and technology and in international commerce has led to the recognition of English as the major international language in the world of practical affairs, with more and more countries making English the first foreign language to be taught and thus producing a vast expansion of English-language-teaching programs all over the world. Those whose native language is English do not sufficiently realize the amount of effort, by teacher and learner alike, that is put into the acquisition of a working knowledge of English by educated first speakers of other languages.

As an alternative to the recognition of particular natural languages as international in status, attempts have been made to invent and propagate new and genuinely international languages, devised for the purpose. Of these, Esperanto, invented by the Polish-Russian doctor L.L. Zamenhof in the 19th century, is the best known. Such languages are generally built up from parts of the vocabulary and grammatical apparatus of the better known existing languages of the world. The relationship between the written letter and its pronunciation is more systematic than with many existing orthographies (English spelling is notoriously unreliable as an indication of pronunciation), and care is taken to avoid the grammatical irregularities to which all natural languages are subject and also to avoid sounds found difficult by many speakers (*e.g.,* the English *th* sounds, which most Europeans, apart from English speakers, dislike). These artificial languages have not made much progress, though an international society of Esperanto speakers does exist.

Language as a cohesive political force

Nationalistic influences on language. Deliberate interference with the natural course of linguistic changes and the distribution of languages is not confined to the facilitating of international intercourse and cooperation. Language as a cohesive force for nation-states and for linguistic groups within nation-states has for long been manipulated for political ends. Multilingual states can exist and prosper; Switzerland is a good example. But linguistic rivalry and strife can be disruptive. Language riots have occurred in Belgium between French and Flemish speakers and in parts of India between rival vernacular communities. A language can become or be made a focus of loyalty for a minority community that thinks itself suppressed, persecuted, or subjected to discrimination. The French language in Canada in the mid-20th century is an example. In the 19th and early 20th centuries Irish Gaelic came to symbolize Irish patriotism and Irish independence from Great Britain. Since independence, government policy continues to insist on the equal status of English and Irish in public notices and official documents, but, despite such encouragement and the official teaching of Irish in the state schools, a main motivation for its use and study has disappeared, and the language is giving ground to English under the international pressures referred to above.

For the same reasons, a language may be a target for attack or suppression, if the authorities associate it with what they consider a disaffected or rebellious group or even just a culturally inferior one. There have been periods when American Indian children were forbidden to speak a language other than English at school and when pupils were not allowed to speak Welsh in British state schools in Wales. Both these prohibitions have been abandoned. Since the Spanish Civil War of the 1930s Basque speakers have been discouraged from using their language in public, as a consequence of the strong support given by the Basques to the republican forces. Interestingly, on the other side of the Franco-Spanish frontier, French Basques are positively encouraged to keep their language in use, if only as an object of touristic interest and consequent economic benefit to the area.

Translation. So far, some of the relatively large-scale effects of culture contacts on languages and on dialects within languages have been surveyed. A continuous concomitant of contact between two mutually incomprehensible tongues and one that does not lead either to suppression or extension of either is translation. As soon as two speakers of different languages need to converse, translation is necessary, either through a third party or directly.

Oral and written translation

Before the invention and diffusion of writing, translation was instantaneous and oral; persons professionally specializing in such work were called interpreters. In predominantly or wholly literate communities, translation is thought of as the conversion of a written text in one language into a written text in another, though the modern emergence of the simultaneous translator or professional interpreter at international conferences keeps the oral side of translation very much alive.

The tasks of the translator are the same whether the material is oral or written, but, of course, translation between written texts allows more time for stylistic adjustment and technical expertise. The main problems have been recognized since antiquity and were expressed by St. Jerome, translator of the famed Latin Bible, the Vulgate, from the Hebrew and Greek originals. Semantically, these problems relate to the adjustment of the literal and the literary and the conflicts that so often occur between an exact translation of each word, as far as this is possible, and the production of a whole sentence or even a whole text that conveys as much of the meaning of the original as can be managed. These problems and conflicts arise because of factors already noticed in the use and functioning of language: languages do not operate in isolation but within and as part of cultures, and cultures differ from each other in various ways. Even between the languages of communities whose cultures are fairly closely allied, there is by no means a one-to-one relation of exact lexical equivalence between the items of their vocabularies.

In their lexical meanings, words acquire various overtones and associations that are not shared by the nearest corresponding words in other languages; this may vitiate a literal translation. The English author and theologian Ronald Knox has pointed to the historical connections of the Greek *skandalon* "stumbling block, trap, or snare," inadequately rendered by "offense," its usual New Testament translation. In modern times translators of the Bible into the languages of peoples culturally remote from Europe are well aware of the difficulties of finding a lexical equivalent for "lamb," when the intended readers, even if they have seen sheep and lambs, have no tradition of blood sacrifice for expiation nor long-hallowed associations of lambs with lovableness, innocence, and apparent helplessness. The English word uncle has, for various reasons, a cozy and slightly comic set of associations. The Latin poet Virgil uses the words *avunculus Hector* in a solemn heroic passage of the Aeneid (Book III, line 343); to translate this by "uncle Hector" gives an entirely unsuitable flavour to the text.

Poetry translation

The translation of poetry, especially into poetry, presents very special difficulties, and the better the original poem, the harder the translator's task. This is because poetry is, in the first instance, carefully contrived to express exactly what the poet wants to say. Second, to achieve this end, the poet calls forth all the resources of the language in which he is writing, matching the choice of words, the order of words, and grammatical constructions, as well as phonological features peculiar to the language in metre, perhaps supplemented by rhyme, assonance, and alliteration. The available resources differ from language to language; English and German rely on stress-marked metres, but Latin and Greek used quantitative metres, contrasting long and short syllables, while French places approximately equal stress and length on each syllable. The translator must try to match the stylistic exploitation of the particular resources in the original language with comparable resources from his own. Because lexical, grammatical, and metrical considerations are all interrelated and interwoven in poetry, a satisfactory literary translation is usually very far from a literal word for word rendering. The more the poet relies on language form, the more embedded his verses are in that particular language, and the harder they are to translate adequately. This is especially true with lyrical poetry in several languages, with its wordplay, complex rhymes, and frequent assonances.

At the other end of the translator's spectrum, technical prose dealing with internationally agreed scientific subjects is probably the easiest type of material to translate, because cultural unification (in this respect), lexical correspondences, and stylistic similarity already exist in this type of usage in the languages most commonly involved, to a higher degree than in other fields of discourse.

Significantly, it is this last aspect of translation to which mechanical and computerized techniques are being applied with some prospects of limited success. Machine translation, whereby, ultimately, a text in one language could be fed into a machine to produce an accurate translation in another language without further human intervention, has been largely concentrated on the language of science and technology, with its restricted vocabulary and overall likeness of style, for both linguistic and economic reasons. Attempts at machine translation of literature have been made, but success in this field, more especially in the translation of poetry, seems very remote at present.

Translation as an art

Translation on the whole is an art, not a science. Guidance can be given and general principles can be taught, but after that it must be left to the individual's own feeling for the two languages concerned. Almost inevitably, in a translation of a work of literature something of the author's original intent must be lost; in those cases in which the translation is said to be a better work than the original, an opinion sometimes expressed about the English writer Edward Fitzgerald's "translation" of *The Rubáiyát of Omar Khayyám,* one is dealing with a new, though derived, work, not just a translation. The Italian epigram remains justified: *Traduttore traditore* "The translator is a traitor."

Messages and codes. Translation serves to extend the communicative value of a text. Sometimes people want to restrict it. Confidential messages, spoken and written, require for their efficacy that they be known to and understood by only the single person or the few persons to whom they are addressed. Such are diplomatic exchanges, operational messages in wartime, and some transmissions of commercial information. Protection of written messages from interception has been practiced for many centuries. Recent developments in telegraphy and telephony have made protection against unauthorized reception more urgent, whether of texts transmitted as speech or as series of letters of the alphabet. Scrambling of telephony is a common expedient; the wave frequencies through which the sounds are to be transmitted are altered at the source so as to be unrecognizable and then reconverted by the intended recipient's receiver. Codes and ciphers (cryptography) are of much longer standing in the concealment of written messages, though their techniques are being constantly developed. Such gains are, of course, countered by developments in the techniques of decipherment and decoding (as distinct from getting hold of the key to the system in use). An important by-product of such techniques has been the reading and interpretation of inscriptions written in otherwise unknown languages or unknown writing systems for which no translation exists. The recent, very significant decipherment of the Linear B script and its recognition as Mycenaean Greek, an early Greek dialect written in a form of orthography quite distinct from the later classical Greek alphabet, was first achieved by the application of cryptographic "code cracking" methods (see also CRYPTOLOGY).

LANGUAGE LEARNING

Role of the brain in language learning

Every physiologically and mentally normal person has learned the main structure and basic vocabulary of his mother tongue by the end of childhood. It has been pointed out that the process of first-language acquisition as a spoken medium of communication is largely achieved from random exposure. There is legitimate controversy, however, over the nature and extent of the positive contribution that the human brain brings to the activity of grammar construction, the activity by which the child develops an indefinitely creative competence from the finite data that make up his actual experience of the language. Creativity is what must be stressed as the product of first-language acquisition. By far the greater number of all the sentences anyone hears and utters during his lifetime are new; that is, they have not occurred before in his personal experience. But individuals find no difficulty at all in understanding at once almost everything they hear nor for the most part in producing sentences to suit the requirements of every situation. This very ease of creativity in man's linguistic competence makes it hard to realize its extent. The only regularly reproduced sentences in most speakers' experience are the stereotyped forms of greeting and leave-taking and certain formalized responses to recurrent situations, such as shopping, cooperative activities in repetitive jobs, the stylized parts of church services, and the like.

Yet, despite this really immense achievement that the progressive mastery of one's first language constitutes, it arouses no comment and attracts no credit. It is simply part of what is expected of one in growing up. Different people may be singled out for praise in certain uses of their language, as good public speakers, authors, poets, tellers of tales, and solvers of puzzles, but not just as speakers. The credit that some individuals acquire in certain communities for "speaking correctly" is a different matter, usually the result of speaking as one's mother tongue a prestigious standard dialect among people most of whom speak another, less favoured one.

Bilingualism. The learning of a second and of any subsequently acquired language is quite a different matter. Except for one form of bilingualism (see below), it is a deliberate activity undertaken when one has already nearly or fully acquired the basic structure and vocabulary of one's first language. Of course, many people never do master significantly more than their own first language. It is only in encountering a second language that one realizes how complex language is and how much effort must be devoted to subsequent acquisition. It has been said that the principal obstacle to learning a language is knowing one already, and it may also be that the faculty of grammar construction exhibited in childhood is one that is gradually lost as childhood recedes.

Ability to learn additional languages

Whereas every normal person masters his mother tongue with unconscious ease, people vary in their ability to learn additional languages, just as they vary in other intellectual activities. Situational motivation, however, appears to be by far the strongest influence on the speed and apparent ease of this learning. The greatest difficulty is experienced by those who learn because they are told to or are expected to, without supporting reasons that they can justify. Given a motive other than external compulsion or expectation, the task is achieved much more easily (this, of course, is an observation in no way confined to language learning). In Welsh schools it is found that English children make slower progress in Welsh when their only apparent reason for learning Welsh is that there are Welsh classes. Welsh children, on the other hand, make rapid progress in English, the language of most further education, the newspapers, most television and radio, most of the better paid jobs, and of any job outside Welsh-speaking areas. Similar differences in motivation have accounted for the excellent standard of English, French, and German acquired by educated persons in the Scandinavian countries and in Holland, small countries whose languages, being spoken by relatively few foreigners, are of little use in international communication. This attainment may be compared with the much poorer showing in second-language acquisition among comparably educated persons in England and America, who have for long been able to rely on foreigners accommodating to their ignorance by speaking and understanding English.

It is often held that children brought up bilingually in places in which two languages are regularly in use are slower in schoolwork than comparable monolingual children, as a greater amount of mental effort has to be expended in the mastery of two languages. This is by no means proved; and, because much of a child's language acquisition takes place in infancy and in the preschool years, it does not represent an effort in the way that consciously learning a language in school does, and indeed it probably occupies a separate part of the child's mental equipment. The question of speed of general learning by bilinguals and monolinguals must be left open. It is quite a separate matter from the job of learning, by teaching at home or in school, to read and write in two languages; this undoubtedly is more of a labour than the acquisition of monolingual literacy.

Types of bilingualism

Two types of bilingualism have been distinguished, according to whether the two languages were acquired from the simultaneous experience of the use of both in the same circumstances and settings or from exposure to each language used in different settings (an example of the latter is the experience of English children living in India during the period of British ascendancy there, learning English from their parents and an Indian language from their nurses and family servants). However acquired, bilingualism leads to mutual interference between the two languages; extensive bilingualism within a community is sometimes held partly responsible for linguistic change (see below). Interference may take place in pronunciation, in grammar, and in the meanings of words. Bilinguals often speak their two languages each with "an accent"; *i.e.,* they carry into each certain pronunciation features from the other. The German word order in "He comes tomorrow home" has been reported as an example of grammatical interference; and in Candian French the verb *introduire* has acquired from English the additional meaning "introduce, make acquainted" (which in metropolitan French is *présenter*).

Literacy. The acquisition of literacy is something very different from the acquisition of one's spoken mother tongue, even when the same language is involved, as it usually is. Both skills, speaking and writing, are learned skills, but there the resemblance ends. The child learns his first language at the start involuntarily and mostly unconsciously from random exposure, even if no attempts at teaching are made. Literacy is deliberately taught and consciously and deliberately learned. There is current debate on the best methods and techniques for teaching literacy in various social and linguistic settings. Literacy is learned through speech, by a person already possessed of the basic structure and vocabulary of his language.

Such facts should be very obvious, but the now-accepted, though fairly recent, standard of near-universal literacy in technologically advanced countries, along with the fact that in second-language learning one usually acquires speech and writing skills at the same time, tends to bring these two parts of language learning under one head. Literacy is manifestly a desirable attainment for all communities, though not necessarily in all languages. It must be borne in mind that there are many distinct languages spoken in the world today by fewer than 1,000 or 500 or even 50 persons. The capital investment in literacy, including teaching resources, teacher time and training, printing, publications, and so forth, is vast, and it can be economically and socially justified only when applied to languages spoken and likely to continue to be spoken by substantial numbers over a wide area.

Effects of literacy

Literacy is in no way necessary for the maintenance of linguistic structure or vocabulary, though it does enable people to add words from the common written stock in dictionaries to their personal vocabulary very easily. It is worth emphasizing that until relatively recently in human history all languages were spoken by illiterate speakers and that there is no essential difference as regards pronunciation, structure, and complexity of vocabulary between spoken languages that have writing systems used by all or nearly all their speakers and the languages of illiterate communities.

Literacy has many effects on the uses to which language may be put; storage, retrieval, and dissemination of information are greatly facilitated, and some uses of language, such as philosophical system building and the keeping of detailed historical records, would scarcely be possible in a totally illiterate community. In these respects the lexical content of a language is affected, for example, by the creation of sets of technical terms for philosophical writing and debate. Because the permanence of writing overcomes the limitations of auditory memory span imposed on speech, sentences of greater length can easily occur in writing, especially in types of written language that are not normally read aloud and that do not directly represent what would be spoken. An examination of some kinds of oral literature, however, reveals the ability of the human brain to receive and interpret spoken sentences of considerable grammatical complexity.

In relation to pronunciation, writing does not prevent the historical changes that occur in all languages. Part of the apparent irrationality of English spelling, such as is found also in some other orthographies, lies just in the fact that letter sequences have remained constant while the sounds represented by them have changed. For example, the *gh* of "light" once stood for a consonant sound, as it still does in the word as pronounced in some Scots dialects; and the *k* of "knave" and "knight" likewise stood for an initial *k* sound (compare the related German words *Knabe* and *Knecht*). A few relatively uncommon words, including some proper names, are reformed phonetically, specifically to bring their pronunciation more in line with their spelling. Spelling prounciations, as these are called, are a product of general literacy. In London, the pronunciation of "St. Mary Axe" as if it were spelled "Simmery Axe" is now decidedly old-fashioned. "St. John" and "St. Clair" survive as proper names with their old pronunciations, in the latter case helped by the presence of the alternative spelling "Sinclair."

WRITTEN LANGUAGE

Historically, culturally, and in the individual's life, writing is subsequent to speech and presupposes it. Aristotle expressed the relation thus: "Speech is the representation of the experiences of the mind, and writing is the representation of speech" (*On Interpretation*). But it is not as simple as this would suggest. Alphabetic writing, in which, broadly, consonant and vowel sounds are indicated by letters in sequence, is the most widespread system in use today, and it is the means by which literacy will be disseminated, but it is not the only system, nor is it the earliest.

Evolution of writing systems. Writing appears to have been evolved from an extension of picture signs: signs that directly and iconically represented some thing or action and then the word that bore that meaning. Other words or word elements not readily represented pictorially could be assigned picture signs already standing for a word of the same or nearly the same pronunciation, perhaps with some additional mark to keep the two signs apart. This sort of device is used in children's word puzzles, as when the picture of a berry is used to represent, say, the second half of the name Canterbury. This opens the way for what is called a character script, such as that of Chinese, in which each word is graphically represented by a separate individual symbol or character or by a sequence of two or more such characters. Writing systems of this sort have appeared independently in different parts of the world.

Character scripts

Chinese character writing has for many centuries been stylized, but it still bears marks of the pictorial origin of some characters. Chinese characters and the characters of similar writing systems are sometimes called ideograms, as if they directly represented thoughts or ideas. This is not so. Chinese characters stand for Chinese words or, particularly as in modern Chinese, bits of words; they are the symbolization of a particular language, not a potentially universal representation of thought. The ampersand (&) sign, standing for "and" in English printing, is a good isolated example of a character used in an alphabetic writing system.

Character writing is laborious to learn and imposes a burden on the memory. Alternatives to it, in addition to alphabetic writing, include scripts that employ separate symbols for the syllable sequences of consonants and vowels in a language, with graphic devices to indicate consonants not followed by a vowel. The Devanagari script, in which classical Sanskrit and modern Hindi are written, is of this type, and the Mycenaean writing system, a form of Greek writing in use in the 2nd millennium BC and quite independent of the later Greek alphabet, was syllabic in structure. Japanese employs a mixed system, broadly representing the roots of words by Chinese characters (the Japanese learned writing from the Chinese in and after the 5th century AD) and the inflectional endings by syllable signs. These syllable signs are an illustration of the way in which a syllabic script can develop from a character script: certain Chinese characters were selected for their sound values alone and, reduced in size and complexity, have been standardized as signs of a particular consonant and vowel sequence or of a single vowel sound.

The Greek alphabet came from the Phoenician script, a

Greek alphabet

syllabic-type writing system that indicated the consonant sounds. By a stroke of genius, a Greek community decided to employ certain consonantal signs to which no consonant sound corresponded in Greek as independent vowel signs, thus producing an alphabet, a set of letters standing for consonants and vowels. The Greek alphabet spread over the ancient Greek world, undergoing minor changes. From a Western version sprang the Latin (Roman) alphabet. Also derived from the Greek alphabet, the Cyrillic alphabet was devised in the 9th century AD by a Greek missionary, St. Cyril, for writing the Slavic languages.

Spelling. Alphabetic writing is not and cannot be an exact representation of the sequence of sounds or even of the sequence of distinctive sounds in the spoken forms of words and sentences. "Consonant" and "vowel" mean different things when applied to letters and to sounds, though there is, of course, much overlap. The *y* at the beginning of "yet" stands for a consonant sound; at the end of "jetty" it stands for a vowel sound. In "thick" and "thin" the sequence *th* represents a single sound, not a *t* sound followed by an *h* sound. In "kite" the *e* represents no sound directly but distinguishes the vowel between *k* and *t* from the vowel in "kit." These disharmonies arise from a number of causes. Economy in the use of letters is one factor. In addition, spoken forms are always changing over the centuries, whereas writing, particularly since the invention of printing, is very conservative. At one time the *e* at the end of words such as "kite" did stand for a vowel sound. This sound was lost between the 14th and 16th centuries, a time when other changes in the pronunciation of such words also occurred. The notorious *ough* spellings in English, standing for different sounds and sound sequences in "rough," "cough," "dough," "plough," "ought," and other such words, have arisen from historical changes that have driven spelling and pronunciation further apart.

This, of course, does not mean that spelling reforms are out of the question. Spelling reform has been talked of in relation to English for many centuries without much effect; but in some countries—for example, Norway and Holland—official action has prescribed certain reforms to be made, and these have then been taught in school and have gradually found their way into printed works. The sheer volume of printed matter preserved for use and consultation in the modern world adds much weight against the convenience otherwise accruing from reforms designed to correct the historically produced disharmonies between spelling and pronunciation.

Representing grammatical distinctions in spelling

Moreover, it is not always most useful for spellings to represent exactly the sound sequences in a word and nothing else; this is the task for which phoneticians have devised transcriptions. As far as the sounds themselves are concerned, the plural signs of "cats," "dogs," and "horses" are different: the final sound of "cats" is like the initial sound of "sink," that of "dogs" like the initial sound of "zinc," and the plural of "horse" is indicated by a sound sequence rather like that in "is." But they are all indicated in writing by one and the same letter and always have been, because only one grammatical distinction, that of singular as against plural, is involved, and at this point in the language the actual differences in the sounds, important elsewhere, are irrelevant.

Letters, insofar as they stand for sounds, stand for consonants and vowels. But other sound features are involved in languages. In English words the location of the stress is important, and the words "import" as a noun and "import" as a verb are distinguished by this alone. All languages make use of sequences of rises and falls in pitch, called intonation, as part of spoken communication. These phenomena are unrepresented in orthography except for certain punctuation marks such as ? and ! and sometimes by italicization and underlining.

This is not a weakness in orthography. Writing is normally intended to be read and when necessary read aloud by people who already know the language and are therefore able to supply from their own competence the required detail. For specific purposes such as foreign-language teaching, as well as for the specific study of pronunciation and speech sounds in phonetics and phonology, various forms of transcription have been devised to indicate unambiguously by written signs the precise form of the spoken utterance, without regard to other considerations.

Written versus spoken languages. For these reasons one should distinguish the grammar of a written language (*e.g.*, written English) from the grammar of the corresponding spoken language (spoken English). The two grammars will be very similar, and they will overlap in most places; but the description of spoken English will have to take into account the grammatical uses of features such as intonation, largely unrepresented in writing, and the description of written English must deal adequately with the greater average length of sentences and some different syntactic constructions and word forms characterizing certain written styles but almost unknown in ordinary speech (*e.g.*, "whom" as the objective form of "who").

Ancient, or dead, languages

In studying ancient (dead) languages one is, of course, limited to studying the grammar of their written forms and styles, as their written records alone survive. Such is the case with Latin, Ancient Greek, and Sanskrit (Latin lives as a spoken language in very restricted situations, such as the official language of some closed religious communities, but this is not the same sort of Latin as that studied in classical Latin literature; Sanskrit survives also as a spoken language in similarly restricted situations in a few places in India). Scholars may be able to reconstruct something of the pronunciation of a dead language from historical inferences and from descriptions of its pronunciation by authors writing when the language was still spoken. They know a good deal about the pronunciation of Greek and Latin and a great deal about the pronunciation of Sanskrit, because ancient Indian scholars left a collection of extremely detailed and systematic literature on its pronunciation. But this does not alter the fact that when one teaches and learns dead languages today, largely for their literary value and because of the place of the communities formerly speaking them in our own cultural history, one is teaching and learning the grammar of their written forms. Indeed, despite what is known about the actual pronunciation of Greek and Latin, Europeans on the whole pronounce what they read in terms of the pronunciation patterns of their own languages.

Under present conditions, with universal literacy either an accepted fact or an accepted target, it is assumed that, wherever it is convenient or useful, writing may be employed for any purpose for which speech might have been used and by all sections of the community. This has not always been so. Literacy was until the 19th century the privilege of the few. In other periods and cultures, writing was the preserve of certain defined groups, such as the priesthood and the official class, and it was restricted to certain purposes, such as the annals of important events, genealogical tables, and records of inventories of things and persons. It is highly probable that writing first developed for particular types of use by particular groups of specialists within communities and subsequently, because of its obvious utility, spread outside these limits.

For further accounts of writing systems in greater detail, see WRITING.

Linguistic change

Every language has a history; and, as in the rest of human culture, changes are constantly taking place in the course of the learned transmission of a language from one generation to another. This is just part of the differences between human culture and animal behaviour. Languages change in all their aspects, in their pronunciation, word forms, syntax, and word meanings (semantic change). These changes are mostly very gradual in their operation, becoming noticeable only cumulatively over the course of several generations. But, in some areas of vocabulary, particular words closely related to rapid cultural change are subject to equally rapid and therefore noticeable changes within a generation or even within a decade. In the 20th century the vocabulary of science and technology is an outstanding example. The same is also true of those parts of vocabulary that are involved in fashionable slangs and jargons, whose raison d'être in promoting group, particu-

Changes in vocabulary

larly age-group, solidarity depends on their being always fresh and distinctive. Old slangs date, as any reading of a novel or visit to a film more than 10 years old is apt to show. The rapid obsolescence of young people's slangs is equally to be seen in the unsuccessful efforts of some well-intentioned older persons who vainly attempt to cultivate the speech styles of present-day youth groups in a misdirected attempt to bridge "the generation gap" (this last phrase is an example of mid-20th-century pseudoscientific slang).

DIVERSIFICATION OF LANGUAGES

Changes through time. In the structural aspects of language, their pronunciation and grammar, and in vocabulary less closely involved in rapid cultural movement, the processes of linguistic change are best observed by comparing written records of a language over extended periods. This is most readily seen by English speakers through setting side by side present-day English texts with 18th-century English, the English of the Authorized Version of the Bible, Shakespearean English, Chaucer's English, and the varieties of Old English (Anglo-Saxon) that survive in written form. Noticeably, as one goes back in time, the effort required in understanding increases, and, while people do not hesitate to speak of "Shakespearean English," they are more doubtful about Chaucer, and for the most part Old English texts are as unintelligible to a modern English speaker as, for example, texts in German. It is clear that the differences involved include word meanings, grammar, and, so far as this can be reconstructed, pronunciation.

Similar evidence, together with what is known of the cultural history of the peoples concerned, makes clear the continuous historical connections linking French, Spanish, Portuguese, Italian, and Romanian with the spoken ("vulgar") Latin of the western Roman Empire. This group constitutes the Romance subfamily of languages and is an example of how, as the result of linguistic change over a wide area, a group of distinct, though historically related, languages comes into being.

In the transmission of a language from parent to child, slight deviations in all aspects of language use occur all the time, and as the child's speech contacts widen he confronts a growing range of slight differences in personal speech forms, some of them correlating with social or regional differences within a community, these speech differences themselves being the results of the transmission process. As a consequence, the child's speech comes to differ slightly from that of his parents' generation. In urbanized communities an additional factor is involved: children have been shown to be effectively influenced by the speech habits of their peer groups once they have made contacts with them in and out of school.

Cumulative changes in languages

Such changes, though slight at the time, are progressively cumulative. Since ready intercommunication is a primary purpose of language, as long as a community remains unitary, with strong central direction and a central cultural focus, such changes will not go beyond the limits of intercomprehensibility. But in more scattered communities and in larger language areas, especially when cultural and administıative ties are weakened and broken, these cumulative deviations in the course of generations give rise to wider regional differences. Such differences take the form of dialectal differentiation as long as there is some degree of mutual comprehension but eventually result in the emergence of distinct languages. This is what happened in the history of the colloquial Latin of the western Roman Empire, and it can be assumed that a similar course of events gave rise to the separate Germanic languages (English, German, Dutch, Danish, Norwegian, Swedish, and some others), though in this family the original unitary language is not known historically but inferred as "Common Germanic" or "Proto-Germanic" and tentatively assigned to early in the 1st millennium BC as the period before separation began.

This is how language families have developed. Most but not all of the languages of Europe belong to the Indo-European family, so-called because in addition it includes the classical Indian language Sanskrit and most of the modern languages of northern India and Pakistan. It includes as subfamilies the two families just mentioned, Romance and Germanic, and several others. It is assumed that the subfamilies, and from them the individual languages of the Indo-European family, are ultimately derived from a unitary language spoken somewhere in eastern Europe or western Asia (its exact location is still under debate), perhaps 5,000 years ago. This unitary language has itself been referred to as "Indo-European," "Proto-Indo-European," the "common parent language," or the "original language" (*Ursprache*) of the family. But it must be emphasized that, whatever it may have been like, it was just one language among many and of no special status in itself. It was certainly in no way the original language of mankind or anything like it. It had its own earlier history, of which virtually nothing can be inferred, and it was, of course, very recent in relation to the time span of human language itself. What is really special about such "parent" or "proto-" languages is that they represent the farthest point to which our available techniques and resources enable us to reconstruct the prehistory of our attested and living languages. Similarly constituted families of languages derived from inferred common sources have been established for other parts of the world; for example, Altaic, covering Turkish and several languages of Central Asia, and Bantu, containing many of the languages of central and southern Africa. For further details of these and other language families see LANGUAGES OF THE WORLD.

"Parent" or "proto-" languages

If enough material in the form of written records from past ages were available, it would be possible to group all the world's languages into historically related families. In addition, an answer could perhaps be posited to the question of whether all languages are descended from a single original language or whether languages emerged independently among several groups of early peoples (the rival theories of monogenesis and polygenesis, a controversy more confidently disputed in the 19th century than today). In actual fact, written records, when they are available, go back only a fraction of the time in which human speech has been developed and used, and over much of the globe written records are nonexistent. In addition, there are no other linguistic fossils comparable to the fossils of geological prehistory. This means that the history and prehistory of languages will not be able to go back more than a few thousand years BC and will be much more restricted in language areas in which few or no written records are available, as in much of Africa and in South America. Many languages will remain not related with certainty to any family. Nevertheless, the methods of historical linguistics, involving the precise and systematic comparison of word forms and word meanings (see further LINGUISTICS), have produced remarkable results in establishing language families on the same basis as Indo-European was established, in far less favourable fields. But any attempt by these means to get back to "the origin of language" or to reconstruct man's original language, if indeed there was one, is quite beyond the reach of science and will remain so.

Changes through geographical movement. The fundamental cause of linguistic change and hence of linguistic diversification is the minute deviations occurring in the transmission of speech from one generation to another. But other factors contribute to the historical development of languages and determine the spread of a language family over the world's surface. Population movements naturally play a large part, and movements of peoples in prehistoric times carried the Indo-European languages from a relatively restricted area into most of Europe and into northern India, Persia, and Armenia. But language and race are by no means the same thing, and the spread of the Indo-European languages resulted, in the main, from the imposition of one of them on the earlier population of the territories occupied. In the historical period, within Indo-European, the same process can be seen at work in the western Roman Empire. Latin superseded the earlier, largely Celtic languages of the Iberian Peninsula and of Gaul (France) not through population replacement (the number of Roman soldiers and settlers in the empire was never large) but through the abandonment of these languages by the inhabitants over the generations as they

Population movements

found in Latin the language of commerce, civilization, law, literature, and social prestige.

Conquest does not always lead to the supersession of a language. Greek survived centuries of Turkish rule and indeed remained a focus of national feeling, as has happened elsewhere in history. Much depends on the various circumstances and on the mutual attitudes of those involved; what must be kept quite clear is the difference between movements of peoples and the spread of languages. When linguistically homogeneous people enter and occupy a virtually empty area, as with most of Australia, the two movements coincide.

Languages do not just spread and compete with each other for territorial use. They are in constant contact, and every language bears evidence of this throughout its history. Modern Greek is full of words of Turkish origin, despite efforts made at various times since independence to purify the language by official action. The Norman Conquest and a period in which French was the language of the ruling class in England effected great changes on English and contributed a very substantial number of French words to English vocabulary; hence the quantity of near synonymous pairs available today: "begin, commence"; "end, finish"; "kingly, royal"; "fight, combat"; and so on.

Tendencies against change. These historical processes take place without any direct volition on the part of speakers as regards the language itself. Latin was learned as part of personal advancement, not for its own sake. Loans were incorporated almost without their being noticed, along with the concomitant cultural changes and innovations. Deliberate action directly related to a language does occur. The creation of pidgins involves some degree of linguistic consciousness on the part of their first users. More deliberate, however, have been various attempts at preserving the purity of a language, at least for some uses, or at arresting the processes of change. The care bestowed on the preservation of the Sanskrit used in religious ritual in ancient India and recent attempts to free Modern Greek from much of its Turkish vocabulary have already been noticed. For a period, under Nazi rule, efforts were made to replace some foreign words in the German language by words of native origin, and there have been movements to replace later accretions in English by words derived from Old English forms. In the long run, such attempts never succeed in preventing or reversing change; at best they preserve collaterally supposedly purer forms and styles for certain purposes and in certain contexts.

Attempts to prevent or reverse linguistic change

With the picture painted above of the tendency for languages to fragment first into dialects and then into separate languages, it might be thought that dialects are relatively late in appearance in the history of a language family. This impression is reinforced by the fact that most nonstandard dialects are unrepresented as such in writing, and so comparatively little is known about dialectal differences within most languages as one goes back in time. In this respect the very detailed knowledge of the Ancient Greek dialect situation is quite untypical.

In fact, dialect divisions must have been a feature of linguistic communities as early as there is any knowledge of them. Dialect splitting is fostered by isolation and loss of contact between groups within a speech community, and the sparse populations of earlier days, often nomadic and spread over large areas relative to their numbers, will have encouraged this process. It is simply the case that all but literate dialects have been lost in the past, and an artificial homogeneity is attributed to most ancient languages and to the so-called reconstructed parent languages of families.

Present-day conditions tend toward the amalgamation of dialects and the disappearance of those spoken by relatively few people. Urbanization, mass travel, universal education, broadcasting, ease of communication, and social mobility all foster rather large regional and social dialects, with special occupational types of language within them, in place of the small, strictly localized dialects of earlier times. This is one reason for the urgency with which dialect studies are being pursued in many Western industrialized countries, such as England and parts of the United States. If work is not done soon, many dialects may perish unrecorded (see also the section *Dialects* above).

For the same reasons, dialect divisions that earlier would have widened into distinct languages are now unlikely to do so. One may compare the emergence of the separate Romance languages from once unitary Latin with the splitting of South American Spanish and Portuguese into different dialects of these two languages. These dialectal divisions are not now expected to widen beyond the range of intercomprehensibility. These same conditions, together with the spread of literacy, are leading to the extinction of languages spoken by relatively small communities. Such is the fate of most of the North American Indian languages, and Irish, Welsh, and Scots Gaelic may ultimately survive only as learned second languages, preserved as cultural focuses for their communities. But in situations like this, both past and present, the intervening period of extensive bilingualism and the concomitant use of two languages has its effect on the changes taking place in the dominant language, which is influenced by the phonetic and grammatical composition of the speakers' former language.

Extinction of languages

LANGUAGE TYPOLOGY

Language families, as conceived in the historical study of languages, should not be confused with the quite separate classifications of languages by reference to their sharing certain predominant features of grammatical structure. Such classifications give rise to what are called typological classes.

In fulfilling the requirements of open-ended creativity imposed on language by human beings, grammatical structure has things in common in all known languages, particularly at the deeper levels of grammar. All known languages have words or wordlike elements combined in accordance with rules into sentences; all known languages distinguish in some way nounlike and verblike sentence components; and all known languages have the means of embedding or subordinating one sentence within another as an included clause (*e.g.*, "the sun set" and "we returned home": "When the sun set we returned home"; "Joan was playing tennis" and "Joan twisted her ankle": "Joan, who was playing tennis, twisted her ankle," or "while she was playing tennis, Joan twisted her ankle"). Descriptive analyses of all the languages of the world have not yet been prepared, and, of course, there is information about only a minute number of those that are no longer spoken—namely, those few that were written. But there is enough known to make the assertion of such universal features as have been given with fair confidence. These are often referred to as language universals; their nature and extent is the subject of current discussion and research.

Within these very general guidelines, however, languages exhibit various types of structure. This can most readily be seen by comparing the relations between the forms of words and their syntactic functions in different languages. Such a comparison is the basis of three broad types of language that have been distinguished since the beginning of the 19th century. They are, in fact, more like characteristics than types, in that most languages contain traces of all three, in different proportions.

Broad types of language structure

Classical Chinese made little or no use of word-form variation, such as is found, for example, in Latin, for grammatical purposes. Sentence structure was expressed by word order, word grouping, and the use of specific grammatical words, or particles. Such languages have been called isolating or analytic. Modern Chinese languages are much less analytic than is often believed; probably, Vietnamese is the most fully representative of this type today. Some languages string together, or agglutinate, successive bits, each with a specific grammatical function, into the body of single words. Turkish is a typical agglutinative language: compare Turkish *evleri* "houses" (accusative case), in which *ev* is the root meaning "house," *-ler* marks plurality, and *-i* is the sign for accusative, with Latin *domūs,* in which *-ūs* combines the representation of accusative and plural without the possibility of assigning either category separately to one part of the word ending. Latin is in this respect an inflectional, or fusional, language. In a more extreme example, Latin *ī* "go!" cumulatively represents in one fused form the verb meaning "go," active voice, imperative mood, second person, and singular number, each a grammatically distinct category.

English, like many other languages, is representative of all three types. In its use of word order alone to distinguish grammatical differences ("the dog chased the cat"; "the cat chased the dog") it resembles Classical Chinese rather than Latin. In a word form such as "manliness," in which each bit can be assigned a grammatical function ("man" the basic noun, *-li-* the adjective formative, and *-ness* the abstract noun formative), it makes use of agglutination, whereas plurals such as "men" and "geese" and past tenses such as "came" and "ran" fuse distinct grammatical categories into a word form in which only arbitrarily can one allot some sound segments, or letters, to one and some to the other.

Assigning languages to different types in this way involves a delicate procedure of balancing one part of the grammar against another and deciding which type of structure predominates and how well the other types are represented. Languages predominantly of each of the types are found in communities at with all levels of civilization and with all types of culture.

Change in grammatical structures

In the course of transmission, grammatical structures change, just as do pronunciation and meanings, and in time the cumulative effect may be the transference of a language from one overall type to another, although it remains descended from the earlier language and therefore is just as much part of the same historical family. Latin is very different typologically from French in its grammatical structure, but French is nevertheless the form that Latin took in France in the course of time. In the matter of the grammatical relevance of word order, the absence of case inflections in nouns, and the use of verbal auxiliaries instead of single word tense forms, French is more like English, a distant cousin within the Indo-European family, than it is like Latin, its immediate progenitor (compare French *j'ai donné,* English "I have given," Latin *dedī*). The two sorts of language classification, historical and typological, serve different purposes and are differently based. Language families group languages together on the basis of descent; *i.e.,* unbroken transmission from an earlier common parent language. The evidence is, in the main, systematic correspondences among the shapes of words of similar meanings (*e.g.,* Greek *patēr,* Latin *pater,* French *père,* German *Vater,* English "father"). Languages are put into typological classes, with the reservation already mentioned, on the basis of certain overall similarities of structure irrespective of historical relations. Though these two classifications may coincide with some languages, as is the case to a great extent in the Bantu family, they do so only contingently; being based on different data and oriented differently, they do not logically or necessarily imply each other.

In a way, these two systems of classification involve the two most important aspects in which languages must be seen for them to be properly understood: as products of a continuous historical process and also as self-sufficient systems of communication in any one period. Both as a component of cultural history and as a central part of culture itself, language is able to reveal, more than any other human activity and achievement, what is involved in mankind's distinctive humanity. (Ro.H.R.)

BIBLIOGRAPHY. A bibliography for LANGUAGE is likely to overlap at least partially with the bibliography for LINGUISTICS. This bibliography draws attention to some books that may usefully be consulted without prior specialist knowledge, and that develop in further detail the major topics introduced in the article.

EDWARD SAPIR, *Language* (1921), one of the most attractive books on language ever written; LEONARD BLOOMFIELD, *Language* (1933), longer and more technical than Sapir's book, but a classic work that remains unmatched in its breadth of coverage, though Bloomfield's treatment of semantics is now felt to be somewhat dated through his strict adherence to behaviourist principles; JEAN F. WALLWORK, *Language and Linguistics* (1969), a short, simple, and modern introduction to the study of language, and *Language and People* (1978), a brief discussion of the social dimensions of language; HERBERT H. and EVE V. CLARK, *Psychology and Language* (1977), an introduction to theories of language acquisition; DWIGHT L. BOLINGER, *Aspects of Language* (1968), very useful wide-ranging survey of current approaches to the subject; DAVID ABERCROMBIE, *Elements of General Phonetics* (1967), and PETER N. LADEFOGED, *A Course in Phonetics* (1975), excellent introductions to the linguistic study of human speech; JOHN LYONS, *Introduction to Theoretical Linguistics* (1968), deals with linguistics rather than with language, but is an important textbook that introduces the reader to some of the most significant developments in the theory of grammar and semantics, as does the same author's later work *Language and Linguistics: An Introduction*; N. MINNIS (ed.), *Linguistics at Large* (1971), a collection of papers on different aspects of language, several of which treat at greater length some of the topics mentioned in the article; ROBERT H. ROBINS, *A Short History of Linguistics*, 2nd ed. (1979), a brief historical account of the study of language from antiquity to the present day. DELL HYMES (ed.), *Language in Culture and Society: A Reader in Linguistics and Anthropology* (1964), contains a number of useful articles on the relations between language and man's life in society. ROY HARRIS, *The Language Myth* (1981), explores the relations between language and thought; GEORGE A. MILLER, *Language and Speech* (1981), tries to explain language from the point of view of biology; ERIC WANNER and LILA R. GLEITMAN (eds.), *Language Acquisition: The State of the Art* (1982), researches how children acquire language; HANS AARSLEFF, *From Locke to Saussure: Essays on the Study of Language and Intellectual History* (1982), challenges established language theories; DAVID LIGHTFOOT, *The Language Lottery: Toward a Biology of Grammars* (1982), examines the place of language in the system of human cognition and perception; JEREMY CAMPBELL, *Grammatical Man: Information, Entropy, Language, and Life* (1982), addresses language and information theory; DEREK BICKERTON, *Roots of Language* (1981), examines origins of languages; and GRAHAM D. MARTIN, *The Architecture of Experience: A Discussion of the Role of Language and Literature in the Construction of the World* (1981), is a special study.

Languages of the World

Languages may be classified either genetically or typologically. A genetic classification assumes that certain languages are related in that they have evolved from a common ancestral language. This form of classification employs ancient records (such as those for Latin) as well as hypothetical reconstructions of the earlier forms of languages, called protolanguages. Because information on the genetic affiliations of languages is sufficiently extensive, world surveys of languages are necessarily oriented in that way—sometimes exclusively so and sometimes in conjunction with typological classifications. Typological classification is based on similarities in language structure. There is not enough known concerning individual frames of reference in language typology to permit a worldwide typological classification.

Before the conclusive demonstration that unwritten languages could be classified genetically, they were often relegated to a typological classification, which at one time was denigrated by scholars. Since 1917, however, the prestige of some kinds of typology has risen—in particular, that of grammatical typology. The best known typological frame of reference represents the grammar of a language, either as a whole or as a subsystem. Once a genetic classification has been established, typological classification may be superimposed on it in order to show change of language type, as from a predominantly inflectional language (*e.g.,* Proto-Germanic) to a predominantly isolating one (modern English), or to show features that are shared by languages in neighbouring branches in the same family (*e.g.,* Celtic and Germanic in Indo-European). The ultimate grammatical typology is that which treats subsystems that are, in some sense, universal to all human languages.

Lexical typologies, based on similarities in vocabulary structure, have been used in cognitive anthropology and psycholinguistics (*e.g.,* perception of colours and use of colour terms). The sociolinguistic frame of reference in typology provides classifications for varieties of language in terms of their functions and their ways of identifying social groups and cultural spaces; in addition, it brings order and integration to problems concerning national standards that are faced by new nations that have many nonstandard and nonwritten languages as well as languages that make use of writing.

A few points of terminology should be explained before surveying and classifying the world's languages. Language family is the label often used for a conservative genetic classification, one that can be attested only when an abundance of cognates (related words) is available. Phylum is the label for a liberal genetic classification that is attested with fewer cognates; it encompasses language families. Although a given phylum will have greater extension than any of the families included in it, only fragments of phonology will be reconstructable in the protolanguage. In actual linguistic usage, however, the term family is often employed to refer to a phylum; *e.g.,* the Hamito-Semitic family, the Sino-Tibetan family.

The label language isolate is used for a language that is the only representative of a language family, as Basque or the extinct Sumerian language; the presumptive but unknown sister languages of isolates are dead and unrecorded. A language isolate may be classified, along with normal language families, under the rubric of an extensive phylum (*e.g.,* Korean is sometimes classified as a member of the Ural-Altaic phylum) or left wholly unclassified (*e.g.,* Ainu in Japan). The label pidgin-creole is used for a language that has had so much vocabulary change that cognates for reconstructing the protolanguage from which it descended cannot be found. A pidgin is a contact language used for communication between groups having different native languages. When a pidgin becomes the native language of a community it is customarily called a creole.

This article begins with a survey of world languages based on geographical regions of unequal size: huge and sprawling areas for the peripheral regions of Africa, Oceania, and the Americas but relatively compact areas for the focal regions within the Euroasiatic world. Nine regions—six in Eurasia, in all of which writing and standard languages are widespread—constitute a convenient basis for comparison and contrast. The larger part of the article consists of more detailed examinations of the languages of the world arranged by genetic affinities.

This article is divided into the following sections:

INTRODUCTORY SURVEY

LANGUAGES OF EUROPE

The great majority of the languages spoken in Europe are of Indo-European and Uralic (especially Finno-Ugric) affiliation. In terms of numbers of speakers, the people in Europe who speak the languages of these families are now fewer than those in non-European countries who also speak such languages. If a language is to be localized primarily in the region in which most people speak it, then Europe is no longer the chief locale of Spanish, for example, but rather Latin America.

An unusually small degree of genetic diversity is found among European languages: there are fewer language families in Europe than in any other continental-sized region of the world. In addition, literary traditions that have resulted in the preservation of earlier forms of present-day languages are found to a high degree among these languages. Every European language with a writing tradition has developed at least one standard that is recognized nationally, and the national standard often coexists with recognized regional standards.

A few European languages are used internationally, as lingua francas, but there is a low degree of pidgin-creole usage in Europe today.

Typological classifications have been superimposed on genetic classifications of European languages in particular. For example, the Italic branch of Indo-European languages may be grouped with the Greek, Celtic, and Germanic branches on the basis of certain structural features, as can Armenian with Greek and Indo-Iranian, and so forth.

Variables in determining the number of separate languages

Indo-European languages. The languages of seven of the nine extant branches of the Indo-European language family are spoken in Europe. Variability in determining the number of particular languages reflects variation in the criteria used (*e.g.,* mutual intelligibility between neighbouring dialects, known common or separate history versus sociocultural factors such as separate literary traditions or status as national languages of politically independent units), as well as the time period for which the criteria are applied. Thus, for example, it is possible to say on linguistic grounds that there are nine extant languages in the Romance subgroup of the Italic branch: Portuguese, Spanish, Catalan, French, Romansh, Ladin, Friulian, Italian, and Romanian. In applying the criterion of separate literary tradition, the list would be expanded by the addition of Provençal and Sardinian. To apply the criterion of status as a national language would reduce the list because Ladin, Friulian, Provençal, and Sardinian are not national languages; but the picture is complicated by the fact that Sardinia was once politically independent, and Andorra, in which Catalan is spoken, has not always been independent.

Similar difficulties in counting separate languages exist for all the branches in which several languages are spoken. In terms of areas of high mutual intelligibility (which do not entirely reflect historical development), there are only five modern Germanic languages: English, Frisian, Netherlandic-German (including Afrikaans and Yiddish), Insular Scandinavian, and Continental Scandinavian. If literary tradition and national criteria are considered, the number is increased by the division of Netherlandic-German into Standard High German, Low German, Dutch-Flemish, Afrikaans, Luxemburgian, and Yiddish; the division of Insular Scandinavian into Icelandic and Faeroese; and the division of Continental Scandinavian into Norwegian (New Norwegian, or Nynorsk, and Dano-Norwegian, or Bokmål), Danish, and Swedish.

For the Slavic languages there are 13 literary standards, but between the nuclei formed by these norms there are scarcely any linguistic boundaries because transitional dialects connect adjacent areas. In terms of intelligibility and to some extent in terms of shared features, the Slavic literary norms can be grouped into three zones: East Slavic (Russian, Belorussian, Ukrainian), West Slavic (Polish, Kashubian, Low Sorbian or Lower Lusatian, High Sorbian or Upper Lusatian, Czech, Slovak), and South Slavic (Slovene, Serbo-Croatian, Macedonian, Bulgarian).

Language boundaries are more clear-cut for the modern living languages in the remaining European branches of Indo-European: Celtic (the physical separation of the speakers of the languages contributes to the separate identification of Welsh, Breton, Irish Gaelic, and Scottish Gaelic), Baltic (the literary and political separation coincide with the separation of Lithuanian and Latvian), Greek (the separate historical development and lack of mutual intelligibility separate Modern Greek from Tsakonian), and Albanian (the political unity of Albania contributes to the single-language identification of its two divergent dialects).

Dialects of two languages in the Indo-Iranian branch of Indo-European are or were also spoken in Europe: the Jassic dialect of Ossetic, an Iranian language, formerly spoken in Hungary; and the European dialect of Romany, which was spread by Gypsies all across Europe and into America. It may be, however, that only in Wales, Finland, and the Balkans does Romany still serve as a native language.

Extinct languages

A number of earlier Indo-European languages that died out without descendants are known from written records and comments by contemporaries; these include several in the Italic branch (as Oscan, Umbrian, Faliscan, Venetic), at least two in the East Germanic group (Gothic and that spoken by the Burgundians, Vandals, and others), and three in the Celtic branch (Gaulish, Cornish, and the only recently extinct Manx). The classification as Indo-European or non-Indo-European for many other extinct languages remains uncertain because of the scarcity of data; *e.g.,* Pictish spoken in Great Britain. One form of Minoan, that represented by the Mycenaean Linear B syllabary, was shown to be an archaic Greek dialect when deciphered.

For more information on the Indo-European languages of Europe, see below *Celtic languages; Italic languages; Romance languages; Baltic languages; Slavic languages; Germanic languages; English language; Albanian language.*

Finno-Ugric languages. In addition to the Indo-European languages, all but two of the languages of the Finno-Ugric branch of the Uralic family are also spoken in Europe. As in the case of Indo-European languages, variation exists in the enumeration of separate languages. For example, several varieties of Lapp are mutually unintelligible but are often classified as dialects of a single Lapp language. The various types have also been classified according to geographic areas or national boundaries (Norwegian, Swedish, Finnish, and Russian Lapp). In Baltic-Finnic, the Finno-Ugric subgroup most closely related to Lapp, the Finnish, Karelian, Veps, Ingrian, Estonian, Livonian, and Votic languages are often linked by transitional dialects between the central areas of a given pair. The other Finnic languages, Mari (Cheremis) and Mordvin, and both of the languages of the Permic subgroup (Udmurt, or Votyak, and Komi, or Zyryan) are spoken much further to the east, in the central area of eastern European Russia.

Geographic classifications

To the north of these, from the mouth of the Northern Dvina River eastward into North Asia there are speakers of Nenets (Yurak), a language belonging to the Samoyedic branch of Uralic. The remaining Finno-Ugric language of Europe, which belongs to the Ugric subgroup, is Hungarian. See below *Uralic languages.*

Other languages. *Maltese.* Maltese, spoken on Malta, is an Arabic dialect, so long isolated from other dialects of Arabic and so heavily influenced by Italian that the resultant loss of mutual intelligibility with other Arabic speakers might justify classifying it as a separate Semitic language.

Basque. Basque, spoken in the Pyrenees in Spain and France, is the only other living language of western Europe that does not belong to the Indo-European family. Numerous inconclusive attempts have been made to link Basque genetically with other languages. See below *Language isolates: Basque language.*

Turkic languages. In addition to Turkish, spoken by a number of people in Bulgaria and elsewhere in the Balkans, several languages of the Turkic language group (considered as a branch of the Altaic language grouping) are spoken entirely in eastern Europe. Chuvash, the most divergent Turkic language, is found mainly in the Chuvash Autonomous Soviet Socialist Republic, Tatar in the Tatar A.S.S.R. and adjacent areas and in Romania and Bulgaria, Bashkir in the Bashkir A.S.S.R., Gagauz in the Ukrainian S.S.R. and Moldavian S.S.R. and in the Balkans, and Karaim in the southern Ukraine and Lithuania (spoken by about 6,000 people). Most or all of the speakers of Crimean Turkish were removed to the Uzbek S.S.R. after World War II. See below *Altaic languages.*

Extinct languages. The existence of a number of long-extinct non-Indo-European languages of Europe is known through the records of the Greeks and Romans and also through the preservation of varying amounts of written records of them. The most extensive records are those in the still undeciphered Etruscan, which is known to have been spoken in Italy from the 8th century BC to the 4th century AD (see below *Language isolates: Etruscan language.* Several languages were spoken in the Iberian peninsula, of which Iberian (preserved in a few inscriptions and many coins) was spoken along the Ebro River and at one time as far east as the Rhône River. Also probably non-Indo-European was the language of the undeciphered Minoan Mycenaean Linear A inscriptions found on the island of Crete.

LANGUAGES OF SOUTH ASIA

The genetic classification of the languages of India, Bangladesh, Pakistan, and the border states (*e.g.,* Nepal, Sikkim, and Bhutan) includes two subgroups—Indo-Aryan (Indic) and Iranian—of a single branch of Indo-European (called Indo-Iranian), some indigenous language families (as Dravidian), a few language isolates (as Burushaski), and some Sino-Tibetan languages.

Indo-Iranian languages. Except for Romany and the few Dardic languages spoken in Afghanistan, all of the languages of the Indo-Aryan (Indic) subgroup of the Indo-Iranian branch of Indo-European are spoken in South Asia. It is difficult to identify language boundaries in the Indo-Aryan group because, between any pair of literary standards, "transitional" dialects grade into one another, with no clear-cut language barriers. The problem is further complicated by the enormous dialect differentiation in most of the Indo-Aryan languages. In terms of lack of mutual intelligibility between literary standards, there are more than 20 Indo-Aryan languages. Although Sanskrit is a classical Indo-Aryan language, preserved in writing, it also enters so deeply into the vocabulary of present-day languages as to become, in some cases, the salient mark differentiating two dialects of one language. Thus, Hindi of India differs linguistically from Urdu of Pakistan chiefly in that the former may be heavily Sanskritized in vocabulary and the latter not.

Indo-Aryan (Indic) languages

Ethnolinguistic loyalties may also increase the number of languages distinguished; *e.g.,* the separate recognition of Bengali and Assamese. Most of the Indo-Aryan languages are spoken by many millions of speakers; *e.g.,* Bengali-Assamese, West Hindi, Bihari, East Hindi, Marathi, Lahnda, Maithili, Gujarati, Oriya, Sinhalese (in Sri Lanka [formerly Ceylon]), Sindhi, and Nepali. There are also large numbers of speakers of Indo-Aryan languages (especially West Hindi) in South Africa, the South Pacific, and the American Guianas.

The languages of the Dardic subgroup differ sufficiently from the other Indo-Aryan languages as to be sometimes classified as Iranian rather than Indo-Aryan or as a separate sub-branch coordinate with the Indo-Aryan and Iranian sub-branches. Kashmiri, spoken in Jammu and Kashmir, is the only Dardic language with a literary tradition. Shina is also spoken in Jammu and Kashmir; other Dardic languages, which are spoken mostly in Pakistan, have relatively few speakers.

Speakers of at least four Iranian languages are also found in South Asia, including Pashto speakers in Pakistan and Baluchi speakers in Pakistan and India (see below *Indo-European languages: Indo-Iranian languages*).

Dravidian languages. Although the greatest concentrations of Dravidian speakers are in southern India, the more than 20 languages of this family are widespread in India, and one language, Brahui, is isolated in Pakistan, separated from its nearest sister language by 800 miles. Four Dravidian languages have long literary traditions and are spoken by many millions: Telugu, Tamil, Malayalam, and Kannada. Tamil speakers are also found in Sri Lanka, Malaysia, Indonesia, Burma, Vietnam, South Africa, and in scattered island and coastal areas around the world. Among other, less widespread Dravidian tongues of India are Gondi, Tulu, Kurukh, and Kui. No convincing remote relationships between the Dravidian family and other families have been proposed (see below *Dravidian languages*).

Munda languages. The 16 or so Munda languages are all spoken in India. Some scholars classify them as a separate language family; others point to similarities with certain languages of Southeast Asia and include them in an Austro-Asiatic grouping with Mon-Khmer, Vietnamese, and Nicobarese. Santali is the Munda language with the greatest number of speakers (a few million); Mundari, Ho, Sora, Kharia, and Korku have significantly fewer speakers (see below *Austro-Asiatic languages*).

Nehali, Khasi, Burushaski

Other languages. Nahali, spoken by a few hundred people in the Nimar District of Madhya Pradesh, and Khasi, spoken by about 500,000 people in the Khāsi and Jaintia Hills District of central Assam, may both be language isolates—*i.e.,* the sole known members of their families. Both seem to be remotely related to the Austro-Asiatic languages, however, and some scholars have tentatively classified Nahali as a Munda language. The other language isolate of South Asia, Burushaski, spoken by some 30,000 people in Pakistan, is without even remote known relatives.

Sino-Tibetan languages. Speakers of languages of most of the branches of the Sino-Tibetan language family are to be found in South Asia. All the languages of the Bodo (Bodo-Garo) branch of the Bodo-Naga-Kachin language group are spoken in Assam. Naga (Tangsa) languages are spoken in scattered locations from eastern Nepal into Burma. The Kachin languages are centred in north Burma, but some dialects are spoken in Assam, where there are also some speakers of Kukish (Kuki-Chin) languages and of Burmish (Burmese-Lolo) languages. Dialects of the various divisions of the Tibetan language are distributed from Kashmir to Bhutan and southward into India (*e.g.,* Balti, Sharpa, Lhoke, Spiti). Speakers of close to 50 Gyarung-Mishmi (or Himalayan) languages are found from northeastern India to northern Assam, with their greatest concentration in Nepal. There is some scholarly disagreement as to how the numerous Sino-Tibetan languages should be classified into branches and groups (see below *Sino-Tibetan languages*).

Tai languages. Some speakers of Khamti, a Tai language, live in Assam, where Ahom, another Tai language, is still used as a ceremonial language in religious rituals but is no longer spoken.

Use of English. In all parts of postcolonial South Asia, including Sri Lanka (formerly Ceylon), some people know English; these speakers, although relatively few in number, are the people most likely to travel to a state in which a South Asian language unknown to them is spoken. Hence, English is de facto the current interstate and international language of South Asia, although many Indians would prefer to adopt another language, such as Hindi or a Dravidian language, as the national language.

LANGUAGES OF NORTH ASIA

The languages of North Asia are those spoken from the Arctic Ocean to South Asia and China and from the Caspian Sea and Ural Mountains in the west to the Pacific Ocean in the east. In genetic classification, most languages of North Asia belong either to the Uralic family, to one of the three families of the Altaic language grouping (Turkic, Mongolian, and Manchu-Tungus) or to Indo-European. The genetic affiliations of the Paleosiberian languages, spoken exclusively in this region, are uncertain at present. Scholars have hypothesized that some of the languages may have once been American Indian languages whose prehistoric speakers backtracked from the New World into North Asia. That it is possible for pre-industrial man to go back and forth over Arctic waters is shown by the residence of those Eskimos who are now found on both the Russian and Alaskan shores of the Bering Strait. It has been claimed that all languages indigenous to North Asia, except the Paleosiberian ones and the recently intrusive Russian language, are genetically related in a Ural-Altaic phylum. This liberal classification is questioned by many scholars. Whether or not the three Altaic language groups are related to the Uralic languages, there is no doubt that all the so-called Ural-Altaic languages share many typological features, such as vowel harmony, agglutination (a type of word formation in which word elements are added together but still retain a separate, definite meaning), and a restriction against combining a plural noun with a quantifier (as though, in English, the plural noun "girls" had to appear as a singular noun in the phrase "five girl," rather than "five girls").

Ural-Altaic phylum

Uralic languages. Languages from two branches of the Uralic family are spoken in North Asia; their speakers are few in number. Two of the Ugric languages (the subgroup of the Finno-Ugric branch that includes Hungarian)—Mansi (Vogul) and Khanti (Ostyak)—are spoken on the Ob River and its southwestern tributaries. All of the languages of the Samoyedic branch are spoken in North Asia: Nenets (Yurak), speakers of which are scattered from the mouth of the Yenisey westward to the mouth of the Northern Dvina; Enets (Yenisey), also spoken around the Yenisey; Nganasan (Tavgi), spoken on the Taymyr Peninsula in Siberia; and Selkup (Ostyak), spoken south of Enets between the Taz and Tym rivers. Another southern Samoyedic language, Kamas (Sayan), had only one speaker in 1963 (see below *Uralic languages*).

Altaic languages. *Turkic languages.* The Turkic languages are remarkable for their lack of diversity in spite of their wide occurrence in all the Euroasiatic regions except South Asia and Southeast Asia. Several of the numerous Turkic languages might be considered as a single language if it were not for the fact that mutual intelligibility between groups is impaired by differential borrowing from the various unrelated languages encountered in different regions. Thus, a Turkish speaker from Turkey might understand or largely understand an Uzbek speaker with more ease than the Uzbek speaker would understand Turkish, which has many loanwords from Persian and Arabic. Educated speakers of Turkic languages are able to read books in other Turkic languages after some adjustment to their varying spelling conventions and sound correspondences. Because such differences are identified with different Turkic ethnic groups, it is customary to identify the larger of these ethnic groups as speaking different Turkic languages, although some degree of intelligibility exists between them, as it does between Uzbek, Bashkir, and Tatar. In addition, some languages have dialects that are transitional between two recognized language groups; *e.g.,* some dialects of Kara-Kalpak are said to be transitional to Turkmen, and others are said to be transitional to Uzbek.

Distribution of Turkic in North Asia

In North Asia, Turkic languages are distributed from the southern extension of North Asia northeastward through central Siberia and include Turkmen in the Turkmen S.S.R., Iran, and Afghanistan; Uzbek in the USSR, mostly in the Uzbek S.S.R., and in Afghanistan; Kirgiz in the Kirgiz S.S.R. and in neighbouring areas from Afghanistan to China; Kara-Kalpak in the Kara-Kalpak A.S.S.R.; and Kazakh in the Kazakh S.S.R. Six or seven Turkic groups immediately north of western Mongolia are much smaller both in terms of numbers of speakers and the area over which they spread; in northern Siberia the Yakut extend from the Yakut A.S.S.R. to include the dialect spoken to the west by the Dolgan. Other Turkic languages are spoken in Southwest Asia, Europe, and East Asia.

Mongolian languages. The Mongolian, or Mongol, languages are dispersed throughout Central Asia from Afghanistan to Manchuria, occupying large parts of North Asia and East Asia. The problem of recognizing language boundaries (*i.e.,* of distinguishing separate languages) in the Mongolian family is complicated by the fact that differences between dialects are exaggerated in areas where Mongolian speakers have borrowed features of different unrelated languages but are minimized in areas where one dialect is spoken as a lingua franca over a wide area. Among the Mongolian languages are Mogol, spoken in Afghanistan, where it has been influenced by Iranian and Turkic languages; Monguor, spoken in Kansu Province of China and in Tibet, with noticeable effects of both Tibetan and Chinese in the language (Paoan, spoken in Kansu Province, is linguistically close to Monguor, but may be a different language); and Daghur, spoken mainly in Inner Mongolia and heavily influenced by Tungus languages. Additional languages include Ordos in Inner Mongolia, Kharachin in China, Oyrat in the Sino-Russian border area from the Kirgiz S.S.R. to the Altai Mountains, and Buryat from the Buryat A.S.S.R. into Inner Mongolia. Some degree of mutual intelligibility exists between some of these, but this may in part be the result of the lingua franca use of Khalkha, the official language of the Mongolian People's Republic.

Manchu-Tungus languages. Speakers of the Manchu-Tungus languages are scattered from central interior Siberia to the shores of the seas of Japan and Okhotsk, including the Kamchatka Peninsula and Sakhalin Island. Those

Difficulties in Manchu-Tungus classification

not near the coast live generally along the banks of the major rivers—the Yenisey, Tunguska, Khatanga, Lena, Amur, and Sungari. Detailed information on most of the Manchu-Tungus languages is scanty, and language names usually coincide with politico-cultural groups, rather than being based on a comparison of linguistic features or knowledge of mutual intelligibility. Borrowing that resulted from contact with speakers of Samoyedic (Uralic) languages to the west and northwest, Mongolian languages and Chinese to the south, and the various Paleo-Siberian languages to the north and east has further complicated the subclassification of the Manchu-Tungus languages by increasing the superficial differences among them. Most speakers of Manchu-Tungus languages are bilingual in the official language of their country, and many are replacing their native languages by Russian or Chinese. After the Manchus in China, the next best known and numerous of the Manchu-Tungus peoples are the Evenks (whose name is sometimes applied as a generic term for all the Tungus tribes). Other groups include the Evens (or Lamuts), Nanais (or Golds), and other tribes with only 2,000 or fewer members. For more information on the Turkic, Mongolian, and Manchu-Tungus languages, see below *Altaic languages.*

Paleo-Siberian languages. Most of the so-called Paleo-Siberian people now live in northeasternmost Siberia in the area between the East Siberian Sea and the Sea of Okhotsk, including the Kamchatka Peninsula, and along the coast of the Sea of Okhotsk as far south as the Amur River, and on Sakhalin Island; peoples of another Paleo-Siberian group live far to the west along the middle and upper Yenisey River. The languages of the Paleosiberian people form four groups that are not only not related to each other but also have not been demonstrated to be related to any other genetic groups. The northernmost and most widespread of these linguistic groups and the only one that includes more than one living language is the Luorawetlan family, which consists of Chukchi, Kamchadal, and Koryak. Some scholars now classify Kerek and Aliutor as separate languages; these have otherwise been classified as dialects of Koryak.

The Yukaghir family includes one living language, Yukaghir (spoken by a few hundred people south of the Arctic Circle on tributaries of the Kolyma River and in the tundra between the Indigirka and Alazeya Rivers), and one language—Chuvantsy—that was spoken until the 20th century on the Anadyr River. Gilyak, spoken on Sakhalin Island and in the coastal and inland Amur River country of the mainland, has no known linguistic relatives. Ket (or Yenisey-Ostyak) is the only language of the Yeniseian or Yenisey-Ostyak family that is still spoken. The speakers of Ket live along the upper and middle Yenisey River, as did the speakers of its sister languages, Kott (Cottian-Manu), which became extinct in the 19th century, and Assan (Asan) and Arin, both of which became extinct in the 18th century (see below *Paleo-Siberian languages*).

Indo-European languages in northern Asia, in addition to Russian, introduced only relatively recently, include the Iranian languages in the southwestern extension of North Asia (Tadzhik Persian in the Tadzhik S.S.R. and Baluchi in the Turkmen S.S.R.) and the long-extinct Tocharian, which penetrated into Central Asia as far as Chinese Turkistan (see below *Indo-European languages: Tochavian language*).

Role of Russian and Soviet policy

Writing and literacy in North Asia. The earliest stimulus toward writing in North Asia was from China. The latest stimulus, from Soviet Russia, has brought literacy to those Altaic peoples whose languages were unwritten in tsarist times. In contrast to the written Altaic languages, the Paleosiberian languages in general remain preliterate. The Soviet educational policy is to encourage the use of native languages for education and for teaching preliterates to write. The standard form of writing Tadzhik, for example, is in the Cyrillic alphabet, and knowledge of this alphabet facilitates later learning of Russian, which is used for supranational communication in modern North Asia. Hence, Russian is the modern lingua franca for North Asia today. In the Mongol Empire of the 13th century, Turkic languages were used as languages of administration across North Asia from the Caspian Sea to Manchuria and, initially, in adjacent Euroasiatic regions conquered by Genghis Khan.

LANGUAGES OF SOUTHWEST ASIA

Languages spoken today in the area from Iran westward to the Mediterranean (in Iran, Iraq, Saudi Arabia, Jordan, Syria, Lebanon, and Israel) are Semitic, Indo-European, or Turkic. The languages in the two marginal sub-areas of Southwest Asia (in Afghanistan and in Turkey and the Caucasus between the Black and Caspian seas) far exceed the languages of Europe in genetic diversity.

At one time, Sumerian, now preserved in writing, was spoken as the first language of civilization in the ancient Near East; this language was neither Semitic nor Indo-European (see also *Sumerian language*). Early literary traditions and literacy for the elite began in this central area of Southwest Asia and extended from the Sumerian, Old Persian, and Akkadian literatures to Asia Minor (Hittite) in the north and to the Nile (Egyptian) in Africa. Akkadian and Persian seem to have been the first two languages put to wide international use.

Indo-Iranian languages. Almost all of the score of living languages of the Iranian subgroup of the Indo-Iranian branch of Indo-European are spoken in Southwest Asia and occasionally extend beyond into neighbouring regions. Persian has three separate literary standards that are not confined to the countries in which they centre (Iran, Afghanistan, and Tadzhik S.S.R.). More than half of the speakers of Pashto live in Afghanistan and the rest in South Asia. Kurdish is spoken in an area extending southward from southern Armenian S.S.R. into Turkey, Syria, Iran, and Iraq. Perhaps three-fifths of the speakers of Baluchi live in Iran and southern Afghanistan. Several other Iranian languages (or dialects) have many fewer speakers; these include Lurī and Bakhtyārī, spoken only in Iran, and Munjī and Shughnī, spoken largely in Afghanistan, with only a few of their speakers in Pakistan or the Tadzhik S.S.R. One Iranian language, Yaghnābī, is spoken only in the Tadzhik S.S.R. Three Iranian languages are spoken almost entirely in the Caucasus: Tat, Talysh (with some speakers in Iran), and Ossetic.

The Dardic subgroups

The half a dozen Nuristani (Kafiri) languages spoken in Afghanistan, sometimes classified as members of the Dardic subgroup of Indo-Aryan, have more recently been classified by some scholars as constituting a separate branch of Indo-Iranian. In addition, some Lahnda (Indo-Aryan) speakers also live in Afghanistan. Two very divergent dialects of another Indo-Aryan language, Romany, are spoken in Southwest Asia—Armenian Romany and Asiatic Romany (the dialect of the Palestinian Gypsies). For more information on the Iranian and Indo-Aryan languages, see below *Indo-European languages: Indo-Iranian languages.*

Other languages. The sole language of another branch of Indo-European, Armenian, is spoken predominantly in the Armenian S.S.R., Georgian S.S.R., and Azerbaijan S.S.R., but also in Syria, Lebanon, Iran, Turkey, and in communities of Armenians in the United States, Egypt, and France (see below *Indo-European languages: Armenian language*).

The long-extinct languages of the Anatolian branch of Indo-European, including Hittite, were once spoken in Southwest Asia (see below *Indo-European languages: Anatolian languages*).

Five Turkic languages are spoken primarily in Southwest Asia: Turkish, spoken in Turkey and surrounding countries largely to the north; Azerbaijani, spoken in the Azerbaijan S.S.R. and Iran; Kumyk, Karachay, and Nogay, spoken in the Caucasus. Three Turkic languages spoken predominantly in North Asia are also spoken in Southwest Asia: Uzbek, Turkmen, and Kirgiz.

One language of the Mongolian family is spoken in Southwest Asia—Mogol in Afghanistan; and Brahui, a Dravidian language, has a small fraction of its speakers in Afghanistan and Iran.

Caucasian languages. In addition to the Indo-European and Turkic languages spoken in the Caucasus, there are the over 30 languages belonging to three Caucasian language families. These may be remotely related to each other in a Caucasian phylum, in which the Northeast Caucasian family is more clearly related to the Northwest Caucasian family than the South Caucasian family is to either. Georgian, a South Caucasian language, is the most widely known Caucasian language, with speakers in the Georgian S.S.R., the Azerbaijan S.S.R., and adjacent parts of Turkey and Iran; it is the only Caucasian language with a long literary tradition. Other South Caucasian (Kartvelian) languages are Laz (Chan) and Svan. The Northwest Caucasian (Abkhazo-Adyghian) languages include Kabardian (Circassian), Abkhaz, Abaza, Adyghian, and Ubykh (almost extinct, now spoken by only a few people in Turkey). The approximately 25 languages of the Northeast Caucasian (Nakho-Dagestanian) family are spoken by people living mostly in the Dagestan A.S.S.R. These languages include Chechen, Ingush, Avar, Dargwa, Lakk, Lezgian, and Tabasaran, all of which (except Chechen and Avar) have fewer than 500,000 speakers. There is some scholarly disagreement concerning the classification of the Caucasian languages (see below *Caucasian languages*).

Semitic languages. Five Semitic languages are still spoken in Southwest Asia: Arabic; Hebrew, primarily in Israel; dialects of East Aramaic, still spoken in Israel, Syria, Iran, Iraq, and the Armenian S.S.R.; West Aramaic dialects, still spoken in Lebanon and Syria; and Modern South Arabic, spoken in southern Saudi Arabia and on nearby islands. Of the extinct Semitic languages, the best known are Phoenician, Akkadian (Babylonian and Assyrian), Moabite, and Ugaritic (see below *Hamito-Semitic languages*).

LANGUAGES OF EAST ASIA

Conservative versus liberal classification of the languages

Languages in East Asia are those traditionally spoken in China, Japan, and Korea; *i.e.,* those that occupy the region between North Asia and Southeast Asia. A conservative genetic classification reflects immense genetic diversity for East Asia by claiming that Ainu, Japanese, and Korean are neither related to each other nor to any other language in East Asia and that the Chinese languages (or dialects) belong in one family, Miao-Yao languages in another, and Tai languages in still another. A liberal genetic classification leaves Ainu isolated, includes Korean and Japanese in the Altaic family, and classifies some or all of the other groups as Sino-Tibetan.

Altaic languages. Languages from three of the major families of North Asia are spoken in China. Uighur, a Turkic language, is spoken in Sinkiang and Kansu Provinces of China as well as in the U.S.S.R. and southwestern Mongolia. Another Turkic language, Kirgiz, has some speakers in China. Manchu is the best known of the Manchu-Tungus languages and that with the longest literary tradition (dating from as early as 1599). After the Manchus established the last Chinese dynasty in 1643, their language was gradually replaced in most parts of China by Mandarin—except for formal and ceremonial occasions—but it is still spoken in scattered localities in Manchuria and in Chinese Turkistan.

Striking similarities in syntax have led some linguists to postulate a remote relationship between the Altaic languages and Korean and, less frequently, Japanese.

Korean, Japanese, and Ainu. Korean is spoken in Korea as well as by sizable populations in China and Japan (see below *Korean language*).

The Japanese language family includes, besides Japanese, several mutually unintelligible dialects spoken on the Ryukyu Islands by people who are bilingual in Japanese. Japanese is spoken by more than 121,000,000 people in Japan and by small groups in Taiwan, Brazil, and the U.S., especially in Hawaii (see below *Japanese language*).

Ainu, the remaining language in insular East Asia for which not even a remote relationship with other languages seems likely, is spoken by approximately 16,000 people on Hokkaido Island of Japan, on Sakhalin Island, and on the Kuril Islands.

Chinese language (dialects). Most important in terms of numbers of speakers and their influence on the other languages in East and Southeast Asia are the Chinese languages (often called dialects). In terms of mutual intelligibility among adjacent dialects, there are several Chinese languages: Mandarin, Wu, Cantonese, Hsiang, Kan-Hakka, and Min (or North Min and South Min). Mandarin is the native language of over 70 percent of the Chinese and is spoken as a second language by many of the native speakers of the other languages, both Chinese and non-Chinese, in China. It has traditionally been the language of administration. Although speakers of two different Chinese languages may not be able to understand one another when they talk, communication between them is possible in writing; conversely, the same written message is read aloud differently by speakers of different Chinese languages. The functional advantages of Chinese writing explains its perseverance for four millennia, but these advantages are partly offset by the difficulties each generation must experience in learning the thousands of character signs that are needed for literacy. Traditionally most Chinese were supposed to be illiterate, but with simplified characters and romanization, the majority of the people in the People's Republic are now literate. The Chinese languages are notable for their enormous numbers of speakers, and Mandarin has the largest number of speakers of any of the world's languages (about 610,000,000 native speakers).

A remote relationship in one family (Sino-Tibetan) has been postulated for the Chinese languages and all the other non-Altaic families that have languages spoken in China. There is no doubt that all these languages bear many similarities to Chinese, but it is unknown to what extent such similarities might be the result of borrowing rather than common origin. A remote relationship in an Austro-Tai phylum has been proposed for two of these families (Tai and Miao-Yao) and Austronesian.

Remote relationships in East Asia

Tai and Miao-Yao languages. All of the languages of the Kam-Sui language group, which is related to the Tai family, are spoken in China (in Kweichow, Hunan, and Kwangsi Provinces), with some dialects extending into Southeast Asia. Speakers of Miao-Yao languages are scattered over south central China and extend into Vietnam, Laos, and Thailand. Dialects of the Miao language include Red Miao, White Miao, Green or Blue Miao, and the more divergent Black Miao. The Yao languages are Yao (also called Man or Mien), Laka, and Punu.

Tibeto-Burman languages. The Tibetan, or Tibetic, language group includes at least two Tibetan proper languages spoken in Tibet, Nepal, Sikkim, and India: Central Tibetan, including Lhasa, the standard dialect of Tibet, and Western Tibetan. In addition there are many other languages in Sikkim, Nepal, Assam, India, and Bangladesh that are closely related to Tibetan proper.

Languages that are more distantly related to Tibetan in a Tibeto-Burman branch of the Sino-Tibetan family are spoken in East Asia over the borders of Burma; these languages, often called Burmic, include dialects of the Burmese-Lolo (Burmish) subgroup (including Burmese) and the Kachin subgroup. For more information on the Chinese, Tibetan, and Burmic languages, see below *Sino-Tibetan language.*

Three general types of syntax, which partly overlap the liberal genetic classification, can be distinguished among languages in East Asia. First, Ainu is isolated syntactically as well as genetically. The second type is shared by Korean and Japanese. All Chinese languages are strikingly alike in syntax, and this third type is approximated among some non-Chinese languages of the Sino-Tibetan family and among some languages of Southeast Asia whose genetic classification is tentatively indeterminate.

LANGUAGES OF SOUTHEAST ASIA (INCLUDING AUSTRONESIAN)

Southeast Asia is generally taken to be a region that includes both a mainland subregion, south of China and east of India, and an insular subregion, which includes the insular half of Malaysia, all of Indonesia, and the islands of the Philippines. Virtually all the languages of insular Southeast Asia belong to a single language family—Austronesian (Malayo-Polynesian). Mainland Southeast Asia, on the other hand, has various representatives from the Austro-Asiatic, Tai, and Sino-Tibetan language groups. Hence, genetic diversity is greater in mainland than in in-

Genetic diversity in mainland Southeast Asia sular Southeast Asia. Austronesian languages extend out of Southeast Asia to the most distant culture areas in Oceania (Polynesia and Micronesia), where they are the only languages known aboriginally. One modern Austronesian language (Malagasy) is even spoken on the African side of the Indian Ocean in Madagascar.

Curiously enough, it is in Melanesia, between the Bismarck Archipelago and New Hebrides, that the most diverse Austronesian languages are spoken today; this provides grounds for the conjecture that the Proto-Austronesian language was spoken there millennia ago and that the daughter languages diversified as their speakers migrated over half the world, with Malay and Cham backtracking eventually to the mainland of Southeast Asia, out of which the ancestors of Proto-Austronesian speakers must have come.

In general, the name of the country and the name of the national language are the same in both insular and mainland regions of Southeast Asia. Thus, Pilipino (based on Tagalog) is the name of the national language of the Republic of the Philippines, even though Pilipino is learned as a second language by most Filipinos. The fear in all of Southeast Asia of indirect neocolonial domination motivates continued distrust of the old languages of colonialism—English, French, Dutch, Spanish—and now also of Japanese and Russian. A pidgin-creole—Neo-Melanesian, or Melanesian Pidgin English—is used as a lingua franca by speakers of Austronesian and other languages from southern Papua through Melanesia into Micronesia.

Though the languages in the mainland subregion of Southeast Asia are genetically diverse, they show widespread ranges of the same typological features, such as the use of distinctive tones and classifiers, among unrelated or remotely related languages.

Mon-Khmer languages

Austro-Asiatic languages. The Mon-Khmer group includes more than 50 languages—more than any other family that is centred primarily in or entirely in Southeast Asia. Mon-Khmer languages are spoken from Burma to Vietnam. In Cambodia, Khmer (Cambodian) is the official language; its speakers are also found in Thailand. Mon is also spoken in Thailand as well as in Burma.

The language of mainland Southeast Asia with the greatest number of speakers is Vietnamese, spoken in Vietnam and by smaller numbers of speakers in Cambodia, Thailand, and Laos. Muong, spoken in the central highlands of northern Vietnam, is recognized as a separate, but related, language.

Classified by some scholars as a northern group of Mon-Khmer languages are several languages spoken in Burma (east of Mandalay), northwest Thailand, northern Laos, and to a lesser extent in northern Vietnam and in China. These are sometimes classified by others as a separate Palaung-Wa, or Salween, family, including Khmu, spoken in Laos and extending into Thailand, and Palaung, spoken in Burma.

Three small groups of related languages in Malaya (sometimes called Malaccan) are considered to be related to the Austro-Asiatic languages. They are the Jahaic, or Semang, languages, spoken in the inland area of northern Malaya and across the border in Thailand; the Senoic, or Sakai, languages, with speakers south of Kuala Lumpur on the coast and inland further south; and the Semelaic, or Jakun, languages, spoken south of the Senoic languages (see below *Austro-Asiatic languages*).

Tai and Sino-Tibetan languages. At least a dozen languages of the Tai language family are spoken in Southeast Asia: Thai, or Siamese, in Thailand; Lao, in Thailand, Laos, and Cambodia; Yuan, in Thailand; Shan, in Burma; Black Tai (Tai Noir), in Laos and Vietnam; Khün and Khamti, in Burma; and White Tai (Tai Blanc), Tay, Nung, Tho, and Kelao (Ch'i-lao), all in Vietnam.

Over 9,000,000 Chinese are distributed throughout Southeast Asia; of these, more than 5,000,000 are in Thailand, 4,000,000 in Malaysia, and smaller numbers in Burma, Cambodia, Vietnam, 54,000 in 1979 and Laos.

Of the other language groups in the Sino-Tibetan family in Southeast Asia, the Burmese-Lolo (Burmish) group has the widest distribution and the greatest number of speakers. Burmese is spoken as a second language by perhaps 90 percent of those in Burma who have another first or native language. The Lolo languages are spoken in Burma, Thailand, Laos, and Vietnam; they include Lisu, Lahu, Akha, Mung, Punoi, Pyen, and others, a few of which extend into Assam. Karen languages are spoken in Burma and Thailand and include Sgaw, Pho, Pa-o (or Taungthu), and Palaychi. Most of the languages of the Kuki-Chin (Kukish) group are spoken in Burma. Kachin languages are also spoken in Burma (see below *Sino-Tibetan languages*).

Nicobarese and Andamanese

Insular language groups. In insular Southeast Asia one small language group, Nicobarese, consisting of the languages spoken on the Nicobar Islands, is, in a liberal classification, classified as Austro-Asiatic. The other insular family, Andamanese, consisting of the languages spoken on the Andaman Islands by perhaps fewer than 200 people, has only recently been supposed to be remotely related to the Papuan languages of Melanesia.

Austronesian languages. There are perhaps 500 languages in the Austronesian (Malayo-Polynesian) family, spoken in Malaysia and the Indonesian archipelago; the Philippines; parts of Vietnam, Cambodia, and Taiwan; on the main island groups of the South and Central Pacific; on New Guinea; and on Madagascar. According to one classification, these languages include, in addition to small subgroups, at least two large subgroups: Western Austronesian (or Indonesian) and Eastern Austronesian (often called Oceanic), which includes the Polynesian languages and some of the Melanesian and Micronesian languages. Those Austronesian languages spoken on the Southeast Asian mainland (Malay in Malaysia, Cham and eight other languages mostly in Vietnam, with speakers of some of them also in Cambodia) belong to a Western Indonesian subgroup, which includes Javanese, Sundanese, and Malay, including Bahasa Indonesia, the national language of Indonesia. Closely related to the Western Indonesian subgroup is the subgroup comprising around 100 languages of the Philippines and a few languages of northern Borneo and northern Celebes (including Tagalog, which includes Pilipino, the national language of the Philippines, and Cebuano, Hiligaynan, and Ilocano). Classed with the West Indonesian and Philippine languages are a small group of languages of Celebes (*e.g.,* Buginese and Makasarese), a few languages of Borneo, and Malagasy (used on Madagascar).

Languages of Polynesia

The languages of Polynesia, including Maori in New Zealand, Tongan, Tahitian, and Hawaiian, form a subgroup that is part of a larger Eastern Oceanic subgroup of over 100 languages, which includes besides the Polynesian languages such languages as Fijian and a number of languages of the Solomon Islands. At least seven of the languages of Micronesia (including Gilbertese, Trukese, and Ponapean) form another subgroup. There are more than 100 Austronesian languages in New Guinea and more than 100 Austronesian languages, not counted as Eastern Oceanic, that are spoken on smaller islands of Melanesia. Those few with as many as 10,000 speakers are all used as lingua francas in wider areas than those of their native speakers (Dobu in the D'Entrecasteaux Islands, Banoni in southwestern Bougainville, Panayati in the Louisiade Archipelago). Among the Austronesian languages still spoken on Taiwan are Ami, Atayalic, Paiwan, and Bunan. There is some scholarly disagreement concerning the classification of the Austronesian languages (see below *Austronesian languages*).

NON-AUSTRONESIAN LANGUAGES OF OCEANIA

In effect, the non-Austronesian language areas of New Guinea and Australia together constitute a wedge in the midst of three Austronesian areas: Polynesia to the east, Micronesia to the north, and Indonesia to the west. A few non-Austronesian languages are found on the Indonesian islands nearest to New Guinea (on Halmahera as well as on Timor and Alor).

An exceptionally liberal genetic classification claims that the many non-Austronesian languages in Melanesia and the few in Indonesia all belong to one phylum. Conservative classifications recognize several or even many different language families and avoid the older name for them (Papuan), because it might suggest either that the unrelated families of non-Austronesian languages are branches of one Papuan family or else that non-Austronesian languages are found only in New Guinea (where Papua is the name of a country). On the other hand, no classification is challenged when it is said that all Australian languages are ultimately related and that they are related neither to Austronesian nor to non-Austronesian languages outside of Australia.

Multilingual societies of Melanesia

In Melanesia, in essence the non-Austronesian world beyond Indonesia, there is much contact between Austronesian and non-Austronesian languages. Many of the Melanesian societies are multilingual, especially those in New Guinea; in addition to their native language, speakers often learn a few secondary languages—those of their immediate neighbours or, most frequently, Neo-Melanesian (a pidgin-creole with an English-based lexicon), or both. In part of Papua, New Guinea, Police Motu, a pidgin based on an Austronesian language, is used as a lingua franca far beyond the territory of the few thousand native speakers of Motu. In Australia the same interest in mastering a multiplicity of languages is widespread, and Aborigines have developed another English-based pidgin-creole, quite different from Neo-Melanesian. Another parallel between Australian languages and the non-Austronesian languages north of Torres Strait is the disinclination of both to recognize or develop any one dialect of a language as a standard.

Papuan languages. About 740 Papuan or non-Austronesian languages extend from the Santa Cruz Islands north and west into the Solomons and the Bismarck Archipelago, across New Guinea to Halmahera, Timor, and Alor. Until the late 1950s all discussions of the languages of New Guinea that treated more than small, closely related groups of languages stressed the fact that the hundreds of languages spoken in a comparatively small area seemed to be completely unrelated to each other except for a few groups of immediate neighbours. Until then, little was actually known about more than a few of the languages of New Guinea. This situation was changed in the 1960s, with the publication of further survey work in the Highlands Provinces, which stated explicit relationships among a large group of languages.

Since the initial recognition of this fairly large group of related languages in the Highlands, called the East New Guinea Highlands phylum, more and more languages of New Guinea have been found to be at least remotely related to it (in a Central New Guinea phylum). It must be kept in mind, however, that it is difficult to enumerate languages and families among the Papuan languages because too little information has been obtained for most of the languages to identify language boundaries or make possible detailed comparisons. There remain a number of families and isolated languages that seem not to be related to other Papuan languages. A new liberal classification presented by the U.S. linguist Joseph Greenberg in 1971, however, treats all the Papuan languages as genetically related in an Indo-Pacific phylum, which also includes Andamanese. Most Papuan languages are spoken by only a few hundred to a few thousand speakers (see also *Papuan languages*).

Indo-Pacific phylum

Australian aboriginal languages. All of the aboriginal languages of Australia are remotely related to each other. A few dozen of the 260 or so Australian languages still spoken account for 90 percent of the total number of speakers; these include the Aranda languages, Tiwi, Walbiri, and Western Desert, two languages spoken by over a score of separately named small groups scattered over a territory 900 miles long. Scores of languages are now spoken by fewer than six people each. The greatest diversity among the languages is found in extreme northern and northwestern Australia (Arnhemland and the Kimberley District); a single remaining family (Pama-Nyungan), with 177 languages, is distributed over the rest of Australia (see below *Australian Aboriginal languages*).

In grammatical typology the non-Austronesian languages north of Torres Strait are heterogeneous, while the Australian languages are syntactically homogeneous and almost identical in patterns of sound combinations. Both Australian languages and non-Austronesian languages have dialects that are linked in a chain such that speakers at either end do not understand the vocabulary of speakers at the other end, although speakers of adjacent dialects can understand each other.

The two or more languages that were spoken on Tasmania until the later part of the 19th century are not related to Australian languages, but may belong to the Indo-Pacific phylum.

LANGUAGES OF AFRICA

European languages in Africa

Languages that came into Africa from another homeland include, among others, all the European languages associated with 19th-century colonialism. Although the majority of countries in Africa regained their freedom in the 1960s, they continue to use the European languages of the colonial period along with the numerous languages indigenous to Africa. Languages from Southwest Asia preceded the languages of European colonization: migrations of peoples to North Africa brought the Ethiopians almost three millennia ago and the Arabic speakers many centuries ago. The Phoenician circumnavigation of Africa in ancient times left traces—Phoenician coins—on the coasts but none in the interior. And long ago migrants from Indonesia reached Madagascar, 250 miles off the African coast. Before and during the colonial period, Arab and Indian traders reached East Africa, where today a few Indo-Aryan languages are spoken among Asian businessmen. The interior of Africa was not known to any non-Africans before the colonial period, but its prehistory can now be partially reconstructed. For example, there is evidence that the homeland of the protolanguage of the numerous Bantu languages was in Cameroon or an adjacent area in West Africa (or in both areas); that a prehistoric migration brought the Bantu speakers to Central and East Africa; and that these Bantu forced the speakers of Bushman and Hottentot languages to leave their homeland around Lake Victoria and move south to the Kalahari.

In all the postcolonial nations today, either English or Arabic or French serves both as an international language and as a functioning national language. The question still unresolved for many African nations concerns which of their indigenous languages to develop through writing and to standardize as the official language or languages of education and of the political state. The numerous pidgin-creoles, as Krio, are recent and colonial in inspiration; Sango in the Central African Republic is surely indigenous but not so surely a pidgin-creole. Most of the dozen or so languages used in trade, as Swahili in East Africa and Hausa in West Africa, tend to have great changes in vocabulary like pidgin-creoles, but they are not classified as pidgin-creoles; instead they are varieties of normal languages that function as lingua francas. Lingua francas of one sort or another are a prerequisite for the markets found throughout tribal and peasant Africa.

Despite the genetic diversity in South Africa and the even greater diversity in West Africa, a part of each of these subregions can be shown, on the basis of typology, to be a linguistic area. Thus, most linguists have found that most languages in West Africa distinguish vocabulary items and word elements by tone; in South Africa the clicks characteristic of Khoisan languages are also found among neighbouring Bantu languages like Xhosa and Zulu. The early use of typology to anticipate genetic classification, however, led to the claim that Africa was full of mixed languages—*e.g.,* Mbugu in Tanzania. But Mbugu, despite having borrowed Bantu prefixes and culture words from Bantu, can be shown to have a single line of origin—to have descended from a single protolanguage (Proto-Cushitic)—on the basis of its grammatical constituents (pronouns and verb forms) and basic vocabulary items that are cognate with other Cushitic languages.

Five branches of the Hamito-Semitic language family in Africa

Hamito-Semitic languages. The Hamito-Semitic (Afro-Asiatic) language family (considered a phylum by some) includes five branches spoken across North Africa from Mauritania to Somalia and beyond into Southwest Asia: Chadic, Semitic, Cushitic, Berber, and the now extinct Egyptian-Coptic. The Chadic branch consists of over 100 languages spoken in Nigeria, Niger, Cameroon, Ghana, Chad, and the Central African Republic. By far the most widespread is Hausa, estimated to be spoken by as many as 22,000,000 people, for many of whom it is a second language.

Five Semitic languages are spoken in Africa, if modern colloquial Arabic is counted as a single language throughout its range across North Africa and the Arabian Peninsula and if Gurage in Ethiopia is also counted as a single language. The Semitic languages in Ethiopia include Amharic, Tigrinya, Tigre, and Gurage (but the people grouped as Gurage may be speaking several separate languages).

Cushitic languages are spoken in Ethiopia, Somalia, The Sudan, Tanzania, and Kenya. The languages with the greatest number of speakers are Gallinya in Ethiopia, Somali in Somalia and Ethiopia, Sidamo, Hadya, and Afar-Saho. Some scholars consider a group of languages traditionally classified as Cushitic to be a separate branch of Hamito-Semitic, called Omotic. Spoken in Ethiopia, they include Walamo, with far more speakers than the other Omotic languages, Ari, Shako, Zaysse, and others with only a few thousand or a few hundred speakers.

The languages of the Berber branch are spoken from the western desert of Egypt west to the Atlantic and extend to Senegal on the coast and to northern Nigeria in the interior. One language that may have been Berber, Guanche, was formerly spoken on the Canary Islands. Berber languages include Shluh, spoken in Morocco; Tamashek (Tuareg) in Algeria, Libya, Niger, and Mali; and Tamazight in Morocco and Algeria (see below *Hamito-Semetic languages*).

Nilo-Saharan languages. The Nilo-Saharan languages in central interior Africa include the Chari-Nile languages and others that are not closely related to each other or to the Chari-Nile group. (The validity of this grouping has been questioned.) The largest Chari-Nile division, Eastern Sudanic, includes more than 60 languages spoken from Chad to Kenya and Tanzania; it includes a group of languages often classified as a separate family or branch (Nilo-Hamitic), which appears in some classifications in the Hamito-Semitic family rather than the Nilo-Saharan.

Among the major Eastern Sudanic languages are Teso in Uganda and Kenya, Dinka in The Sudan, Luo in Kenya and Tanzania, and Lango in Uganda. Only three of the 30 or so languages of the Central Sudanic subgroup of Chari-Nile are spoken by groups of about 100,000 people: Sara in Central African Republic and Chad, Lugbara in Uganda and Zaire, and Mangbetu in Zaire.

Among the Nilo-Saharan languages that are not classified as Chari-Nile is the Saharan group. Kanuri, its largest member, is spoken by more than 4,000,000 people in Nigeria, Niger, Cameroon, and Chad. In the Maba group, Masalit is spoken by 60,000 people in The Sudan. Songhai, often classified as a language isolate, is spoken by 1,000,000 people in Niger, Mali, Upper Volta, Nigeria, and Benin (Dahomey). Fur, also sometimes considered as an isolate, is spoken by about 400,000 people, mostly in The Sudan.

Niger-Congo languages. Languages in the Niger-Congo (or Niger-Kordofanian) family are spoken all across Africa from Mauritania to Kenya and south into South Africa. There are almost 900 Niger-Congo languages, which have been classified into six genetic subgroups. The Bantu languages (of the Benue-Congo subgroup) far outnumber those of any other family in Africa, both in terms of number of languages and in terms of total number of speakers (160,000,000). Fifteen Bantu languages are each spoken by more than 3,000,000 people; the following each have more than 5,000,000 speakers: Rwanda, Shona, Kongo, Luba-Lulua, Xhosa, and Zulu. Other subgroups in the Niger-Congo family include only a few dozen languages, as those in the Mande subgroup in West Africa, which are spoken from Mauritania to Ghana (including Bambara, Mende, and Vai). The Gur (Voltaic) languages, spoken from Mali and the Ivory Coast to Nigeria, include Mossi, with 3,500,000 speakers, and numerous other languages with significantly fewer speakers. The West Atlantic languages, spoken from Senegal to Nigeria, include Fulani, spoken by 11,500,000 people, Wolof (over 2,000,000 speakers), Temne (over 1,000,000 speakers), and several other languages of less numerical import. Of the languages of the Adamawa-Eastern subgroup, spoken from The Sudan to Cameroon, only Sango, through its use as a lingua franca, may be known by more than 1,000,000 people. The Kwa subgroup of Niger-Congo includes Twi (Akan; 9,000,000 speakers), Yoruba (in Nigeria and extending into Dahomey; 18,000,000 speakers), and Igbo (also known as Ibo; in Nigeria; with 15,000,000 speakers). Some scholars link the Kordofanian languages of North and South Kurdufān provinces in The Sudan with the Niger-Congo languages in a Niger-Kordofanian phylum.

Bantu languages

Khoisan languages. The Khoisan family consists of about four dozen languages spoken in southern Africa and two click languages (Sandawe and Hadza) spoken in Tanzania that are not closely affiliated with any one group in the Khoisan family. Uncertainties in the number of languages and the number of language groups arise from the profusion of labels for various groups and the lack of detailed linguistic comparisons among large numbers of them. Most of the Khoisan languages have been considered to be on the verge of extinction, if not known to be already extinct, but recent estimates of the numbers of peoples grouped on the basis of their culture (Hottentot and Bushman) show many thousands of speakers. The Khoisan language estimated to have the most speakers is Nama, with about 130,000. For more information on the Nilo-Saharan (Chari-Nile), Niger-Congo, and Khoisan languages, see below *African languages.*

LANGUAGES OF THE AMERICAS

Languages indigenous to the Americas were brought from Asia by the forebears of modern American Indians (including Eskimos), who left Asia after the dog was domesticated but before other animals were domesticated. Something is known about the culture of these Indians but nothing about their languages, which are known only after contact with European languages.

European languages in the Americas

Today there are six European languages in the Americas that serve as languages of both education and government administration. (Several Indian languages, however, function in this dual role—Guaraní of Paraguay, Greenlandic of Greenland, and Quechua and Aymaran of Peru.) These official languages and their number of primary political divisions are Spanish (18), Portuguese (1), Dutch (2)—1 in Latin America and 1 in the Caribbean; English (2 in North America and 11 in the Caribbean); French (1 in North America and 3 in the Caribbean); and Danish (1 in Greenland). Before the colonial period in Latin America and during the first century or two of that period, the following American Indian languages could also be classed as official or semi-official: Nahuatl (Nahua), the language of the Aztecs in Mexico and Central America; Chibcha-Muisca in Colombia; Quechua, the language of the Incas, in the Andean area; Tupí in Brazil; and Guaraní in and around Paraguay. In addition to American Indian languages, two pidgin-creole languages are official in their own political divisions, Sranan (Taki-Taki) in Suriname and Papiamento in Curaçao. Other pidgin-creoles in the Caribbean, such as Haitian Creole, are being increasingly written.

Genetic diversity among languages of continental-sized areas can be expressed in terms of the number of minimum genetic classes taken as the usual basis for discussion by specialists of that area. Research may lead to a downward (or upward) revision, and a new number of minimum genetic classes is used as a basis for further discussion. For North America (north of Mexico) and for the 20th century, the basis for discussion has shifted three times so far: from about 50 families in the classification of the U.S. scholar J.W. Powell to six phyla in the classification of the U.S. anthropological linguist Edward Sapir, which was revised at the 1964 Conference on North American Indian Languages by splitting and reclassification (*e.g.,* of Sapir's Hokan-Siouan) and by merging (*e.g.,* the Muskogean family and a few isolates were added to Algonkian in the new Macro-Algonkian phylum). This third classification is summarized below. Proposals for a minimum number of genetic classes in South America range from more than 100 families to three phyla (in a recent liberal classification).

The Plains Indian sign language (hand talking) is still known, but Chinook Jargon and other pidgin-creoles in North America fell into disuse as soon as American Indians became bilingual in English, French, or Spanish.

North and Central American Indian languages. For North America north of Mexico, the summary of culture areas (before any American Indians were relocated by Europeans) by the U.S. anthropologist Harold E. Driver is a convenient basis on which to superimpose the various ways in which language classifications (genetic and typological) combine with cultures that are ecologically adapted to each of ten areas—the Arctic, Subarctic, Northwest Coast, Plateau, Plains, Prairies, East, California, Great Basin, and Southwest. The three variables (genetic, typological, and cultural) coincide approximately in the Arctic (the one language family, Eskimo-Aleut, does not include typologically diverse languages, but it does spread over a culture area that is not entirely homogeneous). In the Subarctic two language families are represented, Algonkian and Athabascan, which are distinct typologically as well as genetically. Northwest Coast and adjacent Plateau languages are genetically very diversified but surprisingly homogeneous in a diffusional kind of phonological typology. The languages in the treeless Plains and the midwestern Prairies are genetically and typologically diverse; all the language families represented, except Caddoan, are intrusive in the sense that their homelands lie outside the Plains and Prairie areas.

Ten culture areas of North American Indians

Language families in the East give an impression of a little typological similarity combined with considerable genetic diversity. On the opposite coast, California is surprisingly homogeneous in culture and in language typology but heterogeneous in genetic classification of languages. There are few languages and only two language families represented in the Great Basin, which is homogeneous in all respects. The adjacent Southwest is anomalous in all three variables considered here. Where it is culturally homogeneous, as between Pueblo societies, it is genetically and typologically diverse in language: four different language families are represented in Pueblo societies. Non-Pueblo societies of the Southwest are diverse culturally as well as linguistically.

Eskimo-Aleut. The three languages of the Eskimo-Aleut family are still spoken in their prediscovery areas from Greenland to Siberia and also on Komandor Island between the Aleutians and Kamchatka (see below *Languages of the Americas: Eskimo-Aleut languages*).

Athabascan. About 20 languages of the Athabascan family are still spoken in four different culture areas: the Yukon and Mackenzie areas of the Subarctic (the centre of Athabascan diversity, with 17 living languages, including Chipewyan-Slave-Yellowknife and Carrier), the Northwest Coast (where only Hupa, Tolowa, and Chasta Costa may still be spoken), the Southwest (where the Navajo dialect of what may be considered a single Southwestern Apachean language has more speakers than any other Indian language north of Mexico), and the Plains (where two Athabascan languages are more recently intrusive—Sarcee [Sarsi] from the Subarctic and Kiowa Apache from the Southwest). Three language isolates spoken in the Northwest Coast (Eyak, Tlingit, and Haida) are remotely related to Athabascan in the Na-Dené phylum, but Eyak is so much more closely related to the Athabascan family that it might be considered a divergent member of the family.

Algonkian. The Algonkian family includes 13 languages still spoken, which belonged in the culture areas of the eastern Subarctic (*e.g.,* Cree, Ojibwa, Micmac, Malecite), the Prairies (*e.g.,* Fox, Potawatomi), the Plains (*e.g.,* Blackfoot, Cheyenne, Arapaho), and the East (where most Algonkian languages became or are now becoming extinct, with only the removed Shawnee and Delaware surviving in Oklahoma). Remotely related to the Algonkian languages in a Macro-Algonkian phylum are languages spoken further to the south in the East—the Muskogean family (including Choctaw-Chickasaw and Creek-Seminole) and several language isolates that are no longer spoken, as well as two almost extinct languages of the Northwest Coast that are more closely related to Algonkian (Wiyot and Yurok).

Macro-Algonkian phylum

Macro-Siouan. The Macro-Siouan phylum is named for its most extensive component, the Siouan family, the extant languages of which belong in the Plains and Prairies, including Dakota, Crow, Winnebago, and Omaha-Osage. (The Siouan languages of the East, such as Ofo and Biloxi, are no longer spoken.) Less widely distributed than Siouan is the Iroquoian family (six languages, largely of the East, including Cherokee and Mohawk), the Caddoan family (Caddo in the East, Wichita and Pawnee in the Prairies), and two language isolates of the East (Catawba and Yuchi), more closely related to Siouan than to the other families in Macro-Siouan.

Hokan. The Hokan phylum includes several small families and a number of language isolates scattered from the Northwest Coast through California, with extensions into the Great Basin and the Southwest, and as far south as Meso-America. Hokan languages spoken by the greatest numbers of speakers include those in two families in Mexico, the Tlapanecan and Tequistlatecan, and in the Yuman family in Arizona and California.

Penutian. The Penutian phylum is the only group of languages in North America for which relationships with languages in South America have been traced convincingly. The Penutian languages are thus distributed from the Northwest Coast and Plateau areas through California (with a possible extension into the Southwest) and Meso-America into Bolivia, Chile, and Argentina. Many of the Penutian languages north of Mexico are either no longer spoken or are spoken by fewer than 50 people. In Meso-America, however, many native languages have a considerable number of speakers; *e.g.,* Mixe, in the Zoque family, has over 77,000 speakers, and the Mayan family includes some languages with several hundred thousand speakers, as Maya, Quiché, Kekchí, Cakchiquel, and Mam.

Aztec-Tanoan. The Aztec-Tanoan phylum consists of two families: the Tanoan (Kiowa-Tanoan) family with three languages in the Southwest, including those spoken by the Taos and the Santa Clara, and one language in the Plains (Kiowa); and the Uto-Aztecan family, with about a score of languages spoken from the Plateau and California into Meso-America, with relatively late extensions into the Plains. California Uto-Aztecan languages include Cahuilla and Luiseño; Great Basin languages include Paiute and Shoshoni, with the Ute and Comanche dialects in the Plains; Southwestern languages include Hopi and Pima-Papago; Meso-American languages include Nahuatl, the language of the descendants of the Aztecs. The million speakers of the several varieties of Nahuatl far outnumber the total number of speakers of all the other Uto-Aztecan languages.

The Tanoan and Uto-Aztecan families

Oto-Manguean. Languages of one North American phylum are located entirely in Meso-America—the Oto-Manguean phylum, consisting of five small families. The languages with the largest number of speakers are Otomí, Mixtec, and Zapotec.

Unaffiliated languages. In North America one large family (the Salish family in the Northwest Coast and Plateau) and several smaller families and language isolates (as the Wakashan family in the Northwest Coast and Tarascan in Meso-America) remain undetermined in phylum affiliation. Remote relationships that have been proposed for some of these are in conflict with other proposed relationships, with no overwhelming evidence presented for any one of the proposals.

South American Indian languages. Language names for South America are much more numerous than those for North America, but information on actual languages is generally sporadic and often lacking entirely. Even when the list of names is reduced to 350 for languages said to be still spoken, the data to which the names refer consist, for the most part, of brief word lists; or nothing more may be known than the fact that a tribe X is said to speak differently from a tribe Y. Though it is possible to know that certain languages are probably closely related, it is not always possible to say how closely; *i.e.,* whether they might be dialects of, or occasionally just different names for, the same language. At the opposite extreme of genetic relationship, it is clear that there are large groups of remotely related languages, but the paucity of data makes possible conflicting proposals. For at least one group of languages, those of the high cultures of South America—the Inca and the Aymara—and some of their neighbours, the problem of establishing genetic relationship is complicated by the problem of sorting out borrowings among them.

Andean-Equatorial phylum

The Andean-Equatorial phylum includes the greatest number of still-spoken languages (almost 200) and the three South American Indian languages with the greatest number of speakers (Quechua, Guaraní, and Aymara). The living Andean-Equatorial languages comprise some 14 families and several language isolates. The Arawakan family includes the biggest numbers of languages—around 100—and has the widest distribution: across northern South America from French Guiana to Colombia and southward as far as Paraguay; formerly, Arawakan languages were also spoken in Central America and the islands of the Caribbean. Most Arawakan languages are spoken by not more than a few hundred people. More than two dozen languages of the Tupian family are still spoken over a large part of South America, principally south of the Amazon River. Tupian languages include Guaraní (Tupí-Guaraní), which is spoken in a number of dialects by about 3,400,000 people in Paraguay, Brazil, Argentina, and Bolivia. Quechua, of the Quechumaran group, is spoken by around 7,000,000 people in Peru, Ecuador, Colombia, Bolivia, Argentina, and Chile. Some Quechua dialects are so divergent that they might be regarded as separate languages. The other Quechumaran language group, Aymaran, is spoken by more than 1,000,000 people in Peru and Bolivia. Most other languages in the Andean-Equatorial phylum are spoken by only a few thousand persons.

The Ge-Pano-Carib phylum includes almost as many languages still spoken as the languages of the Andean-Equatorial phylum, but the former are all spoken by relatively small tribes, so that the total number of speakers of these languages is only a small fraction of the number of speakers of Andean-Equatorial languages. In terms of numbers of languages, the largest family in the Ge-Pano-Carib phylum is the Cariban (Carib) family, with some 60 languages still spoken in Venezuela, the Guianas, Brazil, and Colombia. Cariban languages were also formerly spoken in the Caribbean islands. Most Cariban languages have no more than 1,000 speakers. The other large family in the phylum, the Macro-Ge family, includes more than 25 languages in Brazil.

The languages of the Macro-Chibchan phylum, of which 39 may still be spoken, are distributed from Guatemala and Honduras southward into, and possibly beyond, Peru. The largest component of the phylum is the Chibchan family, of which 16 languages are still spoken from Nicaragua to northwest Colombia—these include Cuna, spoken on the Mulatas (San Blas) Islands of Panama as well as on the mainland of Panama and Colombia; Guaymí, spoken in Panama; and Páez in Colombia. For further information on the Indian languages of the Americas, see below *Languages of the Americas: North American Indian languages; Meso-American Indian languages; South American Indian languages.* (C.F.V./F.M.V.)

INDO-EUROPEAN LANGUAGES

Indo-European is the name of a family of languages that by 1000 BC were spoken over most of Europe and in much of Southwest and South Asia; from the second half of the 15th century the Indo-European tongues have spread to most other inhabited parts of the world. In German the family is called *Indogermanisch,* which has led to the occasional use of "Indogermanic" in English. The term Indo-Hittite is used by scholars who believe that Hittite and the other Anatolian languages (see below) are not just one branch of Indo-European but rather a branch coordinate with all the rest put together; thus, Indo-Hittite is used for a family consisting of Indo-European proper plus Anatolian. As long as this view is neither definitively proved nor disproved, it is convenient to keep the traditional use of the term Indo-European.

Indo-Hittite hypothesis

Languages of the family. The well-attested languages of the Indo-European family fall fairly neatly into the ten main branches listed below; these are arranged according to the age of their oldest sizable texts.

Anatolian. Now extinct, Anatolian was spoken during the 1st and 2nd millennia BC in what is presently Asian Turkey and northern Syria. By far the best known of its members is Hittite, the official language of the Hittite Empire, which flourished in the 2nd millennium. Very few Hittite texts were known before 1906, and their interpretation as Indo-European was not generally accepted until after 1915; the integration of Hittite data into Indo-European comparative grammar has, therefore, been one of the principal developments of Indo-European studies in this century. The oldest Hittite texts date from the 17th century BC, the latest from the 13th. For more information, see below *Anatolian languages.*

Indo-Iranian. Indo-Iranian comprises two main sub-branches, Indo-Aryan (Indic) and Iranian. Indo-Aryan languages have been spoken in what is now northern and central India and Pakistan since before 1000 BC. Aside from a very poorly known dialect spoken in or near northern Iraq during the 2nd millennium BC, the oldest record of an Indo-Aryan language is the Vedic Sanskrit of the Ṛgveda, the oldest of the sacred scriptures of India, dating roughly from the centuries around 1000 BC. Examples of modern Indo-Aryan languages are Hindi, Bengali, Sinhalese (spoken in Sri Lanka), and Romany, the language of the Gypsies.

Iranian languages were spoken in the 1st millennium BC in present-day Iran and Afghanistan, and also in the steppes to the north, from modern Hungary to Chinese Turkistan. The only well-known ancient varieties are Avestan, the sacred language of the Zoroastrians (Parsees), and Old Persian, the official language of Darius I (ruled 522–486 BC) and Xerxes I (486–465 BC) and their successors. Some modern Iranian languages are Persian, Pashto (Afghan), Kurdish, and Ossetic. For more information on the Indo-Iranian languages, including the Kafiri group, which occupies a special position, see below *Indo-Iranian languages.*

Minoan Linear B texts

Greek. Greek, despite its numerous dialects, has been a single language throughout its history. It has been spoken in Greece since at least 1600 BC, and, in all probability, since the end of the 3rd millennium. The earliest texts are the Minoan Linear B tablets, some of which may date from as far back as 1400 BC (the date is disputed), and some of which certainly date from around 1200 BC. This material, very sparse and difficult to interpret, was not identified as Greek until 1952. The Homeric epics—the *Iliad* and the *Odyssey*—composed for the most part in the 8th century BC, are the oldest texts of any bulk. For more information, see below *Greek language.*

Italic. The principal language of the Italic group is Latin, originally the speech of the city of Rome and the ancestor of the modern Romance languages: Italian, Romanian, Spanish, Portuguese, French, etc. The earliest Latin inscriptions date apparently from the 6th century BC, with literature beginning in the 3rd century. Scholars are not in agreement as to how many other ancient languages of Italy and Sicily belong in the same branch as Latin. For more information on Latin, the languages derived from it, and the other languages that belong or may belong to the Italic branch of Indo-European, see below *Italic languages; Romance languages.*

Germanic. In the middle of the 1st millennium BC, Germanic tribes lived in southern Scandinavia and northern Germany. Their expansions and migrations from the 2nd century BC onward are largely recorded in history. The oldest Germanic language of which much is known is the Gothic of the 4th century AD. Other languages include English, German, Dutch, Danish, Swedish, Norwegian, and Icelandic. For more information on the Germanic languages, see below *Germanic languages; English language.*

Armenian. Armenian, like Greek, is a single language. The Armenians are recorded as being in what is now eastern Turkey and the Armenian Soviet Socialist Republic as

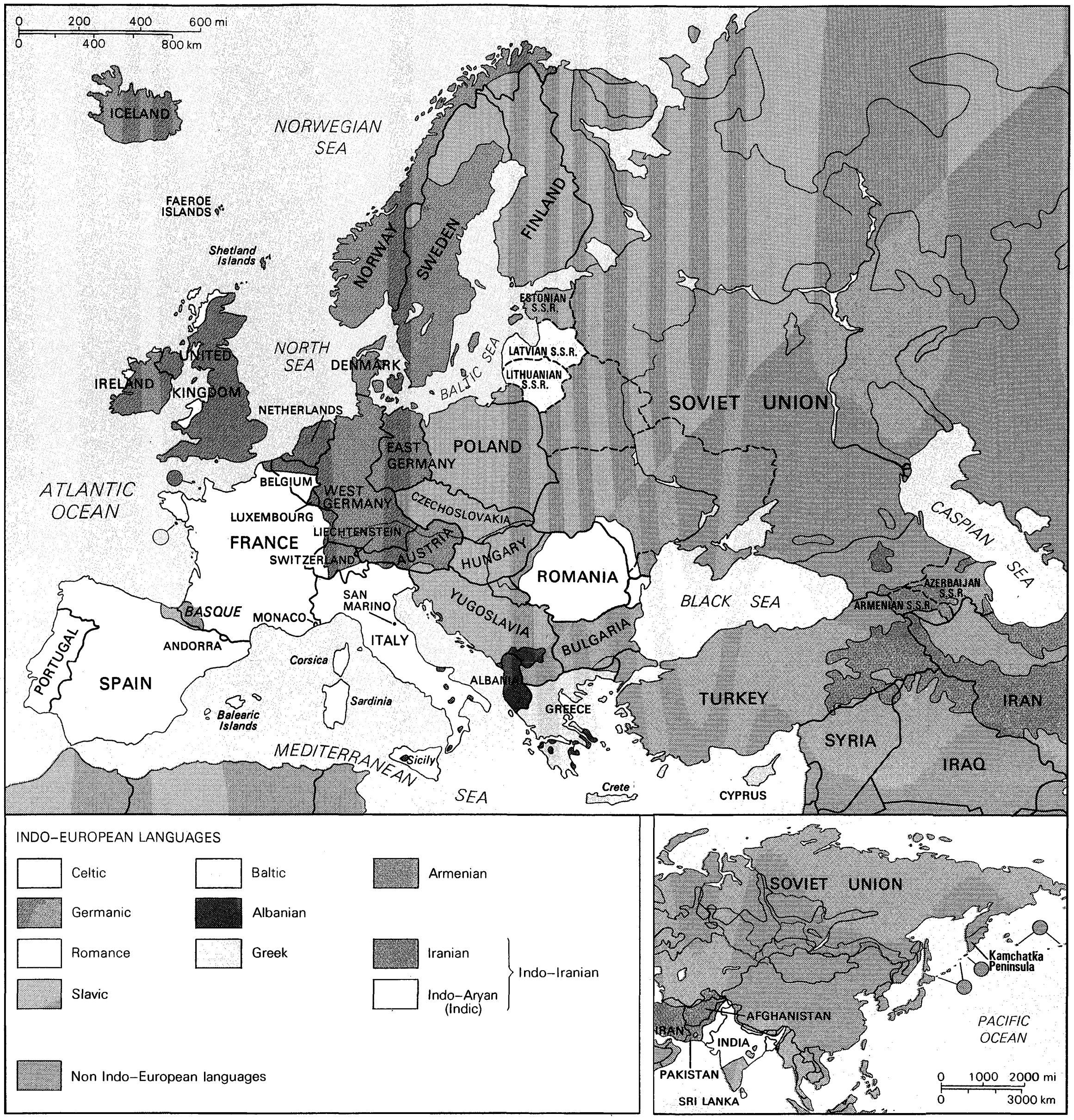


Figure 1: Approximate locations of Indo-European languages in Eurasia in the 20th century.

Adapted from A. Meillet and M. Cohen, *Les Langues du monde* (1952); Editions du Centre National de la Recherche Scientifique, Paris

early as the 6th century BC, but the oldest Armenian texts date from the 5th century AD. For more information, see below *Armenian language.*

Tocharian. Tocharian, now extinct, was spoken in present-day Chinese Turkistan in the 1st millennium AD. Two distinct languages are known, labelled A (Turfanian) and B (Kuchean); many scholars consider Tocharian A and B to be two dialects of the same language. One group of travel permits for caravans can be dated to the early 7th century, and it appears that other texts date from the same or from neighbouring centuries. These languages became known to scholars only in the first decade of the 20th century; they have been less important for Indo-European studies than has Hittite, partly because their testimony about the Indo-European parent language is obscured by 2,000 more years of change, and partly because Tocharian testimony fits fairly well with that of the previously known non-Anatolian languages. For more information, see below *Tocharian language.*

Tocharian A and B

Celtic. Celtic was spoken in the last centuries before the Christian Era over a wide area of Europe, from Spain and Britain to the Balkans, with a group (the Galatians) even in Asia Minor. Very little of the Celtic of that time and the ensuing centuries has survived, and this branch is known almost entirely from the Insular Celtic languages—Irish, Welsh, and others—spoken in and near the British Isles, as recorded from the 8th century AD onward. For more information, see below *Celtic languages.*

Balto-Slavic. The grouping of Baltic and Slavic into a single branch is somewhat controversial, but the exclusively shared features outweigh the old divergences. At the beginning of the Christian Era, Baltic and Slavic tribes occupied a large area of eastern Europe, east of the Germanic tribes and north of the Iranians, including much of present-day Poland and the western Soviet Union. The Slavic part of this area was probably fairly small, perhaps centred in what is now southern Poland. But in the 5th century AD the Slavs began expanding in all directions, until now the Slavic languages are spoken over the greater part of eastern Europe and northern Asia. The Baltic-speaking area, however, has contracted, so that Baltic languages are presently confined to the two Soviet Socialist Republics of Lithuania and Latvia.

The earliest Slavic texts, written in a dialect called Old Church Slavonic, date from the 9th century AD; the oldest substantial material in Baltic comes from the end of the 14th century, and the oldest connected texts from the 16th century. For more information, see below *Baltic languages; Slavic languages.*

Albanian. Albanian, the language of the present-day

Table 1: Widely Shared Indo-European Terms*

	Hittite	Sanskrit	Greek	Latin	English	Armenian	Tocharian B	Old Irish	Lithuanian	Albanian
I	uk	ahám	egō	ego	I	es			àš	
me	ammuk	mā́m	eme	mē	me	is		-m	manè	mua
thou		tuvám	su	tū	thou	du	twe	tú	tù	ti
thee	tuk	tvā́m	se	tē	thee	k'ez	ci	-t	tavè	ty
who?	kuis	kás	tis	quis	who?	ov	k$_{u}$se	cía	kàs	kush
what?	kuit	kím	ti	quid	what?	z-i	k$_{u}$se	cid	kàs	çë
that		tát	to		that	da	te		taĩ	
water	watar	udakám	hudōr		water		war	uisce	vanduõ	ujë
fire	paḫḫur		pūr		fire	hur	puwar			
father		pitár-	pater-	pater	father	hayr	pācer	athair		
mother		mātár-	māter-	māter	mother	mayr	mācer	máthair	mótina	
brother		bhrā́tar-		frāter	brother	elbayr	procer	bráthair	brólis	
sister		svásār		soror	sister	k'oyr	ṣer	siur	seser-	
daughter		duhitár-	thugater-		daughter	dustr	tkācer		dukter-	
son		sūnús	huios		son		soy		sūnùs	
sheep	Luw. ḫawi-	ávis	o(w)is	ovis	ewe			oí	avìs	
cow		gā́v-	bous	bōs	cow	kov	ke$_{u}$	bó	Latv. gùovs	
horse	Hier. Luw. aśuwa-	áśvas	hippos	equus	OE eoh		yakwe	ech	ašvà 'mare'	
pig		sūkarás	hūs	sūs	sow		suwo			thi
dog	Hier. Luw. śuwana-	śvā́n-	kuōn	canis	hound	šun	kwen-	con-	šun-	
wheel		cakrám	kuklos		wheel		kokale 'wagon'			
heart	kart-		kardiā	cord-	heart	sirt		cride	širdìs	
knee	kenu	jā́nu	gonu	genū	knee	cunr	keni	glún		gju
tree,wood	taru	dā́ru	doru		tree			daur 'oak'	OCS drěvo	dru
foot	pat(a)-	pā́d-	pod-	ped-	foot	otn	paiyye			
long	talukis	dīrghás	dolikhos						ìlgas	
new	newas	návas	ne(w)os	novus	new	nor	ñuwe	nue	naũjas	
goes	pa-itsi	éti	eisi	it			yan		eĩna	
is	estsi	ásti	esti	est	is	ē	ste	is	ẽsti	është
eats	etstsi	átti	edei	ēst	eats	utē			ĕ́da	
carries		bhárati	pherei	fert	bears	berē	parän	berid		bie 'brings'
knows		véda	(w)oide		wot	gitē		ro-fitir	OCS věstŭ	
1		ékas	oi(w)os 'alone'	ūnus	one			oín	víenas	një
2	twi-	duvā́	duo	duo	two	erku	wi	dó	dù	dy
3	tri-	tráyas	treis	trēs	three	erek'	trey	trí	trỹs	tre
4		catvā́ras	tettares	quattuor	four	č'ork'	śtwer	cethair	keturì	katër
5		páñca	pente	quīnque	five	hing	piś	cóic	penkì	pesë
6		ṣáṭ	hex	sex	six	vec'	ṣkas	sé	šešì	gjashtë
7	siptam-	saptá	hepta	septem	seven	ewt'n	ṣukt	secht	septynì	shtatë
8		aṣṭā́	oktō	octō	eight	ut'	okt	ocht	aštuonì	tetë
9		náva	enne(w)a	novem	nine	inn	ñu	noí	devynì	nëndë
10		dáśa	deka	decem	ten	tasn	śak	deich	dẽšimt	dhjetë
100		śatám	hekaton	centum	hundred		kante	cét	šim̃tas	
not	natta	ná		ne-	not			ní-	ne-	

*Words lacking in the language named at the top of the column but found in a closely related language are included, with these abbreviations: Luw. = Luwian; Hier. Luw. = Hieroglyphic Luwian; OE = Old English; Latv. = Latvian; OCS = Old Church Slavonic.

republic of Albania, is known from the 15th century AD. It presumably continues one of the very poorly attested ancient Indo-European languages of the Balkan peninsula, but which one is not clear. For more information, see below *Albanian language.*

In addition to the tongues just listed, there are several poorly documented extinct languages of which enough is known to be sure that they were Indo-European and that they did not belong in any of the branches enumerated above (*e.g.,* Phrygian, Macedonian). Of a few, too little is known to be sure whether they were Indo-European or not (*e.g.,* Ligurian).

Reasons for grouping the languages together

Establishment of the family. *Shared characteristics.* The chief reason for grouping the Indo-European languages together is that they share a number of items of basic vocabulary, including grammatical affixes, whose shapes in the different languages can be related to one another by statable phonetic rules. Especially important are the shared patterns of alternation of sounds. Thus the agreement of Sanskrit *ás-ti,* Latin *es-t,* and Gothic *is-t,* all meaning "is," is greatly strengthened by the identical reduction of the root to *s-* in the plural in all three languages: Sanskrit *s-ánti,* Latin *s-unt,* Gothic *s-ind* "they are." Agreements in pure structure, totally divorced from phonetic substance, are, at best, of dubious value in proving membership in the Indo-European family.

Table 1 gives examples of typical vocabulary items widely shared within the Indo-European family that have been decisive in establishing the family. A blank indicates that the language in question does not use the item in accordance with the given meaning or that its word for that meaning is unknown.

Similarities in grammatical endings are shown in Table 2 by samples of noun declension and verb inflection in some of the more archaic languages that have retained the inflectional endings of Indo-European in relatively unchanged form. Note that Old Lithuanian *-i* and *-u* were nasalized vowels, representing a continuation from the earlier forms **-in* and **-un.* (The asterisk marks a form that is not actually found in any document or living dialect but is reconstructed as having once existed in the prehistory of the language.)

The statable phonetic rules referred to earlier are not always obvious without careful observation. Note that the English dental consonants *t, d,* and *th* do not correspond in a straightforward manner to the Greek dental sounds *t, d,* and *th;* that is, English *t* does not occur where Greek *t* appears, nor English *d* where Greek has *d.* But the relationships between the sounds are not random either—English *t* does not correspond to Greek *t* in one word, to *d* in a second, and to *th* in a third, according to no discernible pattern. Rather, where Greek has initial *t,* English has *th,* as in "that" and "three"; where Greek has *d,* English has *t,* as in "tree," "two," and "ten"; and where Greek has *th,* English has *d,* as in "daughter." Note also that phonetic similarity as such is not needed to establish relationship. Thus, many of the Armenian words in Table 1 look quite different from the related words in other Indo-European languages. but here too regular rules of correspondence can be found; *e.g.,* Greek initial *p* corresponds to Arme-

Table 2: Examples of Noun and Verb Inflection

	Hittite paant- (gone)	Sanskrit yant- (going)	Greek iont- (going)	Latin eunt- (going)	Old Lithuanian seser- (sister)
Singular nominative	paant-s	yán	iōn	iēn-s	sesuõ
Singular accusative	paant-an	yánt-am	iont-a	eunt-em	sẽser-į
Singular genitive	paant-as	yat-ás	iont-os	eunt-is	seser-ès
Singular dative	paant-i	yat-é		eunt-ī	sẽser-i
Singular locative	paant-i	yat-í	iont-i	eunt-e	seser-yjè
Plural nominative	paant-es	yánt-as	iont-es	eunt-ēs	sẽser-es
Plural accusative	paant-us	yat-ás	iont-as	eunt-es	sẽser-is
Plural genitive	paant-an	yat-ā́m	iont-ōn	eunt-(i)um	seser-ũ
I go	pai-mi	é-mi	ei-mi	e-ō	ei-mì
You (sg.) go	pai-si	é-ṣi	ei	ī-s	ei-sì
He, she goes	pai-tsi	é-ti	ei-si	i-t	eĩ-ti
We go	pai-wani	i-más	i-men	ī-mus	ei-mè
You (pl.) go	pai-tteni	i-thá	i-te	ī-tis	ei-tè
They go	pa-antsi	y-ánti	i-āsi	e-unt	

nian *h* or zero (a lack of consonant) in the words meaning "fire," "father," "foot," "five."

Linguistic studies of the family. The ancient Greeks and Romans readily perceived that their languages were related to each other, and, as other European languages became objects of scholarly attention in the late Middle Ages and the Renaissance, many of these were seen to be more similar to Latin and Greek than, for example, to Hebrew or Hungarian. But an accurate idea of the true bounds of the Indo-European family became possible only when, in the 16th century, Europeans began to learn Sanskrit. The massive similarities between Sanskrit and Latin and Greek were noted early, but the first person to make the correct inference and state it conspicuously was the English Orientalist and jurist Sir William Jones, who in 1786 said in his presidential address to the Asiatic Society that Sanskrit bore to both Greek and Latin

> a stronger affinity, both in the roots of verbs, and in the forms of grammar, than could possibly have been produced by accident; so strong, indeed, that no philologer could examine them all three without believing them to have sprung from some common source, which, perhaps, no longer exists. There is a similar reason, though not quite so forcible, for supposing that both the *Gothick* [*i.e.,* Germanic] and the *Celtick,* though blended with a very different idiom, had the same origin with the *Sanscrit;* and the old *Persian* might be added to the same family . . . .

The work of Bopp, Rask, and Grimm

The detailed evidence on which Jones based his conclusion was not presented until the 19th century. In 1816 Franz Bopp, the German philologist, presented his *Über das Conjugationssystem der Sanskritsprache in Vergleichung mit jenem der griechischen, lateinischen, persischen und germanischen Sprache* ("On the system of conjugation of the Sanskrit language, in comparison with those of Greek, Latin, Persian, and Germanic"), in which the relation of these five languages was demonstrated on the basis of a detailed comparison of verb morphology (structure). Two years later there appeared the "Undersøgelse om det gamle Nordiske eller Islandske Sprogs Oprindelse" ("Investigation on the Origin of the Old Norse or Icelandic Language"), by the Danish philologist Rasmus Rask, originally written in 1814. This work demonstrated methodically the relation of Germanic to Latin, Greek, Slavic, and Baltic. In 1822 the second edition of the first volume of Jacob Grimm's *Deutsche Grammatik* ("Germanic Grammar") was published; in this grammar were discussed the peculiar Indo-European vowel alternations called *Ablaut* by Grimm (*e.g.,* English "sing, sang, sung"; or Greek *peíth-ō* "I persuade," *pé-poith-a* "I am persuaded," *é-pith-on* "I persuaded"). In addition, Grimm tried to find the principle behind the correspondences of Germanic stop and spirant consonants (the first made with complete stoppage of the breath, and the second made with constriction of the breath but not complete stoppage) to the consonants of other Indo-European languages. The sound changes implied by these correspondences have become known as "Grimm's Law." Examples of it include the stop consonant *p* in Latin *pater* corresponding to the spirant consonant *f* in "father," and the correspondences between English and Greek *t, d,* and *th* discussed above.

Bopp demonstrated in 1838 that the Celtic languages were Indo-European, as had been asserted by Jones. In 1850 the German philologist August Schleicher did the same for Albanian, and in 1877 another German philologist, Heinrich Hübschmann, showed that Armenian was an independent branch of Indo-European, rather than a member of the Iranian subbranch. Since then, the Indo-European family has been enlarged by the discovery of Tocharian and of Hittite and other Anatolian languages, and by the recognition, with the aid of Hittite, that Lycian, known and partly deciphered already in the 19th century, belongs to the Anatolian branch of Indo-European.

The Indo-European character of Tocharian was announced by the German scholars Emil Sieg and Wilhelm Siegling in 1908. The Norwegian orientalist Jørgen Alexander Knudtzon recognized Hittite as Indo-European on the basis of two letters found in Egypt (translated in *Die zwei Arzawa-briefe,* 1902; "The Two Arzawa Letters"), but his views were not generally accepted until 1915, when Bedřich Hrozný published the first report of his own decipherment of the much more copious material that had meanwhile been found in the ruins of the Hittite capital itself.

First full comparative Indo-European grammar

The first full comparative grammar of the major Indo-European languages was Bopp's *Vergleichende Grammatik des Sanskrit, Zend, Griechischen, Lateinischen, Litthauischen, Altslawischen, Gotischen und Deutschen* (1833–52; "Comparative Grammar of Sanskrit, Zend, Greek, Latin, Lithuanian, Old Slavic, Gothic, and German"). But this and August Schleicher's shorter *Compendium der vergleichenden Grammatik der indogermanischen Sprachen* (1861–62; "Compendium of the Comparative Grammar of the Indo-European Languages") were rendered obsolete by the major breakthrough of the 1870s, when scholars realized that sound correspondences are not merely rules of thumb that do not have to be strictly observed, and that apparent exceptions to sound laws can often be accounted for by stating them more accurately or by reconstructing additional different sounds in the parent language. The difference between Gothic *d* in *fadar* "father" and *þ* in *broþar* "brother," for example, both corresponding to *t* in Sanskrit, Greek, and Latin, proved to be correlated with the original position of the accent, a discovery known as Verner's Law (named for the Danish linguist Karl Verner). Thus, *d* appears when the preceding syllable was originally unaccented (*fadar* : Greek *patér-,* Sanskrit *pitár-*), and *þ* occurs when the preceding syllable was originally accented (*broþar* : Greek *phrā́ter* "member of a clan," Sanskrit *bhrā́tar-*).

The knowledge and opinions that had accumulated by the end of the 19th century are largely incorporated in the German linguist Karl Brugmann's *Grundriss der vergleichenden Grammatik der indogermanischen Sprachen* (2nd ed., 1897–1916; "Outline of Comparative Indo-European Grammar", which remains the latest fullscale treatment of the family.

The parent language. By comparing the recorded Indo-European languages, especially the most ancient ones, much of the parent language from which they are descended can be reconstructed. This reconstructed parent language is sometimes called simply "Indo-European," but in this article the term Proto-Indo-European is preferred.

Proto-Indo-European

Phonology. In Proto-Indo-European there were at least 11 stop consonants. In the following grid these sounds are arranged according to the place in the mouth where the stopage was made and the activity of the vocal cords during and immediately after the stoppage:

	labial	dental	palatal	labiovelar
Voiceless	p	t	ḱ	k^w
Voiced		d	g	g^w
Voiced aspirated (?)	bh	dh	ǵh	g^wh

Labial denotes a sound made with the lips; dental, with the tip of the tongue against the back of the teeth. The palatals were probably made by contact between the upper surface of the tongue and the hard palate (the roof of the mouth), like Hungarian *ty* and *gy* in *atya* and *Magyar.* The labiovelars were probably made by contact between the upper surface of the tongue and the soft palate (the area behind the hard palate), with a concomitant rounding of the lips. Voiceless designates sounds made without vibration of the vocal cords; voiced sounds are pronounced with vibration of the vocal cords. The exact pronunciation of the "voiced aspirates" is uncertain.

There may also have been a voiced labial stop, _b_, but correspondences pointing to this are few, and rarely extend beyond immediately neighbouring languages. Correspondences that some scholars take as evidence for a set of plain velar consonants (made with the back of the tongue touching the soft palate), _k, g, gh,_ are partly, perhaps entirely, the result of special developments of labiovelars and palatals in specific positions. The evidence for a set of voiceless aspirated stops _ph, th, ḱh, kh,_ k^wh is extremely weak. (Aspirated consonants are sounds accompanied by a puff of breath.)

There was one sibilant consonant, _s_, with a voiced alternant, _z_, that occurred automatically next to voiced stops. The existence of a second apical spirant, _þ_ (presumed pronunciation like that of _th_ in English "thin"), is extremely uncertain.

Laryngeal theory

Most scholars now agree that the parent language had one or more additional stop or spirant consonants, for which the label laryngeal is used. These consonants, however, have mostly disappeared or have become identical with other sounds in the recorded Indo-European languages, so that their former existence had to be deduced mainly from their effects on neighbouring sounds. Hence, the laryngeal sounds were not suspected until 1878, and even then they were rejected by most scholars until after 1927, when Kuryłowicz showed that Hittite often has _ḫ_ (perhaps a velar spirant like the _ch_ in German _ach_) in places where a "laryngeal" had been posited on the evidence of the other Indo-European languages. There is still considerable disagreement about how many "laryngeals" there were, what they sounded like, what traces they left, and how best to symbolize them. Probably there were three or four, which can be written H_1, H_2, H_3 (and H_4), and probably some or all of them were palatal or (labio-)velar spirants. The principal traces they left outside Anatolian are in the quality and length of neighbouring vowels, H_2 (and H_4) changing a neighbouring _e_ to _a_, and H_3 changing it to _o_, while all laryngeals lengthened a preceding vowel. In Anatolian, H_2 and H_3 remained as _ḫ_, at least in some positions; H_4 is tentatively set up to account for words with _a_ that lack _ḫ_ in Hittite.

When laryngeals between consonants disappeared, a vowel sometimes remained, as in Greek _stasis_, Sanskrit _sthitis_, Old English _stede_ "a standing (place)" from Proto-Indo-European *_stH_2tis_. Scholars who do not posit "laryngeals" reconstruct a separate Proto-Indo-European vowel _ə_ (called _schwa indogermanicum_) to account for these correspondences.

Finally, there were the nasal sounds _n_ and _m_, the liquids _l_ and _r_, and the semivowels _y_ and _w_. When _y_ and _w_ occurred between consonants, they were replaced by the vowels _i_ and _u_. The nasals and liquids functioning as nuclei of syllables in this position (like the final sounds of English "bottom," "button," "bottle," "butter") are traditionally written _n̥, m̥, l̥, r̥_. Some scholars dispense with these diacritical marks and with the distinction between syllabic _i_ and _u_ and nonsyllabic _y_ and _w_, but this obscures certain distinctions, such as that between _-wn̥-_ in *_ḱwn̥su_ "among dogs," Sanskrit _śvasu_, and _-un-_ in *_tund-_ "shove," Sanskrit _tundate_.

Ablaut and vowel grades

The vowel system of Proto-Indo-European was dominated by a pattern of alternation called ablaut. The alternant (called a grade) that occurs in a given syllable of a given form is only partly predictable from the shape of the rest of the word. The basic vowel of the system was _e_ ("normal grade"), and the changes it could undergo were loss (zero-grade), change to _o_ (_o_-grade), lengthening to _ē_ (lengthened grade), and lengthening plus change to _ō_ (lengthened _o_-grade). The stem _ped-_ "foot," for example, appears as such in Latin _ped-is_ (normal grade) "of a foot," as _-bd-_ in Avestan _fra-bd-a-_ (zero-grade) "fore-foot," as _pod-_ in Greek _pod-es_ (_o_-grade) "feet," as *_pēd-_ in Latin _pēs_ (lengthened grade) "foot" in the nominative singular, and as *_pōd-_ in English "foot" (lengthened _o_-grade).

Ablauting forms whose basic vowel is _a, o, ē, ā,_ or _ō_ in the recorded languages (_e.g.,_ Greek _ag-_ "lead," _op-_"see," _stā-_ "stand") are now believed to have had _e_ preceded or followed by laryngeal in the parent language; _e.g.,_ *H_2_eǵ-_ "lead," *H_3_ek_^{w}- "see," *_ste_H_2- "stand." It is uncertain whether there were additional _o_ and _a_ vowels besides those arising by ablaut and from _e_ next to a laryngeal.

The vowels _i_ and _u_ did not participate in ablaut alternations, but rather functioned primarily as the syllabic realizations of the consonants _y_ and _w_, as in *_leyk_^{w}- "leave," zero-grade *_lik_^{w}-, like *_derḱ-_ "see," zero-grade *_dr̥ḱ-_. Long _ī_ and _ū_ in the recorded languages derive, at least in part, from sequences of _i_ or _u_ plus laryngeal; _e.g.,_ Latin _vīvus_ "alive" from *$g^w iH_3$_wós_.

Thus the parent language had at least the following vowels:

		front	back
high	....	_i_	_u_
mid	....	_e, ē_	_o, ō_

(In forming front vowels, the highest point of the tongue is in the front of the mouth; for back vowels, that point is in the back. High vowels are those in which the tongue is highest—closest to the roof of the mouth; mid vowels are made with the tongue between the extremes of high and low.) Of these vowels, _i_ and _u_ really functioned as consonants, and _ē, o, ō_ were all conditioned alternants of _e_. But as noted above there may also have been _ī, ū, a,_ and a second _o_.

The accent just before the breakup of the parent language was apparently mainly one of pitch rather than stress. Each full word had one accented syllable, presumably pronounced on a higher pitch than the others.

The Indo-European verb

Morphology and syntax. The Proto-Indo-European verb had three aspects: imperfective, perfective, and stative. Aspect refers to the nature of an action as described by the speaker; _e.g.,_ an event occurring once, an event recurring repeatedly, a continuing process, or a state. The difference between English simple and "progressive" verb forms is largely one of aspect; _e.g.,_ "John wrote a letter yesterday" (implying that he finished it) versus "John was writing a letter yesterday" (describing an ongoing process, with no implication as to whether it was finished or not). The Anatolian languages lack a dimension of aspect, and it is not yet clear what the earlier system underlying both Anatolian and the rest of Indo-European was.

The imperfective aspect, traditionally called present, was used for repeated actions and for ongoing processes or states; _e.g.,_ *_sti-ste_H_2- "stand up more than once, be in the process of standing up," *_weǵh-e-_ "be in the process of conveying," *_es-_ "be." The perfective aspect, traditionally called aorist, expressed a single, completed occurrence of an action or process; _e.g.,_ *_cte_H_2- "stand up, come to a stop," *_wēǵh-s-_ "convey." The stative aspect, traditionally called perfect, described states of the subject; _e.g.,_ *_woyd-_ "know," *_ste-sto_H_2- "be in a standing position."

Verb roots were by themselves either perfective (like *_ste_-H_2- "stand") or imperfective (like *_weǵh-_ "convey," *_es-_ "be"). This basic aspect, however, could be reversed by aspect markers; _e.g.,_ reduplication for imperfective, as in *_sti-steH-_ (reduplication is the repetition of a word or part of a word), and _-s-_ for perfective, as in *_wēǵh-s-_. The stative aspect was always marked by the _o_-grade of the root in the indicative singular (as in *_woyd-_ "know"), and usually also by reduplication (as in *_ste-sto_H_2-); it had personal endings different from those of the other two aspects.

From one aspect of a given verb the shape and even the existence of the other two aspects could not be predicted; for example, *_es-_ "be" had only the imperfective aspect. Ways of forming imperfectives were especially numerous and often involved, in addition to their imperfective aspectual meaning, some other notion, such as performing the action habitually or repeatedly (iterative), or causing someone else to perform it (causative). One root could thus have several imperfective stems; so to the root *_er-_ "move" there were at least a causative form, *_r̥-new-_ "set in motion," and an iterative form, *_r̥-sk̯e-_ "go repeatedly."

The Proto-Indo-European verb was also inflected for mood, by which the speaker could indicate whether he was making statements or inquiries about matters of fact;

making predictions, surmises, or wishes about the future or about unreal but imagined situations; or giving commands. Compare English "If John is home now (he is eating lunch)" with the verb "is" in the indicative mood, discussing a matter of fact, with "If John were home now (he would be eating lunch)" with the verb "were" in the subjunctive mood, describing an unreal situation. There were two Proto-Indo-European suffixes expressing mood: _-e-_ alternating with _-o-_ for the subjunctive, corresponding roughly in meaning to the English auxiliaries "shall" and "will," and _-ye_H_1- alternating with _-i_H_1-for the optative, corresponding roughly to English "should" and "would." Verbs without one of these two suffixes were marked for mood and tense by their personal endings.

Mood in Indo-European

These personal endings basically expressed the person and number of the verb's subject, as in Latin _amō_ "I love," _amās_ "you (singular) love," _amat_ "he or she loves," _amāmus_ "we love," and so on. In the imperfective and perfective aspects there were two sets of endings, distinguishing two voices: active, in which typically the subject was not affected by the action, and mediopassive, in which typically the subject was affected, directly or indirectly. Thus Sanskrit active _yajati_ and mediopassive _yajate_ both mean "he sacrifices," but the former is said of a priest who performs a sacrifice for the benefit of another, while the latter is said of a layman who hires a priest to perform a sacrifice for him. In the stative aspect there was no distinction of voice. (Voice indicates the relationship of the action expressed by the verb to the subject of the statement.)

To mark mood and tense, verbs in the imperfective aspect that did not have a mood suffix had three sets of personal endings in both active and mediopassive voices: imperative, primary, and secondary. Verbs with imperative endings belonged to the imperative mood (used for commands); _e.g._, _*s-dhí_ "be," _*és-tu_ "let him be." Verbs with primary endings were marked as non-past in tense and indicative in mood; _e.g._, _*és-ti_ "he is." (Indicative mood signifies objective statements and questions.) Verbs with secondary endings were unmarked for tense and mood, but were most typically used as past indicatives (_e.g._, *g^w_hén-t_ "he slew") and to fill out gaps in the imperative paradigm (_e.g._, _*s-té_ "be" in the plural, _*gəhn̥-té_ "ye slew; slay" in the plural). To mark such forms unambiguously as past indicatives, an augment, usually consisting of the vowel _e_, could be prefixed; _e.g._, _*é-_g^w_hen-t_ "he slew," _*ēst_ (= _*é-es-t_) "he was."

Verbs in the perfective aspect without a mood suffix did not occur with primary endings, and so lacked a non-past indicative tense. Verbs in the stative aspect apparently lacked a distinction between primary and secondary endings, so that a form like _*wóyd-e_ "he knows" meant also "he knew."

Noun cases, number, gender

The inflectional categories of the noun were case, number, and gender. Eight cases can be reconstructed: nominative, for the subject of a verb; accusative, for the direct object; genitive, for the relations expressed by English "of"; dative, corresponding to the English preposition "to," as in "give a prize to the winner"; locative, corresponding to "at," "in"; ablative, "from"; instrumental, "with"; and vocative, used for the person being addressed. For examples of some of these see Table 2. Besides singular and plural number, there was a dual number for referring to two items. Each noun belonged to one of three genders: masculine, to which belonged most nouns designating male creatures; feminine, to which belonged most names of female creatures; and neuter, to which belonged only a few words for individual adult living creatures. The gender of nouns not designating living creatures was only partly predictable from their meaning.

Adjectives were nouns that varied in gender according to the gender of another noun with which they were in agreement, or, if used by themselves, according to the sex of the entity to which they referred; thus, Latin _bonus sermō_ "good speech" (masculine), _bona aetās_ "good age" (feminine), _bonum cor_ "good heart" (neuter), or _bonus_ "a good man," _bona_ "a good woman," _bonum_ "a good thing." The neuter of an adjective was identical with the masculine except for having different endings in nominative and accusative cases. Feminine gender was either completely identical with the masculine or derived from it by means of a suffix, the two commonest being *-_e_H_2- and *-_i_H_2- (*-_ye_H_2-).

Demonstrative, interrogative, relative, and indefinite pronouns were inflected like adjectives, with some special endings. Personal pronouns were inflected very differently. They lacked the category of gender, and marked number and case (in part) not by endings but by different stems, as is still seen in English singular nominative "I"; oblique "my," "me"; plural nominative "we"; plural oblique "our," "us." (The oblique is any case other than nominative or vocative.)

Syntactic features

Some notable features of Proto-Indo-European syntax are: the non-ergative case system, that is, the subject of an intransitive verb is in the same case as the subject (rather than the object) of a transitive verb; concord (agreement) in case, number, and gender between adjective and noun; and use of singular verbs with neuter plural subjects, as in Greek _panta rhei_ "all things flow," with the same verb as _ho potamos rhei_ "the river (masculine) flows," contrasting with _hoi potamoi rheousi_ "the rivers flow" (indicating that neuter plurals were originally collectives and grammatically singular).

Lexicon and culture. Much less is known about the parent language's vocabulary than about its phonology and grammar. Sounds and grammatical categories do not easily disappear or undergo radical change in so many daughter languages that their former existence can no longer be detected. It is relatively easy, however, for an individual word to disappear or shift meaning in so many daughter languages that its existence or meaning in the parent language cannot be confidently inferred. Hence, from the linguistic evidence alone, scholars can never say that Proto-Indo-European lacked a word for any particular concept; they can only state the probability that certain items did exist, and from these items make inferences about the culture and location in time and space of the speakers of Proto-Indo-European.

Thus is it supposed that the Proto-Indo-European community knew and talked about dogs (_*ḱwón-_), horses (_*éḱwo-_), sheep (*H_3_éwi-_), and almost certainly cows (*g^w_ów-_) and pigs (_*suH-_). Probably all these animals were domesticated. At least one cereal grain was known (_*yewo-_), and at least one metal (*H_2_eyos_ or *H_4eyos). There were vehicles (_*woǵho-_) with wheels (*k^w_e_k^w_lo-_), pulled by teams joined by yokes (_*yugo-_). Honey was known, and probably formed the basis of an alcoholic drink (_*melit-_, _*medhu_) related to the English "mead." Numerals up through 100 (_*ḱm̥tóm_) were in use. All this suggests a people with a well-developed Neolithic (characterized by simple agriculture and polished stone tools) or even Chalcolithic (copper-or bronze-using) technology.

Location and date. Linguists have not found a reliable and precise way to determine from linguistic evidence alone the date at which any set of related languages must have begun diverging. The best that can be done is to estimate the degree of difference between the languages in question, taking into account all that is known about them, and then compare this estimate with the estimated degrees of difference within families of languages—such as the Romance family—whose actual time of divergence is approximately known. Using this sort of "dead reckoning," it can be said that the earliest attested Indo-European languages—Anatolian, Indo-Iranian, and Greek—are different enough that the parent language must have been split into several distinct languages well before 2000 BC, but similar enough that the first split into separate languages is not likely to have been much earlier than 3000 BC, and may have been later.

Division of Indo-European into languages between 3000 and 2000 BC

For further progress the linguistic findings must be correlated with those of archaeologists and paleontologists to see if there was a population group within Eurasia that was relatively small and homogeneous before 3000 BC and that underwent considerable expansion and fragmentation beginning about 3000 BC—give or take a few centuries—such that some of its fragments can be ancestral to components of the cultures of the speakers of the various recorded Indo-European languages. The culture of this population group in the centuries around 3000 BC must also correspond to what can be inferred for Proto-Indo-European from the linguistic data.

At present the archaeological evidence seems to find such a group in the Kurgan culture of the south Russian steppe, east of the Dnepr (Dnieper) River, north of the Caucasus, and west of the Urals. According to the Lithuanian-American archaeologist Marija Gimbutas, in *Indo-European and Indo-Europeans* (1970), this culture began spreading west *c.* 4000–3500 BC (Kurgan II), and began to occupy a really wide area stretching from eastern central Europe to northern Iran *c.* 3500–3000 BC (Kurgan III). Allowing a few centuries for the speech of widely separated bands to diverge to the point of becoming distinct languages, this agrees tolerably well with the date suggested by the linguistic evidence for breakup of the parent language. So far the Kurgan culture has been traced back to the 5th millennium BC; its earlier antecedents are still unknown.

Possible relationship to Uralic

Remote relationship of Indo-European to the Uralic languages is very likely. Geographically, the earliest reconstructible locations of the two families are contiguous; lexically, there are strong resemblances in a number of basic words or word parts, including personal, demonstrative, interrogative, and relative pronouns, personal endings of verbs, the accusative case ending *-m,* and such words as those for "water" and "name"; typologically, the families are fairly similar (*e.g.,* both have many suffixes, but few or no prefixes or infixes—elements inserted within words). The resemblances, however, are too few to permit the reconstruction of a common "Indo-Uralic" parent language; the two families must have separated several thousand years before the breakup of Indo-European.

If Indo-European is related to other language families—*e.g.,* to Hamito-Semitic (Afro-Asiatic) or Caucasian—it must have diverged from them much earlier than from Uralic, because the number of cogent resemblances is much smaller. There is no evidence that Indo-European originated by fusion of components from two or more distinct language families.

Characteristic developments of Indo-European languages. As Proto-Indo-European was splitting into the dialects that were to become the first generation of daughter languages, different innovations spread over different territories.

Indo-Iranian, Balto-Slavic, Armenian, and Albanian agree in changing the palatal stops *ḱ, *ǵ, and *ǵh into spirants (*s, ś, th*) or affricates; *e.g.,* Sanskrit *aśri-* "sharp edge." Old Church Slavonic *ostrŭ* "sharp," Armenian *asełn* "needle," Albanian *athëtë* "bitter" beside Greek *ákros* "tip," Latin *acidus* "biting," all from a basic element *H_2ek- "sharp, pointed." (Spirants, also called fricatives, are sounds produced with audible friction as a result of the airstream passing through a narrow, but unstopped, passage in the mouth; *e.g.,* English *s, f, v.* Affricates are sounds that begin as stops, with complete stoppage of the airstream, but are released as spirants, or fricatives; *e.g.,* the *ch* in "church," the *j* in "jam.") The languages that change the palatal stops to spirants or affricates are not separated from one another by any recorded languages that preserve the palatals as stops; so it is therefore inferred that the change to affricates (whence later spirants) occurred just once, and spread over a cohesive dialect area of Proto-Indo-European.

Of the languages that share this change, however, Balto-Slavic shares with Germanic (including English) an *m* in certain case endings where other Indo-European languages, including Indo-Iranian, Armenian, and Albanian, have *bh* or a sound regularly developed from *bh.* Examples of the *m* ending include English "the-m" and Old Church Slavonic *tě-mŭ* "to those ones"; the *bh* and related sounds (*ph, v, b*) are illustrated in the following: Sanskrit *té-bhyas* "to those ones," Armenian *noro-vk‘* "with new ones," Albanian *male-ve* "to mountains," Greek *ókhesphin* "with chariots," Latin *omni-bus* "for all." Because Balto-Slavic and Germanic are neighbours, it is inferred that *m* replaced *bh* in these case endings just once in the parent language, and that the area over which this innovation spread only partly overlapped the area that adopted affricated pronunciation of the palatals.

This pattern is general for changes dating from the time the parent language was breaking up into distinct languages. Each of the resulting languages shares some innovations with some of its neighbours, but only rarely do different innovations shared by two or more branches of Indo-European cover exactly the same territory.

Developments of the separated Indo-European languages

Once the dialects had become differentiated enough to be distinct languages—probably by 2000 BC, at least in most cases—each largely went its own way, and agreements in developments since then are due either to borrowing across language boundaries (as in the notable convergences between Modern Greek, Albanian, Romanian, and the southernmost Slavic languages) or to parallel but independent workings out of the same base material.

Changes in phonology. In phonology, the most striking changes have been loss or reduction in many languages of final or unaccented syllables, and loss in several languages of certain consonants between vowels, often followed by contraction of the resulting vowel sequence. Thus words in modern Indo-European languages are often much shorter than their Proto-Indo-European ancestors; *e.g.,* English "four," Armenian *č‘ork‘,* colloquial Persian *car* "four" from *k^wetwóres; French *vit* (pronounced *vi*) "lives" from *g^wíH$_3$weti; Russian *dvesti* "two hundred" from *duwoy kṃtoy.

Changes in morphology. Because much of the marking of Proto-Indo-European inflectional categories was done in final syllables, loss and reduction of these syllables have often had serious grammatical consequences. In the noun, loss of endings has generally led to loss or great reduction of the case and gender systems, while ways have generally been found to salvage the distinction between singular and plural. In Modern Persian, for example, where all final syllables have been lost, the old case and gender distinctions have disappeared also, but plural number is still regularly marked, either with *-an* (originally the genitive plural ending of some nouns) or with *-ha* (of obscure origin).

In the verb, where more endings originally had two syllables, loss of final syllables has had less serious consequences for morphology. Even here, however, some languages, including English, have totally or almost totally given up the marking of subject by personal endings. Compare English "I, we, you, they love" and "he, she loves" with the Spanish conjugation for "love"—*amo, amas, ama, amamos, amáis, aman*—or the Russian version—*ljubljú, ljúbish, ljúbit, ljúbim, ljúbite, ljúbjat.*

Changes in noun inflection have generally involved simplification. Almost everywhere the dual number has been lost; in many languages the noun genders have been reduced from three to two (as in French, Swedish, Lithuanian, and Hindi), or lost entirely (as in English, Armenian, and Bengali). Only Slavic has complicated the gender system, by imposing on the inherited distinctions contrasts of animate versus inanimate or of personal versus nonpersonal.

Reduction in cases

Everywhere except in the oldest Indo-Iranian languages the original eight Indo-European cases have suffered reduction. Proto-Germanic had only six cases, the functions of ablative (place from which) and locative (place in which) being taken over by constructions of preposition plus the dative case. In Modern English these are reduced to two cases in nouns, a general case that does duty for the vocative, nominative, dative, and accusative ("Henry, did Bill give John the letter?"), and a possessive case continuing the old genitive ("Bill's letter"). In languages such as French and Welsh, nouns are no longer inflected for case at all. In some languages, to be sure, nouns have begun fusing with words placed directly after the nouns to create new case systems, coexisting with relics of the old. Thus, Old Lithuanian had in addition to seven inherited cases an illative (place into), made by adding *-n(a)* to the accusative (*peklosna* "into hell"), an allative (place to, toward), made by adding *-p(i)* to the genitive (*Jesausp* "to Jesus"), and an adessive (place at which), made by adding *-p(i)* to the locative (*Joniep* "in John").

Changes in the verb have been more complex. Besides loss or merger of old categories, many new forms have been created and many old forms have acquired new values. In Ancient Greek the focus of the stative aspect (perfect) has largely shifted from the present state ("he is dead") to the previous event that led to this state ("he has died"). As a result, the perfect came to mean the same as the perfective past (aorist), and has therefore disappeared from Modern Greek. New forms created in Ancient Greek include future and future perfect tenses, based on the desiderative present forms (such as "he wants to walk") of the parent language.

In Germanic the principal new creation was the weak past tense (ending in a *t* or *d*), such as English "loved," "thought," German *liebte, dachte,* made by combining the verb stem with a past tense of the Germanic verb for "do." (The strong past tense formed by vowel alternations, like "sing," "sang," "run," "ran," comes from the proto-Indo-European stative aspect.)

In some languages participles (verbal adjectives) have come to function as finite verbs. Thus in Hindi *mard strī-ko dekhtā* "the man sees the woman," *dekhtā* "sees" is etymologically a participle "seeing," agreeing in number and gender with the subject *mard* "man." In the past tense, *mard-ne strī dekhī* "the man saw the woman," the verb *dekhī* is etymologically a past passive participle "seen," agreeing in gender and number with the object *strī* "woman," and the subject is marked with an instrumental ending.

Vocabulary changes. Changes in vocabulary have been even greater than those in sounds and grammar. Words in modern Indo-European languages have several sources. They may be recognizable loanwords, such as English "skunk," "chain," and "inch" (from Algonkian, French, and Latin, respectively); they may have been formed within the history or prehistory of the language itself, such as English "radar" and "rightness"; they may be of obscure origin, such as English "drink," which is common Germanic but has no cognates outside Germanic, or "boy," which is peculiar to English and Frisian; or they may be inherited words that have changed meaning, such as English "merry" from Proto-Indo-European **mr̥ǵhu-* "short." Only a small fraction of the vocabulary can be traced back to words that can confidently be asserted to have existed in the parent language with approximately their present meaning. The same is true, albeit in a lesser degree, even for the oldest recorded Indo-European languages. None has more than a few hundred words and roots that are clearly inherited from the parent language without essential change of meaning. Table 1 gives examples of words widely retained with little change. Typically they include pronouns; nouns, verbs, and adjectives of relatively simple and ubiquitous meaning; numerals; and simple adverbs and prepositions.

Non-Indo-European influence on the family. Indo-European languages, like all languages, have always been subject to influence from neighbouring languages, both related and unrelated.

Influence of non-Indo-European languages on the sounds and grammar of Proto-Indo-European is not demonstrable, partly because there is no direct evidence about the languages that were in contact with Indo-European before 3000 BC. It can be surmised, however, that some words are loans; *e.g., *peleḱus* "ax," a word for an object likely to be imported or learned of from neighbours with superior technology, and which is not analyzable into a known Indo-European root plus a known Indo-European suffix.

Adoption of loanwords

When Indo-European languages have been carried within historic times into areas occupied by speakers of other languages, they have generally taken over a number of loanwords, as with English and Spanish in the Americas or Dutch in South Africa. Aside from the special case of the pidgin and creole languages, however, there has been very little effect on sounds and grammar. These have been significantly affected within historic times only when an Indo-European language has been spoken in prolonged close contact with non-Indo-European speakers, as with Ossetic (an Iranian language) in the Caucasus, or when its speakers have been very strongly influenced culturally by speakers of a non-Indo-European language, as with Persian, in which Arabic plays much the same role as Latin does in English.

In prehistoric times most branches of Indo-European were carried into territories presumably or certainly occupied by speakers of non-Indo-European languages, and it is reasonable to suppose that these languages had some effect on the speech of the newcomers. For the lexicon, this is indeed demonstrable in Hittite and Greek, at least. It is much less clear, however, that these non-Indo-European languages affected significantly the sounds and grammar of the Indo-European languages that replaced them. Perhaps the best case is India, where certain grammatical features shared by Indo-European and Dravidian languages appear to have spread from Dravidian to Indo-European rather than vice versa. For most other branches of Indo-European languages any attempt to claim prehistoric influence of non-Indo-European languages on sounds and grammar is rendered almost impossible because of ignorance of the non-Indo-European languages with which they might have been in contact. (W.C.)

Anatolian languages

The term Anatolian languages in its most comprehensive use includes both the Indo-European and non-Indo-European languages spoken in Anatolia (Asia Minor) before the Greco-Roman period. The Anatolian languages are known only from texts of the 2nd and 1st millennia BC; the earliest evidence is that of the so-called Cappadocian tablets (19th–18th century BC). The term Asianic is sometimes used as an alternative designation for the Anatolian languages, but, since the discovery in 1915 that Hittite, the main Anatolian language, is an Indo-European language, there has been a tendency to use Asianic in a more restricted sense for the non-Indo-European languages that existed in Anatolia before the entry of the Indo-Europeans. These are called substratum languages.

Hattic (or Hattian), also misleadingly called Proto-Hittite, is the best known substratum language. It is completely unrelated to Hittite and its sister languages as well as to Hurrian, a language also spoken in Anatolia.

The Anatolian group of Indo-European languages consists of Hittite, Palaic, Luwian, Hieroglyphic Luwian, Lydian, and Lycian. Hittite, Palaic, and Luwian are known from 2nd-millennium cuneiform texts found in the excavations in Boğazköy-Hattusa since 1905; Hieroglyphic Luwian is found on scattered inscriptions and seals from Anatolia (mainly the southern area) and northern Syria dating mainly from later times (*i.e.,* between *c.* 1200 and 700 BC, although there are earlier examples from the empire period, *c.* 1400–*c.* 1190 BC). Lydian and Lycian are known from texts in alphabetic script from *c.* 600 to 200 BC. It seems fairly reasonable to add the Carian language of southwest Anatolia to this list as well as other less well documented languages like Sidetic. More to the east, in the Caucasus region centring around Lake Van, Hurrian of the 3rd and 2nd millennia BC was replaced in the 1st millennium BC by the related Urartian language. Both of these languages are definitely non-Indo-European.

Historical background of ancient Anatolia. It is customarily assumed that the Indo-Europeans entered Anatolia around or shortly after 2000 BC, although there are no specific archaeological data that might enable scholars to specify the period of entry or the route the invaders followed. On the basis of the agricultural terminology used in Hittite, it has been suggested that the entry into Anatolia was not a warlike invasion of predominantly male groups. If such had been the case, the influence of substratum languages would have been likely, but, on the contrary, the word stems used are definitely Indo-European. The differences in the terminology used in other Indo-European subgroups indicate that the "Anatolians" seceded from the parent group at an early date, before the common agricultural nomenclature came into being. On the other hand, Hittite shares the Indo-European notion of the hereafter, pictured as a pastureland with grazing cattle "for which the dead king sets out."

Indo-European entry into Anatolia

There is a tendency among linguists to postulate an eastern route of entry into Anatolia by way of the Caucasus, because certain grammatical features—*e.g.,* the loss of the feminine gender—might be explained as having been caused by prolonged contacts with Caucasian languages.

It is likely that the Indo-European forebears of the later speakers of Hittite, Palaic, Luwian, and Lydian entered Anatolia together, following a common route, because the Anatolian languages share a considerable number of losses as well as innovations that presuppose a long common past.

In the central parts of Anatolia, within the bend of the Halys River (modern Turkish, Kizil Irmak), and in the northern regions, Hittite and Palaic were profoundly influenced by Hattic as a substratum language. The Hattian culture also changed the political and religious concepts of the newcomers, and a clear cultural dependency of the Indo-Europeans on the older Hattian population is evident. Some scholars have stressed the likelihood that farther to the south the Luwians might have been conversant with a different substratum. In view of the absence of textual evidence, and because knowledge of the Luwian vocabulary is rather restricted, it is perhaps not surprising that this possible substratum element escapes definition. (For the history of Anatolia in the 2nd and 1st millennia BC, see TURKEY AND ANCIENT ANATOLIA.)

The most important invaders of Anatolia in the "Dark Age" (after 1190 BC) were the Phrygians. Their language is definitely Indo-European, but it bears no relationship to the Anatolian subgroup. Rather, it seems akin to Thracian, Illyrian, or possibly Greek. Greek, in the second half of the 1st millennium BC, and, later, Latin, from the 2nd century onward, entered central Anatolia as languages of a ruling caste. Much earlier—beginning in Mycenaean times—the west coast had attracted Greek settlers. In the first half of the 1st millennium, the southern and northern shores also attracted Greek-speaking peoples. To the east in the Caucasus region, other Indo-Europeans, the Armenian-speaking invaders, penetrated into the former Urartian territory well before the beginning of the Persian period, probably in the 7th and 6th centuries BC. During Persian times, a Persian ruling caste entered eastern and also northeastern Anatolia and was still clearly recognizable in the Hellenistic and Roman periods (*e.g.*, in Bithynia, Pontus, Cappadocia, and Commagene). Late data on names and scattered remarks made by Fathers of the Church indicate that until late Roman and perhaps even Byzantine times, some Anatolian dialects remained in use in certain isolated parts of the interior.

Early research on Anatolian languages

Classification of the languages. Research on the Anatolian languages began in 1821 with the Lycian language and passed an initially fruitful phase in the 1880s with work on Hieroglyphic Hittite (nowadays referred to as Hieroglyphic Luwian). In 1902 the Norwegian Assyriologist Jørgen Alexander Knudtzon's study on the Arzawa letters was published; these were two letters exchanged between a king of Arzawa and Pharaoh Amenhotep III that had been found in the Amarna archive. They were written in the Hittite language in cuneiform writing. In 1915 research reached a climax with the interpretation of Cuneiform Hittite by the Czech Orientalist Bedřich Hrozný. In all four of these highlights, the discovery that the texts in question were Indo-European was either clearly expressed or more discreetly implied. This conclusion was based on both the nominal (noun) declension and the verbal conjugation: the languages had a nominative ending in *-s,* the accusative in *-n,* verbal endings like *-ti* and *-nti* for the 3rd person singular and plural of the present tense, and an imperative form like *estu* "let it be." These features were deemed to be sufficient proof of their Indo-European origin. Study of the Anatolian subgroup of Indo-European thus began with Lycian, the last Anatolian offshoot in the temporal sequence, then passed the intermediary stage of Hieroglyphic Luwian, and reached the 2nd-millennium Hittite language in 20th-century research. For the relationship between members of the Anatolian subgroup, see Figure 2.

The non-Indo-European Hurrian and Urartian languages are related to one another, but modern research indicates that Urartian should not be considered as a direct continuation of Hurrian.

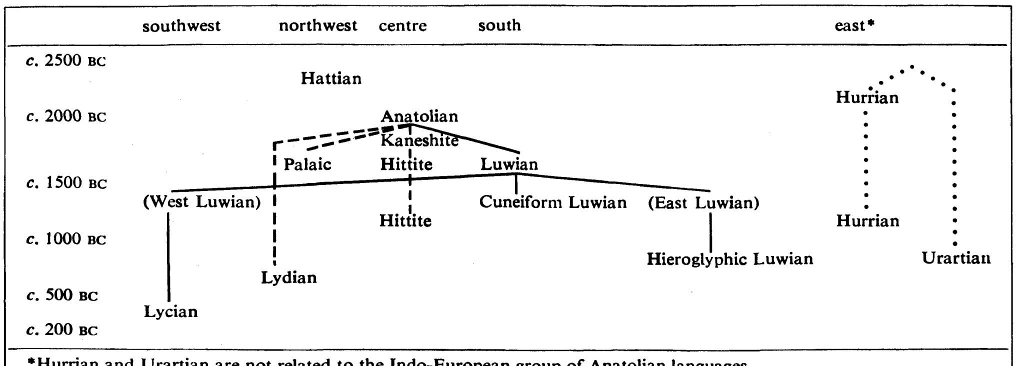
HISTORY AND DEVELOPMENT

Languages using cuneiform writing and Anatolian hieroglyphs. *Hattic.* The Hattic language appears as *hattili* in Hittite cuneiform texts. Called Proto-Hittite by some, it was the language of the linguistic substratum inside the Halys River bend and in more northerly regions. Apparently the Indo-European newcomers of Hittite stock were named with the same designation as their predecessors. All the Hattic material preserved by Hittite scribes belongs to the religious sphere of life: rituals (*e.g.,* connected with the erection of a new building), incantations, antiphons, litanies, and myths. Among the Hattic interpolations in Hittite texts, there are some to which a Hittite translation has been added. It is impossible to ascertain the length of time that the Hattians had been present in Anatolia before the Indo-Europeans entered the country, but it seems certain that during the Hittite New Empire (*c.* 1400–*c.* 1190 BC) Hattic was a dead language.

Records of Hattic

Hattic studies began in 1922 with the work of the German Assyriologist Emil Forrer. In 1935, Hans G. Güterbock, a German-born Orientalist, published a large group of texts containing Hattic material and in so doing completed the publication of the Hattic texts stemming from the Winckler excavations (1905–12). Important studies on the subject have continued to appear since then.

Hittite. The Hittite language is known from the approximately 25,000 tablets or fragments of tablets preserved in the archives of Boğazköy-Hattusa, excavated by German archaeologists beginning in 1905. In Hittite cuneiform texts, the language is referred to as *nesili* (*nasili*) "language of Nesa," or *nesumnili* "language of the Neshite." Earlier Hittite linguistic material may be found in the indigenous proper names and a few loanwords from the local dialect that are recorded in the Cappadocian tablets (the commercial correspondence in Assyrian of Assyrian colonists living in Anatolia, especially in the emporium at Kültepe, near modern Kayseri, between *c.* 1900 and 1720 BC). The data from Kültepe are sometimes referred to as "Kaneshite" (from Kanesh, the old name of Kültepe); this is obviously the modern equivalent of the word *kanisumnili* "language of the Kaneshite" found in a Hittite text. It is possible, or even likely, that Kanesh and Nesa do, in fact, refer to the same entity.



*Hurrian and Urartian are not related to the Indo-European group of Anatolian languages.

Figure 2: Relationship between members of the Anatolian subgroup.

Hittite tablets from places outside of the Hittite capital are rare; only stray examples have been found—*e.g.*, in Tarsus, Alalakh, Ugarit, and Amarna. These findings attest to the growth of a great Hittite empire, especially between *c.* 1400 and *c.* 1190 BC. Old Hittite, the written embodiment of the earliest Indo-European language that has been discovered so far, is known from some tablets preserved in an "old ductus" type of handwriting that was typical of copies from the Old Kingdom period (*c.* 1700–1500 BC). The intermediary "Dark Age" between *c.* 1500 and *c.* 1400 BC is sometimes referred to as the period of the so-called Middle Hittite language. Most of the Old and Middle Hittite texts, however, are preserved in copies from the later empire period.

Archives of Boğazköy-Hattusa

The archives of Boğazköy-Hattusa have been found in various places in the citadel, in the Great Temple complex, and in the "House on the Slope." Although the majority of the texts are concerned with religious subjects (oracle texts, hymns, prayers, myths, rituals, and festival texts), these archives also contain material of historical, political, administrative, literary, and legal character. The cuneiform adopted by the Hittite scribes is a variant of a writing system of Mesopotamian origin that closely resembles the ductus and shapes prevalent in tablets of the 17th century BC (layer VII) from Alalakh (modern Atsana in southeastern Turkey). It is possible that the cuneiform script might have been introduced as a result of the Hittites inducing Syrian scribes to transfer their activities to the Hittite capital during the early part of the Old Kingdom, shortly after 1650 BC. It has also been posited, with good reason, that the newly acquired script was first used to write Akkadian and was only later employed for Hittite as well. In addition to the genres enumerated above, the "scholarly literature" deserves to be mentioned. This consists of the material considered by the scribes to be essential for their training; it includes word lists, omens, and ritual prescriptions, all reflecting an encyclopaedic approach aimed at complete coverage of the subjects concerned. The Sumerian texts found in these archives belong to this class of literature. For treaties and correspondence with foreign powers, Akkadian was used as the diplomatic language of that period. Therefore, both Sumerian and Akkadian formed part of the curriculum of the qualified scribes, these languages belonging to the "eight languages" found in the Hittite archives.

In actual fact, the first decipherer of Hittite was the Norwegian scholar J.A. Knudtzon, who pointed out in 1902 that the language of the so-called Arzawa letters (*i.e.*, Hittite)—found in the Amarna archive—had an apparent affinity with Indo-European. Because the cuneiform script had already been deciphered, Knudtzon, and Bedřich Hrozný after him, were able to "read" their texts. Thus their discovery consisted more in the interpretation than in the actual decipherment of the written material. The first series of German excavations, lasting from 1905 to 1912, produced about 10,000 tablets. It was work on this corpus that familiarized Hrozný with the contents of these tablets and led him to his epoch-making discovery that Hittite was indeed Indo-European (1915).

Palaic. Palaic, which appears as *Palaumnili* "language of the Palaite" in Hittite cuneiform texts, was the language of the region of Pala (probably Blaëne in the Greek period), in northwest Anatolia. During the Old Hittite kingdom, Pala, Luwiya, and Hattusa formed the three major provinces of the Anatolian part of the Hittite territory. From the intermediary "Dark Age" onward, Kaska nomads made their influence felt in northern Anatolia, and this resulted in a decline of importance for this region.

Linguistic affiliations of Palaic

The Indo-European character of Palaic was first advocated by Emil Forrer (1922). Part of the text material is preserved on tablets in "old ductus." The knowledge of the limited vocabulary leaves much to be desired, but parallels—especially in the inflection of the noun, the forms of the demonstrative, relative, and enclitic pronouns, and the verbal endings—vouch for a close relationship to Hittite and Luwian.

Luwian. Luwian (or Luvian), the language of Anatolia's southern coast, is known from texts stemming from three major periods: (1) the Hittite New Empire (*c.* 1400–*c.* 1190 BC); (2) the period of the Neo-Hittite states (*c.* 1190–*c.* 700 BC); (3) the period of the Lycian monumental inscriptions (*c.* 400–200 BC). In addition to the various time periods, there is also a variation in writing system—Mesopotamian cuneiform, Anatolian hieroglyphs, and an alphabet derived from a Greek source—and dialectal differentiation. There are indications that as early as the 15th and 14th centuries BC, there was a West Luwian dialect (the precursor of alphabetic Lycian) and an East Luwian dialect (the forerunner of the later Hieroglyphic Luwian of the Neo-Hittite states). Both of these differed from the Luwian found in the archives of Boğazköy-Hattusa, which was possibly a central dialect.

As in the case of Palaic, the pioneering work on Luwian written in cuneiform was done by Emil Forrer (1922). Following this work, new text materials were published in 1953, closely followed by both grammatical and vocabulary studies as well as a standard dictionary of Cuneiform Luwian (1959).

The Anatolian hieroglyphic system has a long history, with its logographic beginnings dating back to early Hittite stamp seals of the 18th and 17th centuries BC; the youngest texts seem to date from the last quarter of the 8th century BC. The geographical range of the inscriptions is great, stretching from Sipylus and Karabel in the extreme west to Alaca Hüyük and Boğazköy-Hattusa in the north, Malatya, Samsat, and Tell Ahmar (Til Barsib) in the east, and Hama and ar-Rastān in the south. During the "Dark Age" of the 16th and 15th centuries BC, the early writing grew into a fully developed writing system with logograms (word-signs), syllabic values, and auxiliary signs. During the New Empire, the script was already in use for a multitude of purposes (rock inscriptions, seals, and wooden tablets for everyday use in the temple and the army). Whether an example of the empire period such as the Aleppo inscription already reflects the Luwian language is a moot question but seems likely. It is certain that the later inscriptions of the Neo-Hittite states were in Luwian.

Work on Hieroglyphic Luwian

The first attempts to decipher Hieroglyphic Luwian, made by the British archaeologist Archibald H. Sayce, were fortunate in some fundamental details, but it was not until the 1930s that systematic and mutually stimulating research by scholars of several countries led to the establishment of a number of syllabic values for the characters as well as to a correct analysis of the sentence structure of the inscriptions. In his publication of the (bilingual) Hittite royal seals (in 1940, 1942), Hans G. Güterbock bridged the gap between the inscriptions of the empire period and the late Neo-Hittite states; the seals found in the French excavations at Ugarit (in northern Syria) served a similar purpose. The most important recent finding was the discovery in 1947 by Helmuth T. Bossert, a German archaeologist, of the Karatepe bilingual inscriptions, written in Phoenician and Hieroglyphic Luwian.

On many points the Luwian vocabulary is still an enigma. The unity between the various Luwian dialects and the close relationship of Luwian to the other members of the Anatolian subgroup, however, is secured by several linguistic parallels, especially in the singular inflection of the noun, the forms of certain pronouns, the verbal endings, and a number of lexical (vocabulary) correspondences.

Hurrian. In earlier stages of research, the terms Mitanni language and Subarian were used as designations for Hurrian. In Hittite cuneiform texts, *hurlili* "language of the Hurrian" is used. In the last centuries of the 3rd millennium BC, Hurrians were already present in the Mardin region, which, from a geographical point of view, belongs to the North Mesopotamian plain. In Mesopotamian texts (from the time of the Akkad dynasty) some Hurrian personal names and glosses have been found. The customary assumption is that this non-Semitic and also non-Indo-European ethnic group had come from the Armenian mountains. During the beginning of the 2nd millennium BC, the Hurrians apparently spread over larger parts of southeast Anatolia and northern Mesopotamia. Still later, during the intermediary "Dark Age," they are supposed to have infiltrated into Cilicia and the adjacent Taurus and Antitaurus regions (Kizzuwatna in 2nd millennium texts). Before the middle of the 2nd millennium BC, an

Indo-Aryan ruling caste wielded some type of authority over parts of Hurrian territory. Some names and words in ancient Near Eastern texts bear witness to their presence. Among these words are a group of technical terms related to the training of horses that found its way into Hittite treatises on that subject; they are most important from a historical point of view. After Sumerian, Akkadian, Hattic, Palaic, and Luwian, Hurrian and these Indo-Aryan glosses constitute the sixth and seventh additional languages of the Hittite archives.

Sources of Hurrian texts

Hurrian texts have been found in Urkish (Mardin region, *c.* 2300 BC), Mari (on the middle Euphrates, 18th century BC), Amarna (Egypt, *c.* 1400 BC), Boğazköy-Hattusa (Empire period), and Ugarit (on the coastline of northern Syria, 14th century). Amarna yielded the most important Hurrian document, a political letter sent to Pharaoh Amenhotep III. From Mari came a small number of religious texts; from Boğazköy-Hattusa, literary and religious texts; and from Ugarit, vocabularies belonging to the more "scholarly literature" described above and Hurrian religious texts in Ugaritic alphabetic script. Hurrian personal names, found in texts from many sites (Boğazköy-Hattusa, Alalakh, Ugarit, and especially Nuzu), constitute a second linguistic source of major importance.

The research on Hurrian started in the 1890s with simultaneous contributions by several scholars. Subsequently, Bedřich Hrozný (1920) and Emil Forrer (1919, 1922) discovered the presence of Hurrian material in the Boğazköy-Hattusa archives.

Urartian. The terms Chaldean and Vannic have also been used as designations for Urartian during earlier stages of research. Urartian is not a late dialect of Hurrian but a separate language, although both stem from a common parent. During the 9th through 6th centuries BC, Urartian was used in northeastern Anatolia as the official language of the state of Urartu, which centred around the district of Lake Van but also extended over the Transcaucasian regions of modern Russia and into northwestern Iran and at times even into parts of North Syria. The Urartian texts are written in a variant of the Neo-Assyrian script and consist mostly of monumental inscriptions (annals, votive inscriptions related to building and irrigation activities), some small inscriptions on helmets and shields dedicated in the temple, and a few economic cuneiform tablets. Two bilingual inscriptions in Urartian and Assyrian that apparently correspond very closely provided the key to the understanding of the language; the stylistic resemblances to Assyrian texts of the same period guided the further interpretation.

Archibald H. Sayce was the first scholar to devote his attention to Urartian in the 1880s and 1890s and continued his activities until 1932. More important were the philological contributions of the German historian Carl F. Lehmann-Haupt between 1892 and 1935. The first reliable description of Urartian grammar was published by the German Orientalist Johannes Friedrich (1933).

Next to the Urartian texts in cuneiform writing, there also existed an indigenous hieroglyphic script that is still undeciphered and is too meagrely represented to warrant a serious attempt.

Old Phrygian and New Phrygian

Languages using a derivative of the Phoenician or Greek alphabet. *Phrygian.* The Phrygian inscriptions and graffiti may be separated into two groups, the Old Phrygian texts in a typical Phrygian alphabet dating from *c.* 730–450 BC, and the New Phrygian inscriptions (sepulchral texts in the Greek alphabet) stemming from the 1st and 2nd centuries AD. The Old Phrygian texts may be divided into a central group (Midas City and the central area) and an eastern group (found in Gordium), with offshoots in a still more eastern direction marking the utmost Phrygian expansion (inscriptions in or around Hüyük near Alaca, in Boğazköy-Hattusas, and in Tyana). An important recent finding—and the longest Old Phrygian text to date—is the rock inscription near the village of Germanos (modern Soğuk Çam) in Bithynia (found in 1966). The total number of Old Phrygian texts now stands at about 80; more than 50 of these are from Gordium and represent about one-quarter of the available material. There is a consensus of opinion on the Indo-European character of the Phrygian language; most scholars think that Phrygian is somehow connected to the Greek branch of Indo-European languages, although, at an earlier stage, some scholars considered the possibility of a connection with the Anatolian branch of Indo-European, and others proposed a relationship with Thracian and Illyrian.

In a publication of new material from Gordium, the U.S. archaeologist Rodney S. Young cautiously suggested that the Old Phrygian alphabet may be dependent on a prototype in use on the North Syrian or Cilician coasts. The old idea that the Phrygian alphabet was dependent on a Greek one (and not vice versa) need not be abandoned in that case. Historically, such a derivation would present no problems, because the presence of Greek settlements in these areas during the second half of the 8th century BC is amply attested to by both Assyrian annals and late Greek historical sources as well as by archaeological findings. Internal evidence from the Phrygian alphabet, presented by the French linguist Michel Lejeune, serves as proof for some researchers that that alphabet derived from the Greek one.

Lydian. Of the more than 70 Lydian texts (*e.g.,* sepulchral inscriptions, votive texts, many graffiti), more than half have been found by United States excavators at the Lydian capital, Sardis. Two small Greek–Lydian bilingual texts were far less helpful than the famous Aramaic–Lydian text. A few texts (about ten) may go back to the 6th or 5th centuries BC, but many more stem from the 4th century. The Lydian alphabet was derived from an East Greek prototype; the superfluous signs in the Greek alphabet were used for specific Lydian sounds, and additional signs were either borrowed from other "Anatolian alphabets" or freely created.

Research on Lydian

Important results concerning Lydian were reached using a strictly combinatory method; *i.e.,* passages were compared that expressed similar contents in a slightly different manner in order to obtain a better understanding of the language's structure. This stage of the research culminated in a conclusive article by the Italian Piero Meriggi on the Indo-European character of Lydian (1936). Subsequently, other scholars published evaluations of the Lydian data, a dictionary, and a grammar book. The study of Lydian is hampered by many lexicological uncertainties, but there is at least a growing consensus on matters of grammar leading to the common notion that Lydian belongs to the Anatolian subgroup of Indo-European. The final obstacle to this classification as Anatolian was removed in 1959 by the Italian Onofrio Carruba, who proved that Lydian, like the other members of the Anatolian branch, does not possess a separate feminine gender. Lydian shares common features with Hittite, Palaic, and Luwian and should therefore be acknowledged, it seems, as a fourth independent member of the Anatolian subgroup.

Carian. A great number of the more than 100 Carian inscriptions are graffiti found in Egypt that were left behind by Carian mercenaries in the services of Egyptian pharaohs of the Saitic period (664–525 BC). In recent years, more monumental inscriptions have been found in Caria itself, and Carian clay tablets have also been discovered. In the mid-20th century, several scholars concluded that Carian writing consists of a purely alphabetic script and is not a mixed system of both single letters and syllabic signs as was formerly thought. It is a likely but still unproven assumption that Carian may also be classified in the Anatolian subgroup of Indo-European.

Lycian. More than 150 Lycian monumental inscriptions have been found so far, which, with very few exceptions, are sepulchral in character. They are written in an indigenous Lycian alphabet that is based not on an East Greek prototype (as its Lydian replica) but on a West Greek one. Although the Lycian coin legends are still usually dated from the period between 500 and about 360 BC, the tradition of the Lycian monumental inscriptions is now thought to have continued for a longer period, into the 3rd century BC. During the beginning of the research in the first half of the 19th century, extensive use was made of a good bilingual text that offers a faithful Greek translation. In the first phase of research, which ended about 1880, Lycian was investigated by an etymological method by which it was linked up either with Greek or the Iranian languages. A more reliable combinatory method was later introduced, but the most fecund phase in the study of Lycian occurred at the end of the 19th century, when the Scandinavian school of scholars cooperated closely in the publication of several important studies. In 1945, Holger Pedersen published a synthesis of all data that seemed to indicate a relationship of Lycian with Hittite; thus Pedersen proved conclusively that Lycian belongs to the Anatolian branch of Indo-European languages. This conclusion was slightly modified when the British scholar Franz J. Tritsch (in 1950), and, later, the French scholar Emmanuel Laroche showed that Lycian should be more specifically compared to Luwian.

Relationship of Lycian with Anatolian

Sidetic. The historical detail preserved by the Greek historian Arrian that the city of Side on the Pamphylian coast possessed a particular, indigenous language has been strikingly confirmed by legends on Sidetan coins of the 5th (?) through the 3rd (?) centuries and by five inscriptions from the 3rd and 2nd centuries BC (two of which are bilingual). There is a curious likelihood that this alphabet was directly derived from a Semitic writing system rather than from a Greek prototype, but Greek influence was not absent, as is clearly evidenced in the Greek bilingual texts and by a loanword from Greek. The first reliable study of Sidetic was made by Helmuth T. Bossert in 1950. In the case of Sidetic, even the value of a group of signs is still undecided, and research has not yet reached a stage in which a fruitful analysis of the texts and a classification of the language are within sight.

LINGUISTIC CHARACTERISTICS

Non-Indo-European languages. The non-Indo-European Hattic is an agglutinative language; that is, it combines several elements of meaning into a single word. In the conjugation of verbs, it uses prefixes that are attached to the word stems, which are mostly monosyllabic or bisyllabic. Hattic nouns consist of a free number of syllables and have both prefixes and suffixes. There are, however, no formal distinctive features to distinguish nouns and verbs.

Both the Hurrian and the Urartian languages differentiate between stems and suffixes, but there is again no sharp distinction between noun and verb. Many suffixes may be added onto one another in a row, but within the often prolonged suffix series a detailed order is rigidly observed. Among the suffixes added to the noun, several subgroups are distinguished; one group might be compared to the case endings of the Indo-European languages. One of the most characteristic phenomena of this group is the distinction between a subject case (the "nominative") and an "agentive" case. The agentive marks the actor or subject of a transitive verb when the object is expressed by its counterpart, the "nominative." The subject case is characterized by a lack of ending on the stem; it marks the subject in nominal sentences (sentences without verbs) and occurs with intransitive verbs and as the object of transitive verbs.

Subject and agentive cases in Hurrian and Urartian

Phrygian. The New Phrygian texts especially favour the attribution of the Phrygian language to Indo-European. They contain such data as *ios* as relative pronoun (Indo-European **io-s,* Greek *hos*), a demonstrative pronoun that is either comparable to Indo-European **ki-/ko-* or to **so* (an asterisk indicates a hypothetical reconstructed form), and the form *ad-daket* "he adds" related to Latin *addit* and to Greek *é-thē-ka.*

The Anatolian subgroup of Indo-European. *Grammatical characteristics.* Characteristic of the Anatolian languages is the absence of the dual number ("you and I") and the lack of feminine gender in the declension of nominals (nouns, pronouns, and adjectives). There is a division between an animate (common) gender and an inanimate (neuter) gender. In Hittite, a neuter may not be the subject of a transitive action verb; in that case, an *-ant* suffix is added before the neuter nominative ending in *-s.* This *-s* ending persists in the whole subgroup. The case system of Old Hittite is still fairly complicated, but in the subgroup as a whole there is a clear tendency toward a greater simplicity. The presence in Hittite of an archaic irregular class of nouns is a striking characteristic; *e.g.,* there are alternate *r* and *n* stems as in *uttar/uttanas* "word, affair" and *watar/witenas* "water."

The Anatolian inflection of pronouns conforms to the traditional Indo-European pattern by being different from that of the nouns, but, at the same time, it shows some striking peculiarities. Typical Anatolian pronouns are: Hieroglyphic Luwian *amu,* equivalent to Lycian *amu, emu, ẽmu* "I, me" (compare Hittite nominative *uk,* accusative *ammuk*); and Hittite nominative *zik,* accusative *tuk* "you," as compared to *ti, tu* in Palaic. Some of the languages have enclitic pronouns; *i.e.,* pronouns pronounced as being part of the preceding word. A demonstrative pronoun *aba-* ("that," but in some member languages also "this") is present in Hittite, Palaic, Cuneiform and Hieroglyphic Luwian, Lycian (*ebe-*), and Lydian (*ebad* "here, there"), and an interrogative or relative pronoun *kui-* (compare Latin *quis*) is common to Hittite, Palaic, and Cuneiform Luwian. The corresponding terms for *kui-* in Hieroglyphic Luwian, Lycian, and Lydian also seem to be phonetic variants of the same original pronoun.

Anatolian verbs

The Anatolian verbal system is simple: it has two moods (indicative and imperative) and two tenses (present and preterite). There are some traces—either to be classified as debris or as the nucleus for a future development—of an aorist *-s* fixed to the stem; *e.g., kaness-* "to acknowledge" (compare Greek *gi-gnō-sk-ō*); *kalless-* "to call" (compare Greek *kaleō,* aorist *é-kale-s-a*). (The aorist is a verb form denoting action without reference to its duration or completion.) A mediopassive "voice" is present in Hittite (*es-a-ri* "he is seated"; *ki-tta-ri* "he is lying"), Palaic, Luwian, and perhaps in Lydian. (The mediopassive expresses a type of reflexive meaning ["He washes himself"] or passive meaning ["He is being washed"].)

Reduplication (repetition) of the verbal stem occurs in the entire Anatolian subgroup. It adds an iterative or intensive nuance to the meaning, but it does not function in a system of tenses as in Greek. Very typical of the Anatolian subgroup are verbal suffixes like the causative *-nu-* (compare Hittite *war-* "to burn," *warnu-* "to kindle," *harg-* "to perish," *harganu-* "to ruin, to destroy"). In principle, these formations can be built on any verbal stem whose meaning permits such an addition. It should be stressed that in Hittite a normal expression for a "state" consists of a nominal sentence (that is, a sentence without a verb but with a noun, an adverbial expression, or a participle as predicate); sometimes, however, the verb *es-* "to be" is used as the carrier of modal or temporal nuances. The total absence of the Indo-European perfect (describing a "state" resulting from a recently concluded action) becomes very clear by the usage of the adverb *nawi* "not yet," which occurs with a present tense in Hittite (but which would employ a perfect tense, such as "has been," in English and other Indo-European tongues).

Very characteristic of the Anatolian subgroup is a strong preference for the linking together of particles and enclitic pronouns to form "chains" that are placed at the beginning of the sentence or clause. The first component of such a "chain" usually is a stressed part of the sentence or otherwise a sentence connective (like *nu* in Hittite, *a* in Luwian).

Phonological characteristics. In the Anatolian vowel system, *a, e, i,* and *u* are present, but *o* is curiously absent. In Lycian, the Greek value omicron has been used for the Lycian *u,* but in Lydian the existing *o* seems to be a secondary development. A main dialectal criterion is the treatment of Anatolian *e:* in Old Hittite, there still was a differentiation between *e* and *i,* but in later Hittite, an *-e* at the end of a word changed to *-i.* In Luwian, *e* tended to appear as *a.* Vowel gradation (*i.e.,* a change of vowel) that reflects meaning change plays a role in Hittite (*e-es-zi* "he is" versus *a-sa-an-zi* "they are") but was impossible in Luwian because of the sound change. Both Lycian and Lydian possess separate signs for nasalized vowels (*ã* and *ẽ*).

Laryngeal theory and Anatolian data

Advocates of the so-called laryngeal theory (first proposed by the Swiss linguist Ferdinand de Saussure in 1879) have found their postulate partly confirmed by Anatolian data. This theory maintains that the different forms of certain

Table 3: Anatolian Lexical Correspondences

	Hittite	Palaic	Cuneiform Luwian	Hieroglyphic Luwian	Lycian	Lydian
to make, to do	iya-		aya-	aia-	a-/e-	i- (?)
in	anda		anda	anta	ñta/e	(-)ẽn-/(-)ẽt-
to be	es-/as-	es-/as-	as-	as-	es-	e- (?)
house	pir-/parn-		parn-	parn-	*prñn-	bira-
to give	pai-/piya-	pi(ya)-	piya-	pi-	pije-	bi-
up	ser/sara		sarri	logogram	hri	
high, superior	sarazzi-			*sarli- (?)	hrzzi-	serli-
god	siu-/siwi-/ siwa-/siun-/ siuni-/siuna-		massani-	masana/i-	mahani-	civ-
divine	siunali-		massanalli-			civvali-

*Indicates that the following form is an unattested, reconstructed form.

words in the various Indo-European subgroups can be satisfactorily explained only by assuming that all the known Indo-European languages have lost certain guttural sounds (laryngeals) that were originally present in the parent speech. In 1927, both the Polish linguist Jerzy Kuryłowicz and the French scholar Albert L.M. Cuny announced their discovery that in Hittite an *h* sound was preserved in positions in which a laryngeal would have formerly been (compare Hittite *hant-* "front" to English *anti-;* Hittite *pahhur* "fire" but English *pyre*). But the Anatolian evidence for the laryngeal theory is certainly not without problems, and the adherents of the theory consider that other laryngeals disappeared in Hittite as well.

Lexical data. Some examples of correspondences in vocabulary are given in Table 3. It has often been remarked—and not without reason—that although the grammar of the Anatolian languages would be recognizably Indo-European, the vocabulary would be less so. This is usually attributed to the deeply penetrating influences exercised by strange surroundings, not only while the "Anatolians" were "en route" but also after their arrival in Anatolia.

The relationship with the other subgroups. The relationship of the Anatolian branch to the rest of Indo-European has often been defined in the United States on the basis of the "Indo-Hittite hypothesis." That is, Hittite or Anatolian on the one hand and Proto-Indo-European on the other were both supposed to descend from a common parent. This hypothesis attributes too much weight to the Anatolian evidence. It was demonstrated as early as 1938 that the Anatolian branch should be placed on a par with the rest of the Indo-European subgroups and not as a coequal with Indo-European itself. Nowadays, the Indo-Hittite hypothesis is very rarely defended. Another extreme position states that the Hittite-Luwian-speaking group (another designation for the Anatolian subgroup) left the Indo-European parent group comparatively late, after the Greek and Armenian divisions had done so and approximately at the same time as Indo-Iranian. If this theory were true, there would be no need to use the Anatolian data for a thorough revision of the reconstructed Proto-Indo-European language, because these data would be less relevant, at least not more so than Indo-Iranian and Greek, on which the old reconstruction was based. A third opinion—prevalent in the French school of Indo-European studies—holds that the Hittite or, preferably, the "Common Anatolian" data are of special importance, because the Anatolian languages are particularly archaic. According to this theory, similarities in morphology (word elements) between the Celtic, Italic, and Hittite–Luwian groups and Tocharian (an Indo-European language of central Asia) seem to imply that the dialects from which these groups evolved were in peripheral positions in the Indo-European language area and were probably the first to move away from the main group. (Ph.H.J.H.t.C.)

Indo-Hittite hypothesis

Indo-Iranian languages

The Indo-Aryan and Iranian languages together constitute the Indo-Iranian language group, the easternmost major branch of the Indo-European family of languages. Indo-Aryan (Indic) languages are spoken by approximately 660,000,000 persons in India, Pakistan, Sri Lanka (formerly Ceylon), Nepal, Bangladesh (former East Pakistan), and other areas of the Himalayan region. In addition, languages of the Indo-Aryan group are spoken by about 5,000,000 people in Europe, Africa, the Americas, and Oceania: the Gypsy, or Romany, dialects distributed about the U.S.S.R., the Middle East, Europe, and North America are of Indo-Aryan origin. Speakers of Iranian number about 62,000,000 and live in areas extending from Pakistan (former West Pakistan) to Iran, Afghanistan, and the southern U.S.S.R. Among the Indo-European languages, only Mycenaean Greek and Hittite possess older records than those of Indo-Iranian.

The Indo-Iranian tongues have been used as both administrative and literary languages. Old Persian was the administrative language of the early Achaemenian dynasty dating from the 6th century BC; and an eastern Middle Indo-Aryan dialect was the language of the chancellery of King Aśoka in India in the mid-3rd century BC. As literary languages, the Indo-Iranian languages have been used in the texts of some of the world's great religions: Indo-Aryan for Buddhism, Hinduism, and Jainism, and Iranian for Zoroastrian and Manichaean texts. The oldest Zoroastrian texts are in dialects included under the name Avestan. Commerce, conquest, and religion spread the influence of these languages. Indo-Aryan languages, for example, penetrated deep into Southeast Asia; names in Indonesia and other areas and Sanskrit texts in Cambodia reflect this influence.

The close relation between the Iranian and Indo-Aryan groups has never been doubted. They share characteristic features that set them apart as a subgroup of Indo-European. The long and short varieties of the Indo-European vowels *e, o,* and *a,* for example, appear as long and short *a:* Sanskrit *manas-* "mind, spirit," Avestan *manah-*, but Greek *mĕnos* "ardor, force." (In the following examples, ¯ indicates a long vowel; ˘ indicates a short vowel. The spellings used in this article for Indo-Aryan and Iranian forms are traditional transliterations for the most part. In some cases, more accurate phonetic symbols are used. These can be found in the International Phonetic Alphabet.) In instances in which some Indo-European languages have an *a* sound, Indo-Iranian has *i* as a reflex of Indo-European sounds called laryngeals; *e.g.,* Greek *patēr* "father," Sanskrit *pitṛ-*, Avestan and Old Persian *pitar-*. After stems ending in long or short *a, i,* or *u,* an *n* occurs sometimes before the genitive (possive) plural ending *ām* (Avestan *-ąm*); *e.g.,* Sanskrit *martyānām* "of mortals, men" (from *martya-*); Avestan *mašyānąm* (from *mašya-*), Old Persian *martiyānām.*

Linguistic features shared by Indo-Aryan and Iranian

In addition to several other similarities in their grammatical systems, Indo-Aryan and Iranian have vocabulary items in common—*e.g.,* such sacrificial terms as Sanskrit *yajña-*, Avestan *yasna-* "sacrifice"; Sanskrit *hotṛ-*, Avestan *zaotar-* "a certain priest"; and names of divinities and mythological persons, such as Sanskrit *mitra-*, Avestan *miθra-* "Mithra." Indeed, both the Iranians and the Indo-Aryans used the same word to refer to themselves as a people: Sanskrit *ārya-*, Avestan *airya-*, Old Persian *ariya-* "Aryan."

Indo-Aryan and Iranian also differ in many points. Among them, Indo-Aryan has an *i* sound representing an Indo-European laryngeal sound not only in initial syllables but generally also in interior syllables; *e.g.,* Sanskrit *duhitṛ-* "daughter" (*cf.* Greek *thugátēr*). In Iranian, however, the sound is lost in this position; *e.g.,* Avestan *dugədar-, duγdar-.* Similarly, the word for "deep" is Sanskrit *gabhīra-* (with *ī* for *i*), but Avestan *ǰafra-.* Iranian also lost the accompanying aspiration (a puff of breath, written as *h*) that is retained in certain Indo-Aryan consonants; *e.g.,* Sanskrit *dhā* "set, make," *bhṛ,* "bear," *gharma-* "warm," but Avestan and Old Persian *dā, bar,* and Avestan *garəma-.* Further, Iranian changed stops such as *p* before consonants and *r* and *v* to spirants such as *f:* Sanskrit *pra* "forth," Avestan *frā;* Old Persian *fra;* Sanskrit *putra-* "son," Avestan *puθra-,* Old Persian *puṣṣa-* (*ṣṣ* represents a sound that is also transliterated as ç). In addition, *h* replaced *s* in Iranian except before nonnasal stops (produced by releasing the breath through the mouth) and after *i, u, r, k; e.g.,* Avestan *hapta-* "seven," Sanskrit *sapta-;* Avestan *haurva-* "every, all, whole," Sanskrit *sarva-.* Iranian also has both *xš* and *š* sounds, resulting from different Indo-European *k* sounds followed by *s*-like sounds, but Indo-Aryan has only *kṣ; e.g.,* Avestan *xšayeiti* "has power, is capable," *šaēiti* "dwells," but Sanskrit *kṣayati, kṣeti.* Iranian was also relatively conservative in retaining diphthongs that were changed to simple vowels in Indo-Aryan.

Iranian differs from Indo-Aryan in grammatical features as well. The dative singular of *-a*-stems ends in *-āi* in Iranian; *e.g.,* Avestan *mašyāi,* Old Persian *cartanaiy* "to do" (an original dative singular form functioning as infinitive of the verb). In Sanskrit the ending is extended with *a—martyāy-a.* Avestan also retains the archaic pronoun forms *yūš, yūžəm* "you" (nominative plural); in Indo-Aryan the *-s-* was replaced by *y* (*yūyam*) on the model of the 1st person plural—*vayam* "we" (Avestan *vaēm,* Old Persian *vayam*). Finally, Iranian has a 3rd person pronoun *di* (accusative *dim*) that has no counterpart in Indo-Aryan but has one in Baltic.

The original location of the Indo-Iranian group was prob-

Table 4: Modern (New) Indo-Aryan Languages

key: B—Bangladesh, former East Pakistan; I—India; N—Nepal; P—Pakistan (former West Pakistan)

language group, language	where principally spoken*	estimated number of speakers, 1981 (000)†	comments
Eastern group			
Assamese	*Assam*, India	11,800 (I)	official language of Assam, India; also‡
Bengali	*Bangladesh; West Bengal, Tripura,* and Assam, India	85,300 (B); 58,500 (I); 90 (P); 30 (N)	official language of Bangladesh and of West Bengal, Tripura, and Manipur, India; also‡
Oriya	*Orissa,* India	24,700 (I); 23 (B)	official language of Orissa, India; also‡
Northwest group			
Punjabi	*Punjab*, Northwest Frontier Province and Karāchi, Pakistan; *Punjab*, Haryana, Delhi, and Ganganagar district of Rajasthān, India; *Jammu* portion of both Indian- and Pakistani-held portions of Jammu and Kashmir	52,800 (P)§; 17,100 (I); statistics not available for Pakistani-held portion of Jammu and Kashmir	official language of Punjab, Pakistan, and with Urdu, of Jammu section of Jammu and Kashmir
Lahnda	Punjab and Northwest Frontier Province, Pakistan	statistics not available (P)¶, 14 (I)	
Sindhi	*Sind* province and *Las Bela* and other eastern districts of Baluchistan province, Pakistan; Kutch district of Gujarāt, India	10,100 (P); 2,040 (I); mainly immigrants from area now in Pakistan	official language of Sind, Pakistan; also‡
Pahari (a group of languages)			
Eastern Pahari			
Nepal (major east Pahari language)	*Nepal, Sikkim,* Bhutan	7,630 (N); 1,600 (I); 74 (Sikkim 1961), statistics not available for Bhutan	official language of Nepal
Central Pahari			
Kumauni	Himalayan Uttar Pradesh, India	1,500 (I)	
Garhwali	Himalayan Uttar Pradesh, India	1,160 (I)	
Western Pahari (62 languages and dialects according to the Indian census of 1961)	*Himachal Pradesh,* adjoining district of Uttar Pradesh, India; Himalayan districts of Indian- and Pakistani-held Jammu and Kashmir	994 (I); statistics not available for Pakistani-held portion	Pahari is an official language of Himachal Pradesh, along with Hindi
Unclassified and unspecified	(same as Western Pahari, above)	1,500 (I)	
Dardic			
Dard (East Dardic)			
Kashmiri (major language of East Dardic)	*Vale of Kashmir* and adjoining districts to south and west in Indian- and Pakistani-held portions of Jammu and Kashmir	2,995 (I), of which 2,991 are speakers of Kashmiri; 74 (P); statistics not available for Pakistani-held portion of Jammu and Kashmir or for Afghanistan	Kashmiri, along with Urdu, is an official language of the Kashmiri-speaking area of Jammu and Kashmir; also‡
Other Dardic languages:			
Khowari (Central Dardic), Kafiri (West Dardic), and other minor languages	*Gilgit Agency* of Pakistani-held portion of Jammu and Kashmir; adjoining districts of Northwest Frontier Province, Pakistan; and adjoining portion of northeast Afghanistan	208 (P, partial enumeration); statistics not available for Pakistani-held portion of Jammu and Kashmir or for Afghanistan	position of Dardic is disputed; some account for its peculiarities by proposing that it left the Indo-Iranian branch after Indo-Aryan but before all the features particular to Iranian had evolved; others suggest that East and Central Dardic are definitely Indo-Aryan, but that they did not go through the middle Indo-Aryan stage represented in documents; Kafiri occupies a special portion
West and Southwest groups			
Gujarati	*Gujarāt*, Bombay district of Mahārāshtra, India	32,330 (I); 503 (P);	official language of Gujarāt, India; also‡
Marathi	*Mahārāshtra* and eight adjoining districts in three older states of India	51,800§ (I)	official language of Mahārāshtra, India; also‡
Konkani	*Goa*, coastal Mahārāshtra south of Bombay, and coastal Karnataka, India	1,700¶ (I)	
Sinhalese	*Sri Lanka* (Ceylon)	10,986 (census 1981 data of Sri Lanka for Sinhalese group)	official language of Sri Lanka (Ceylon)
Divehi (Maldivian)	*Maldive Islands*	approximately 150	official language of Maldive Islands

*Italic type indicates language is spoken by a majority or plurality of the population in the area; roman type indicates language is spoken by a minority of the population in the area. †Not shown when there are fewer than 10,000 reported speakers in a given country. "Reported" number of speakers is often far different from actual number of speakers. Indian (I) data include census returns for Indian-held portion of Jammu and Kashmir; Pakistani (P) data do not include data for Pakistani-held portion of Jammu and Kashmir; Pakistani data are incomplete for the tribal areas of the Northwest Frontier Province, west Dardic.

Original homeland of Indo-Iranian

ably to the north of modern Afghanistan, in the present-day southern U.S.S.R.—the area called Soviet Turkistan—where Iranian languages are still spoken. From there, some Iranians migrated to the south and west, the Indo-Aryans to the south and east. From geographical references in the earliest Indo-Aryan literary document, the Rigveda, it is clear that the earliest settlement of Indo-Aryans was in the northwest of the Indian subcontinent. Migration did not take place at once; there was doubtless a series of migrations. The date of entry of the Indo-Aryans into the subcontinent cannot be precisely determined, though the beginning of the 2nd millennium BC is plausible and generally accepted.

There is heated controversy concerning the precise linguistic position of the language of the Indo-Iranian family first attested in Middle Eastern cuneiform texts of *c.* 1450–1350 BC. Some borrowed words and proper names appearing in these Hittite-Hurrian documents have been interpreted as belonging either to Indo-Iranian, to an Indic subgroup of Indo-Iranian that had not yet fully split, or to Indo-Aryan proper. Complete scholarly agreement on this issue has not been reached.

The identification of the Harappan peoples of the Indus Valley, whose writing has not yet been satisfactorily deciphered, also awaits further research; with it may come a possible answer as to whether Indo-Aryans encountered these people or whether their civilization had passed by the time the Indo-Aryans arrived on the subcontinent. Whatever the answers to these problems may be, the reasons for the split of the Indo-Aryans and Iranians are not known.

In the following presentation regarding Indo-Aryan documents as evidence for linguistic history, it should be borne in mind that almost all dates are approximations.

Table 4: Modern (New) Indo-Aryan Languages (continued)

language group, language	where principally spoken*	estimated number of speakers, 1981 (000)†	comments
Midland group			
Hindi	*Uttar Pradesh, Madhya Pradesh, Bihār, Haryana, Delhi,* Rājasthān, Punjab, Himachal Pradesh, and in scattered proximate districts of West Bengal and Mahārāshtra, India	173,000§ (I); 249 (B); 225 (N)	co-official language (with English) of the Republic of India and a lingua franca throughout North India; language of official business in area described; Khari Boli, based on a dialect of western Uttar Pradesh to the northeast of Delhi, is considered to be a standard form of Hindi; official language of Uttar Pradesh, Madhya Pradesh, Bihār, Haryana, Rājasthān, and Himachal Pradesh states and of Delhi union territory, India, also‡
Eastern Hindi (incomplete)			
Awadhi (Avadhi)	north central and central Uttar Pradesh, India	153 ‖ (I)	
Bagheli	north central Madhya Pradesh and south central Uttar Pradesh, India	421 ‖ (I)	
Chattisgarhi	east central Madhya Pradesh, India	7,495§ (I)	
Western Hindi (incomplete)			
Braj Bhasa	western Uttar Pradesh and adjacent districts of Haryana, Rajasthan, and Madhya Pradesh, India	32 ‖ (I)	
Bundeli (Bundelkhandi)	north central Madhya Pradesh and southwestern Uttar Pradesh, India	470 ‖ (I)¶	
Other Hindi languages and dialects (Eastern and Western, including Hindustani)	scattered over much of Uttar Pradesh, Madhya Pradesh, and Haryana and in eastern Rājasthān, India	10,066 ‖ (I)	
Urdu	*Karāchi* district and Pakistan in general; all but northeastern and southern peninsular India	35,000¶ (I); 6,700 (P), claimed as an additional language by 5,900 others (P); 240 (B); 54 (N)	official language of Pakistan (before 1971 co-official language of Pakistan; recognized in the constitution of India; a form of Urdu known as Dakhini Urdu (Southern Urdu) is still used in the area around Hyderābād; an official language in both the Indian- and Pakistani-held portions of Jammu and Kashmir; also‡
Bihari (a group of languages)			
Maithili	North Bihār, India; adjacent lowland Nepal	7,500 (I)¶; 1,560 (N)	
Magahi (Magadhi)	central Bihār, India	8,130 (I)	
Bhojpuri	western Bihār and eastern Uttar Pradesh, India	17,570 (I)¶; 1,120 (N)	
Others	Bihār, India	1,560 (I)¶	
Rajasthani (a group of languages			
Mewati	northeast Rājasthān, India	125 ‖ (I)	
Ahirwati	northeast Rājasthān	37 ‖ (I)	
Harauti	southeast Rājasthān	979 ‖ (I)	
Malvi	western Madhya Pradesh and southeast Rājasthān	1,190 ‖ (I)	
Nimadi	southwest Madhya Pradesh	920 ‖ (I) India	
Marwari	western, central, and northern Rājasthān	10,887¶ (I)	
Rajasthani-other and unclassified	over most of Rājasthān, with locally important groups in scattered districts of Mahārāshtra, Andhra Pradesh, and Karnataka, India; and both Indian- and Pakistani-held portions of Jammu and Kashmir	11,140¶ (I); 335 (P); statistics not available for Pakistani-held portion of Jammu and Kashmir	Rājāsthani, an official language of Rājasthān along with Hindi
Bhili (a group of dialects)	southern Rājasthān, western Madhya Pradesh, eastern Gujarāt, and northwest Mahārāshtra, India	1,562 ‖ (I)	
Khandeshi	northwest Mahārāshtra	750 (I)	
Others (inadequately classified or unspecified)			
Tharu	sub-Himalayan Nepal	603 (N)	
Miscellaneous dialects	sub-Himalayan Nepal	891–953 (N), depending on inclusiveness of listing	

languages, and Punjabi. As of 1972, Afghanistan has had no population census, but speakers of various Dardic languages there may number as many as 100,000, while another 10,000–20,000 may speak Indo-Aryan languages. ‡One of the 15 official languages listed in Schedule VIII of the Indian Constitution. §Presumed significant overstatement by census. Punjabi in Pakistan includes Lahnda. Marathi in India includes many speakers of Konkani and possibly also of Khandeshi. "Hindi" (undifferentiated) includes many speakers of all the languages of the Midland group footnoted ‖ or ¶. ‖ Presumed gross understatement by census. ¶Presumed significant understatement by census.

THE INDO-ARYAN LANGUAGES

Languages of the group. Indo-Aryan languages are assigned to three major periods: Old, Middle, and New Indo-Aryan. These periods are linguistic, not strictly chronological. Old Indo-Aryan includes different dialects and linguistic states referred to in common as Sanskrit. The most archaic Old Indo-Aryan is that of sacred texts called Vedas. Classical Sanskrit is the name given to the literary language that represents a polished form of various dialects. The late Vedic dialect described by the grammarian Pāṇini (*c.* 6th century BC) is also commonly called Classical Sanskrit. Middle Indo-Aryan includes both the dialects of inscriptions from the 3rd century BC to the 4th century AD and literary languages. Apabhraṃśa dialects represent the latest stage of Middle Indo-Aryan development. Though all Middle Indo-Aryan languages are included under the name Prākrit, it is customary to speak of the Prākrits as excluding Apabhraṃśa.

New Indo-Aryan is represented by such modern vernaculars as Hindi and Bengali, which began to emerge from about the 10th century AD. These too have earlier and later stages, culminating in the present-day languages.

New Indo-Aryan languages accounted for about 490,000,000 speakers in India, or approximately 74 percent of the population in the early 1980s. Considering the approximately 85,000,000 Bengali speakers in Bangladesh, approximately 63,000,000 speakers accounted for by Punjabi and Sindhi in Pakistan, and 11,000,000 Sinhalese (Sinhala) speakers in Sri Lanka (formerly Ceylon), the total number of New Indo-Aryan speakers is well over 650,000,000. According to the latest Indian census, there are 547 mother tongues of the Indo-Aryan group within the bounds of postpartition (1947) India. Some of these are dialects used by few speakers; others are official state languages having 30,000,000 or 50,000,000 speakers. The major groups of New Indo-Aryan languages are given in Table 4. Structurally and historically, Hindi and Urdu are one, though they are now official languages of different countries written in different alphabets. The term *hindī* (also *hindvī*) is known from as early as the 13th century. The term *zabān-e-urdū* "language of the imperial camp" came into use in about the 17th century. In the south, Urdu was used by Muslim conquerors of the 14th century.

Modern Indo-Aryan tongues and number of speakers

Many of the languages in Table 4 are official state languages, the media of education up to the university level and of official transactions. Hindi, written in the Devanāgarī script, is the co-official language (with English) of the Republic of India and is used as a lingua franca throughout North India. It has varieties according to the mother tongue of the area; *e.g.,* Bombay Hindi and Calcutta Hindi. Each of the major state languages has several other dialects in addition to the standard dialect adopted for official purposes. Including the various dialects down to the village level, it can be said that a chain of communication stretches across North India such that each dialect forms a link with each adjacent dialect. On the level of official languages this is not so: a Gujarati speaker will not readily understand colloquial Bengali.

Official state languages

Historical survey of the Indo-Aryan languages. The points noted above regarding Indo-Aryan migration make it difficult to determine the domain of Proto-Indo-Aryan, the ancestral language of all the known Indo-Aryan tongues, if indeed there was any such single region. All

Researched and compiled by Joseph E. Schwartzberg

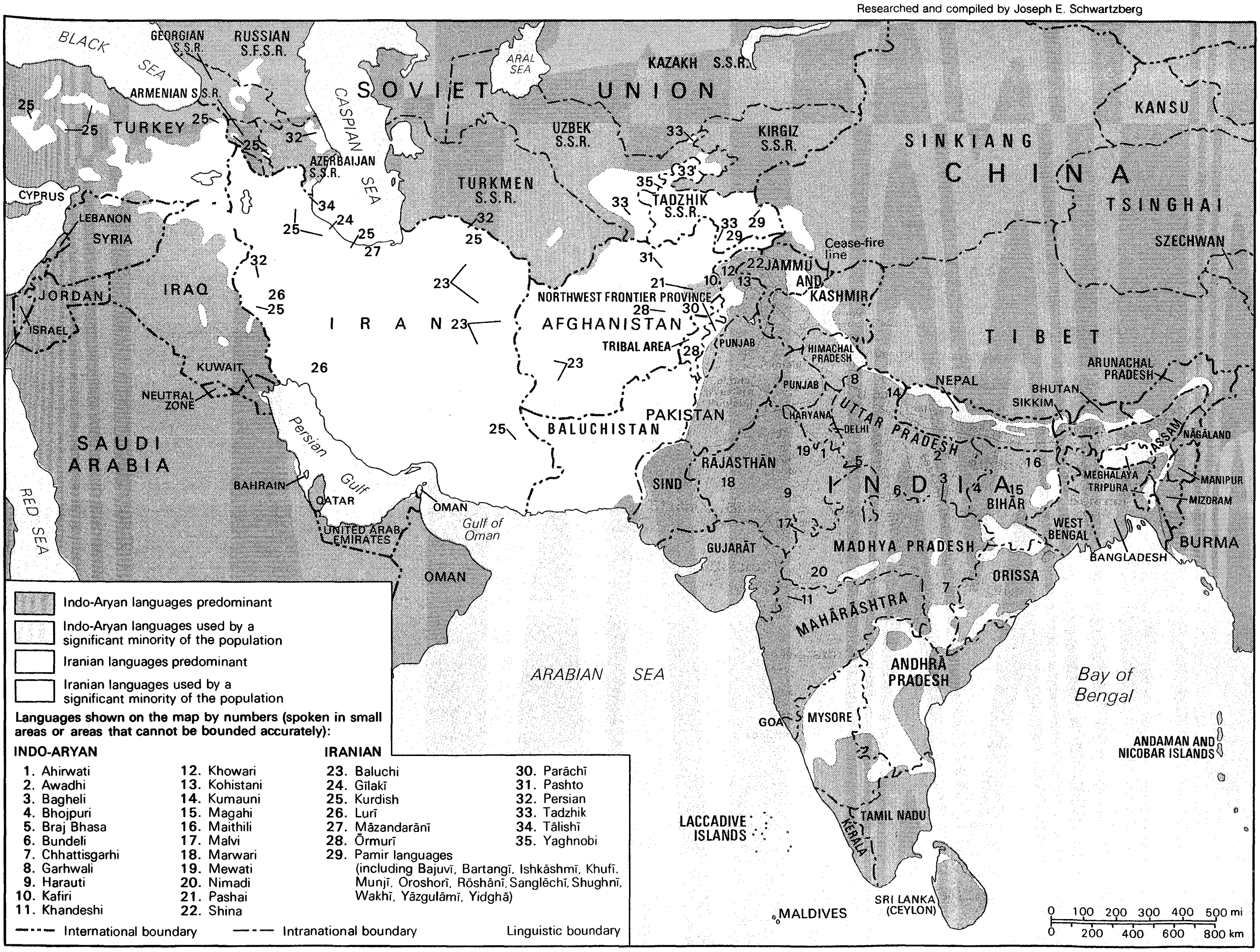


Figure 3: Distribution of the Indo-Iranian languages.

that can be said with certainty is that the Indo-Aryans on the subcontinent first occupied the area comprising most of present-day Punjab (both West and East), Haryana, and the Upper Doab (Ganges–Yamuna interfluve) of Uttar Pradesh. The structure of Proto-Indo-Aryan must have been close to that of early Vedic, with dialectal variations.

Vedic documents

Old Indo-Aryan. The most archaic Sanskrit is that of the Vedas, of which there are four major text groups called Saṃhitās: the Rigveda, Atharvaveda, Sāmaveda, and Yajurveda. The Yajurveda is in turn divided into two main branches, the White (Śukla) Yajurveda and the Black (Krishna) Yajurveda. The Rigveda, Atharvaveda, and Sāmaveda are purely metrical texts mainly used by priests in their ritual. The texts of the Black Yajurveda contain both verses used in ritual sacrifice (called *mantras*) and prose sections that are explanatory in nature, giving mythological explanations of sacrifices and objects used in them, together with etymologies (derivations of words). These sections are known as *Brāhmaṇa* portions. Each Veda also has a particular *Brāhmaṇa* connected with it. The early Vedic texts are pre-Buddhistic; a plausible date accepted for the composition of the Rigveda is between 1200 and 1000 BC, though the exact chronology of these early texts is difficult to establish. The prose passages of *Brāhmaṇas* and of the early *sūtra* (aphoristic texts) period may be called late Vedic. Also of the late Vedic period is the grammarian Pāṇini, author of a treatise called *Aṣṭādhyāyī,* who makes a distinction between the language of sacred texts (*chandas*) and the usual language of communication (*bhāṣā*).

Epic Sanskrit is so called because it is represented principally in the two epics, *Mahābhārata* and *Rāmāyaṇa.* In the latter the term *saṃskṛta* "formed, polished" is encountered, probably for the first time with reference to the language. The date of com-

Table 5: Sanskrit (Devanāgarī Alphabet and Numerals)

vowels and diphthongs			equivalents		approximate* pronunciation
initial	medial	name	EB preferred	alternatives	
अ		akāra	a		f*u*n
आ	ा	ākāra	ā		f*a*ther
इ	ि	ikāra	i		f*i*ll
ई	ी	īkāra	ī		mach*i*ne
उ	ु	ukāra	u		p*u*ll
ऊ	ू	ūkāra	ū		r*u*de
ऋ	ृ	ṛkāra	ṛ	ṛi, *ri*	litte*r*
ॠ	ॄ	ṝkāra	ṝ	ṛī, *ri*	†
ऌ	ॢ	ḷkāra	ḷ	lṛi, *li*	ab*le*
ए	े	ekāra	e	ē	f*a*de
ऐ	ै	aikāra	ai	āi	s*i*te
ओ	ो	okāra	o	ō	b*o*ne
औ	ौ	aukāra	au	āu	n*ow*

consonants and special signs	name	equivalents: EB preferred	alternatives	approximate* pronunciation
Gutturals‡				
क	kakāra	k		*k*in
ख	khakāra	kh		blo*ckh*ead
ग	gakāra	g		*g*o
घ	ghakāra	gh		lo*g h*ut
ङ	ṅakāra	ṅ	ñ	si*ng*
Palatals				
च	cakāra	c	ch, *k*	*ch*in
छ	chakāra	ch	chh, *kh*	pi*tch h*ook
ज	jakāra	j	*g*	*j*ob
झ	jhakāra	jh	*gh*	he*dgeh*og
ञ	ñakāra	ñ	n	ca*ny*on
Retroflexed∮				
ट	ṭakāra	ṭ	*t*	po*t*
ठ	ṭhakāra	ṭh	*th*	an*th*ill
ड	ḍakāra	ḍ	*d*	di*d*
ळ	ḷakāra	ḷ	*l*	‖
ढ	ḍhakāra	ḍh	*dh*	a*dh*ere
ण	ṇakāra	ṇ	*n*	ow*n*
Dentals¶				
त	takāra	t		li*tt*le
थ	thakāra	th		boa*t h*ouse
द	dakāra	d		*th*en
ध	dhakāra	dh		an*d h*e
न	nakāra	n		*no*
Labials♀				
प	pakāra	p		li*p*
फ	phakāra	ph		u*ph*ill
ब	bakāra	b		*b*a*b*y
भ	bhakāra	bh		a*bh*or
म	makāra	m		*maim*
Semi-vowels				
य	yakāra	y		*y*ard
र	repha	r		*r*a*r*e
ल	lakāra	l		*l*i*l*y
व	vakāra	v		*w*e
Spirants♁				
श	śakāra	ś	ç, *s*	*sh*y (palatalized)
ष	ṣakāra	ṣ	sh	*sh*y (retroflexed)
स	sakāra	s		*s*and
ह	hakāra	h		*h*at
Diacritics				
ः	visarga	ḥ		□
ं	anusvāra	ṃ	ṉ	◇
ँ	anunāsika	ṃ	ṁ	◇

Devanāgarī	Arabic	Devanāgarī	Arabic	Devanāgarī	Arabic
०	0	११	11	२२	22
१	1	१२	12	२३	23
२	2	१३	13	२४	24
३	3	१४	14	२५	25
४	4	१५	15	२६	26
५	5	१६	16	२७	27
६	6	१७	17	२८	28
७	7	१८	18	२९	29
८	8	१९	19	३०	30
९	9	२०	20	१००	100
१०	10	२१	21	१०००	1,000

*These pronunciations apply to Sanskrit. The same symbols sometimes have different values in the modern languages. †Same as *ṛkāra,* but lengthened. ‡Pronounced at back of throat. ∮Pronounced with the tongue curled back against the roof of the mouth. ‖ A retroflexed *l,* close to the second *l* in "little." ¶Pronounced with closed teeth. ♀Pronounced with the lips. ♁Breathed. □A diacritical mark indicating aspiration. ◇Diacritical marks indicating nasalization.

position for the core of early Epic Sanskrit is considered to be in the centuries just preceding the Christian era.

Classical Sanskrit is the language of the major poetic works (*kāvya*), drama (*nāṭaka*), tales such as the *Hitopadeśa* and *Pañca-tantra,* and technical treatises on grammar, philosophy, and ritual. It was used not only by the poet Kālidāsa and his predecessors Bhāsa, a dramatist, and Aśvaghoṣa, a Buddhist author, in the first centuries AD but was also continued long after Sanskrit was a commonly used mother tongue; indeed, Sanskrit is a language of learned treatises and commentaries to this day. It is also used as a lingua franca among *paṇḍit*s (Brahmin scholars) from different areas of India.

Development of Sanskrit

Linguistic developments can be traced from the early Vedic of the Rigveda through the later Saṃhitās on to the late Vedic of *Brāhmaṇa* prose and *sūtra*s, culminating in the language described by Pāṇini, which is tantamount to Classical Sanskrit. For example, the nominative plural form ending in *-āsas* (*devāsas* "gods") was already less frequent than *-ās* in the Rigveda and continued to lose ground later; in *Brāhmaṇa, -ās* (*e.g., devās*) is the normal form. There are numerous other changes evident. For example, the instrumental singular form of *-a*-stems ends both in *-ā* and *-ena* (a pronoun ending) in the Rigveda, with the latter form predominating; thus, *vīryā* "heroic might" appears once, and *vīryeṇa* occurs ten times (from *vīrya-* "heroic might, act"). In later Vedic *-ena* is the usual ending. All the early Vedic forms are expressly classed as belonging to the sacred language (*chandas*) by Pāṇini.

The verb also shows chronological differences. For example, the 1st person plural ending *-masi* (*e.g., bharāmasi* "we bear") predominates over *-mas* in Rigvedic but not in the Atharvaveda; *-mas* becomes the normal ending later. Early Vedic distinguishes between the aorist, imperfect, and perfect tenses. The aorist is commonly used to refer to an action that has recently taken place; the imperfect is a narrative tense referring to actions accomplished in the distant past. The perfect form of the verb originally denoted, as in Greek, a state reached; *e.g., bi-bhāy-a* "is afraid" (root *bhī*). From earliest Vedic, however, this was not always the use of the perfect. Although the grammarian Pāṇini distinguished between the three tenses noted (he said the perfect is used to denote an action beyond one's ken), the perfect and imperfect both came to be used as narrative tenses.

There are also future forms of Vedic, formed with suffixes (*-iṣya* and *-sya*) and used from earliest times. A future form, composed of an agent noun of the type *kartṛ-* "doer" and followed, except in the 3rd person, by forms of the verb *as* "be" (*e.g., kartāsmi* [*kartā asmi*] "I will do"), was recognized as in common use by Pāṇini but is rare in early Vedic.

The injunctive of early Vedic

Early Vedic had a category that went out of use by the late Vedic period of *Brāhmaṇas*—the injunctive, which was formally a form with secondary endings lacking the augment, a prefixed vowel. The injunctive could be used to denote a general truth. A general truth can also be signified by the subjunctive, which is characterized by the vowel *a* affixed to the present, aorist, or perfect stem. Later Vedic retained the injunctive only in negative commands of the type *mā vadhīs* "do not slay." The subjunctive also diminished slowly until it was no longer used; for Pāṇini the subjunctive belonged to sacred literature. The functions of the subjunctive were taken over by the form called optative (and the future form).

Noun forms incorporated into the verb system are numerous in early Vedic. Rigvedic has forms with affixes *ya* and *tva* functioning as future passive participles (gerundives); *e.g., vāc-ya-* "to be said," *kar-tva-* "to be performed, done." The Atharvaveda has, additionally, forms with *-(i)tavya* (*hiṃs-itavya-* "to be injured") and *-anīya* (*upa-jīv-anīya-* "to be subsisted upon"). By late Vedic, the type with *tva* had been eliminated; Pāṇini recognized as normal the types *kārya-, kartavya-, karaṇīya-* "to be done." In Indo-Aryan, from earliest Vedic down to New Indo-Aryan, forms called absolutives (or gerunds) are used to denote the previous of two or more actions performed (usually) by one agent: "having done . . . he did"; for example, *pibā niṣadya* "sit down (*niṣadya* "having sat down") and drink." Rigvedic uses *tvī, tvā, tvāya,* (*t*)*ya* to form absolutives, but these were later reduced to two: *tvā* with a simple verb or one compounded with the negative particle, and *ya* with a verb compounded with a preverb (a preposition-like form).

Early Vedic also uses various case forms of action nouns in the capacity of infinitives; *e.g.,* dative singular *-tave* (*dā-tave* "to give"), genitive singular *-tos* (*dā-tos*), both from a noun in *-tu,* which also supplies the accusative ending *-tum* (*dā-tum*). There are other types in early Vedic, but the nouns in *-tu* are important; in late Vedic the accusative *-tum* and the genitive *-tos* (construed with *īś* or *śak* "be able, can") became the norm. According to Pāṇini, forms in *-tum* and dative singular forms of action nouns are equivalent variants: *bhok-tuṃ gacchati/ bhojanāya gacchati* "He is going out to eat."

Chronological and dialectical modifications

That some forms fell into disuse in the course of Indo-Aryan is natural; the above represent both chronological and dialectal modifications. Such change was recognized by Indian grammarians; *e.g.,* Patañjali, of the mid-2nd century BC, noted that perfect forms of the type *ca-kr-a* "you did, have done" (2nd person plural) were not in use at his time; instead, a nominal (adjective) form *kṛ-ta-vant-as* was used, consisting of the past passive participle *kṛ-ta-* and an adjectival suffix *-vant.* Indian grammarians also recognized the existence of different dialects. Pāṇini noted forms used by northerners (*udīcya*) and easterners (*prācya*), as well as various dialectal uses described by grammarians who preceded him.

Earlier documents also afford evidence for dialect variation; *e.g.,* the early Vedic of the Rigveda is a dialect in which the Indo-European *l* sound was for the most part replaced by *r*—*prā* "fill," *pūr-ṇa-* "full." This change accords with Iranian; *e.g.,* Avestan *pərəna* "full." These forms contrast with Latin *plenus* and Gothic *fulls,* with *l.* Other dialects kept *l* and *r* distinct. There are also doublets that have both *r* and *l* in words with Indo-European *r: rohita-/lohita-* "red." The variant with *l* can be assumed to belong to an eastern dialect. This variance accords with Middle Indo-Aryan evidence and the fact that such *l* forms become more numerous in the tenth book (*maṇḍala*) of the Rigveda, which is demonstrably more recent than the most ancient parts of the Rigveda and dates from a time when the Indo-Aryans had progressed farther east than their original location on the subcontinent. The development of retroflex *ḷ-* and *ḷh-* sounds (produced by curling the tip of the tongue upward toward the hard palate) from the retroflex sounds of *ḍ* (*nīḷa-* "nest" from *nīḍa-*) and *ḍh* when occurring between vowels is another feature characteristic of some dialects, including the major dialect of the Rigveda.

Classical Sanskrit and its accentual system

Classical Sanskrit represents a development of one or more such early Old Indo-Aryan dialects. At this state, the archaisms noted above have been eliminated. Moreover, the accentual system of Classical Sanskrit is not the same as that of Vedic, which had a system of pitches; vowels had low, high, or circumflex (first rising, then falling) pitch, and the particular vowel of a word that received high pitch could not be predicted. In Classical Sanskrit, on the other hand, the accent was probably predictable. If the next to the last vowel was long, it received the accent; if not, the vowel preceding it was accented. The Vedic system survived at least to the time of Pāṇini, who described it fully and did not restrict it to sacred language.

For all this simplification, Classical Sanskrit is considerably more complex than Middle Indo-Aryan. In addition to the vowels *a, i,* and *u* (in both long and short varieties), it has *ṛ* and *ḷ* used as vowels. Consonant clusters occur freely, except in word final position, and the system of sound modification conditioned by the context, called *sandhi,* is fully operative. Moreover, in its grammatical system Classical Sanskrit maintains the dual number, seven cases in addition to the vocative form (which marks the one addressed), and a complex set of alternations. For example, to the nominative singular form *agni-s* "fire," correspond the genitive singular *agne-s* "of fire" the nominative plural *agnay-as* "fires," and the instrumental plural *agni-bhis* "with fires," with differing vowels in the second syllable. There are also separate sets of nominal (noun) and pronominal (pronoun) endings. Some nouns and adjectives inflect as pronouns; *e.g., ekasmai,* dative singular masculine-neuter of *eka-* "one."

The verb system of Classical Sanskrit also maintains complex alternations. In the present tense of the type *bhav-a-ti* "becomes, is," the stem (*bhav-a-*) remains unchanged throughout the paradigm except for lengthening of the *-a-* to *-ā-* before *v* and *m.* But other verbs have vowel alternation; *e.g., as-mi* "I am," *s-mas* "we are"; *e-mi* "I go," *i-mas* "we go"; *juhomi* "I pour," *juhumas* "we pour." A distinction is observed between active and mediopassive endings: *jan-ay-a-ti* "engenders" with the active ending *-ti,* but *jā-ya-te* "is born" with the mediopassive ending *-te.* (Mediopassive verb forms are used for the passive, reflexive, and other meanings.)

Classical Sanskrit also has a rich system of nominal and verbal derivatives. Compound words are of the following kinds: copulative (*dvandva*) compounds such as *mātāpitarau* "mother and father" (also elliptic *pitarau* "parents"); the type like *tat-puruṣa-* "his man," in which the first member is equivalent to a case other than nominative; the type like *bahu-vrīhi* "much-rice," in which the object denoted is other than that of any of the members of the compound (*bahur vrīhir yasya* "He who has much rice"); and adverbial compounds (*avyayībhāva*) of the type *upāgni* (*upa-agni*) "near the fire." In addition, there are derivatives with affixes *-tara-* and *-tama,* such as *priya-tara-* "very dear" and *priya-tama-* "most dear" from the adjective *priya-.* Pronouns have derivatives equivalent to case forms; *e.g., tatra* "there," *yatra* "where," and *kutra* "where?" are equivalent to locative forms such as *tasmin, yasmin,* and *kasmin.* These can also be used without a noun.

Among the derivative verbal systems are the causative and the desiderative ("desire to"); the former has an affix *-ay-* (*gam-ay-a-ti* "makes to go," *kār-ay-a-ti* "has do") or, after roots in *-a, -pay-* (*sthā-pay-a-ti* "sets in place"). The desiderative is formed with *-sa-* and reduplication (repetition of a part of the root)—*dī-dṛk-ṣa-te* "desires to see" (root *dṛś*). The desiderative also has an agent noun in *-u*—*dī-dṛk-ṣ-u* "who wishes to see."

The Prākrits

Middle Indo-Aryan. The Sanskrit word *prākṛta,* whence the term Prākrit, is a derivative from *prakṛti-* "original, nature." Grammarians of the Prākrits generally consider the original from which they derive to be the Sanskrit language as described by grammarians going back to Pāṇini. Most modern scholars consider *prākṛta* to refer to the "natural" languages, the vernaculars, as opposed to Sanskrit, the polished language of literature and the educated (*śiṣṭa*). There is also linguistic evidence to support this view. Several forms in the Prākrits are found in Vedic but not in Classical Sanskrit. As Classical Sanskrit is not directly derivable from any single Vedic dialect, so the Prākrits cannot be said to derive directly from Classical Sanskrit.

The most archaic literary Prākrit is Pāli, the language of the Buddhist canon (*c.* 5th century BC) and of the later stories and commentaries of Theravāda Buddhism. Pāli represents essentially a western Middle Indo-Aryan dialect, though there are sufficient easternisms in the canon to have led some scholars to the view that the canon as it exists today is a recast of an original in an eastern dialect. To the Buddhist literature also belongs the *Gāndhārī Dhammapada,* the only literary text written in a dialect of the northwest. The Niya documents, official documents written in Prākrit dating from the 3rd century AD, also belong to the northwest. The earliest inscriptional Middle Indo-Aryan is that of the Aśokan inscriptions (3rd century BC). These are more or less full translations from original edicts issued in the language of the east (from the capital Pāṭaliputra in Magadha, modern Patna in Bihār) into the languages of the areas of Aśoka's kingdom. There are other Prākrit inscriptions up to the 4th century AD, and Sanskrit was not used inscriptionally until the first centuries AD. Literary Prākrits other than Pāli were also used in independent works and in dramas along with Sanskrit.

According to Prākrit grammarians, Mahārāṣṭrī ("From the Mahārāshtra Country") is the Prākrit par excellence. It is the language of *kāvya*s (epic poems) such as the *Rāvaṇavaha* (also called *Setubandha*) from no later than the 6th century AD. Mahārāṣṭrī is also the language of lyrics in Rājaśekhara's *Karpūra-mañjarī* (*c.* 900), the only extant drama written completely in Prākrit, and of verses recited by women in the classical drama of Kālidāsa and his successors, though not earlier. The literary dialect used for conversation among higher personages other than the king and his captains in the drama is Śaurasenī, while Māgadhī is used by lower personages.

The language of the early Jaina canon, the final version of which was made in the 5th or 6th century AD, is called Ardhamāgadhī ("Half Māgadhī"); Jaina also used another literary dialect, called Jaina Māhārāṣṭrī in non-canonical works. The oldest poetic work in this is Vimala Sūri's *Paumacariya* (*c.* 3rd century). Of other Prākrit dialects mentioned by grammarians, Paiśācī (or Bhūta-Bhāṣā, both meaning "Language of Demons") is noteworthy; it is said to be the language of the original *Bṛhatkathā* of Guṇāḍhya, source of the Sanskrit book of stories *Kathāsaritsāgara.*

Buddhist Hybrid Sanskrit

Buddhist works were also written using a language that has been called Buddhist Hybrid Sanskrit. Among these works is the *Mahāvastu,* the core of which is thought to date from the 2nd century BC. This language is a Middle Indo-Aryan dialect of indeterminate origin, which steadily became more Sanskritized in prose sections of later works.

The most advanced stage of Middle Indo-Aryan, Apabhraṃśa, was also used as a literary language. That there was literary creation in Apabhraṃśa by the 6th century is clear from an inscription of King Dharasena II of Valabhī, in which the King praises his father as being adept in Sanskrit, Prākrit, and Apabhraṃśa composition. Moreover, in the fourth act of Kālidāsa's drama *Vikramorvaśīya* there are Apabhraṃśa verses. Because Kālidāsa probably lived in the 3rd or 4th century, literary composition in Apabhraṃśa is earlier still, if these verses are legitimate. There is a great deal of later literature in Apabhraṃśa, for the most part Jaina works; *e.g., Paumacariu* of Svayambhū (8th–9th century), *Harivaṃśa-purāṇa* of Puṣpadanta (10th century), *Sanatkumāra-cariu* of Haribhadra (12th century).

Middle Indo-Aryan is characterized generally by the reduction of the complexities seen in Old Indo-Aryan. The vowel system was reduced by the merger of *ṛ* (and *ḷ*) sounds with vowels and the change of the diphthongs *ai* and *au* to the vowel sounds *e* and *o; e.g.,* Pāli *accha-* "bear" (Sanskrit *ṛkṣa-*), *iṇa-* "debt" (Sanskrit *ṛṇa-*), *uju-* "straight" (Sanskrit *ṛju-*), *pucchati* "asks" (Sanskrit *pṛcchati*), *metti-* "friendship" (Sanskrit *maitrī-*), *orasa-* "breast-born, legitimate" (Sanskrit *aurasa-*). Moreover, *-aya-* and *-ava-* commonly contracted to *-e-* and *-o-; e.g.,* Pāli *jeti* "conquers" (Sanskrit *jayati*), *odhi-* "limit" (Sanskrit *avadhi-*). Final consonants were deleted, with the exception of *-m,* which developed to an *-ṃ* sound before which a vowel was shortened (Pāli *bhāriyaṃ* "wife"; Sanskrit *bhāryām*). Together with the trend toward replacing variable consonant stems by unchanging stems in *-a-,* this change had serious consequences for the grammar. Consonant stems steadily disappeared and were transformed to stems ending in a vowel; *e.g.,* to Sanskrit *śarad-* "autumn," *sarit-* "stream," and *sarpis-* "butter" correspond the Pāli forms *sarada-, saritā,* and *sappi-.* Consonant clusters were also modified in Middle Indo-Aryan; *e.g.,* Pāli *khetta-* "field" (corresponding to Sanskrit *kṣetra-*), Pāli *dakkhiṇa-* "right, south" (Sanskrit *dakṣiṇa*), *aggi-* "fire" (Sanskrit *agni-*), *puṇṇa-* "full" (Sanskrit *pūrṇa*), and *taṇhā-* "thrist" (Sanskrit *tṛṣṇā-*). The shortening of vowels before modified consonant clusters led to the use of short *ĕ* and *ŏ* sounds, which were unknown in Old Indo-Aryan; *e.g.,* Pāli *sĕmha-* "phlegm" (Sanskrit *śleṣman*), *ŏṭṭha-* "lip" (Sanskrit *oṣṭha-*).

Characteristics of literary Māhā-rāṣṭrī

The above phenomena are not restricted to Pāli; they are pan-Middle Indo-Aryan. Differences between Pāli and Aśokan and other Prākrits include the retention of voiceless stops (*i.e., p, t, k*) between vowels in Pāli and Aśokan dialects; other Middle Indo-Aryan dialects modify them. The extreme development appears in literary Māhārāṣṭrī, in which unaspirated stops (pronounced without an accompanying audible release, or pull of breath) other than retroflexes (*ṭ, ḍ*) and labials (*p, b*) were deleted, aspirated stops (pronounced with an audible puff of breath) were replaced by *h,* retroflexes (pronounced by curling the tongue upward toward the hard palate) became voiced, and labials were replaced by *v; e.g., loa-* "world" (Sanskrit *loka-*), *loaṇa-* "eye" (Sanskrit *locana-*), *sāhā-* "branch" (Sanskrit *śākhā-*), *paḍhai* "recites, reads" (Sanskrit *paṭhati*), and *savaha-* "curse" (Sanskrit *śapatha-*).

Essentially on the same level are the dialects of Jaina texts, but in these a *y* glide prescribed by grammarians occurs when a consonant is elided: *vayaṇa-* "face" (Sanskrit *vadana-*); *sayala-* "whole" (Sanskrit *sakala-*). In Śaurasenī, on the other hand, voiceless stops (*e.g., p, t, k*) between vowels are voiced (*e.g.,* become *b, d, g,* respectively); *e.g., ido* "hence" (Sanskrit *itaḥ*); *tadhā* "thus" (Sanskrit *tathā*). Though Pāli and Aśokan are at an earlier level of development with respect to these changes, they share with the rest of the Middle Indo-Aryan dialects the replacement of voiced aspirated sounds between vowels by *h: lahu-* "light, unimportant" from *laghu-; dahati* "gives" (Sanskrit *dadhāti*). Similarly, they share the change of *dy-* to *j: joti-* "light, brilliance" (Pāli *jotati* "shines," Sanskrit *dyotate*). Pāli and Aśokan, however, retain a *y* sound, changed to *j* in most other Prākrits; *e.g.,* the pronoun *ya-* (feminine *yā-*), as in Sanskrit, opposed to *ja-*.

The deletion of stop consonants noted above resulted in vowel sequences within words that were unknown to Old Indo-Aryan. Similarly, the extent of *sandhi* modification was restricted in Middle Indo-Aryan. The Middle Indo-Aryan vowels *ī* and *ū* do not change to *y* and *v* before dissimilar vowels in compounds; *e.g.,* Māhārāṣṭrī *rattīandhaa-* "dark of night" (Sanskrit *rātry-andhaka-*). In addition, the first of two contiguous vowels in different words is subject to deletion; *e.g.,* Pāli *manas'icchasi* (from *manasā icchasi*) "you wish in your mind."

In its grammatical system, Middle Indo-Aryan also reduced complexities. The dual number no longer exists as a separate category; for Sanskrit *dvābhyām* "by two," Middle Indo-Aryan has *dohi*(*ṃ*), with the ending *-hi*(*ṃ*) equivalent to the instrumental plural *-bhis* of Old Indo-Aryan. Among other changes is the replacement of the dative case by the genitive except in particular usages; *e.g.,* the use of forms corresponding to the Old Indo-Aryan dative to denote a purpose.

In Middle Indo-Aryan, nominal and pronominal forms are no longer strictly segregated; *e.g.,* Aśokan *vijitamhi* "in the kingdom" (also *vijite*) has a pronominal ending equivalent to Sanskrit *-smin.*

Middle Indo-Aryan verb system

In the verb system, the contrast between active (*-ti*) and mediopassive (*-te*) endings was obliterated. Further, the Old Indo-Aryan distinction between aorist, imperfect, and perfect forms was eliminated. With few exceptions, the sigmatic aorist (an aorist form with *s*) provides the only productive preterite of early Middle Indo-Aryan: Aśokan *ni-kkhamisu* "they set out" (Sanskrit *nir-a-kramiṣur*). In later Prākrits verbally inflected preterites were generally eliminated; in their place was used the past participle. For example, in Śaurasenī *devi uva-visa, mahārāo vi ā-ado* "Sit down, my queen, the king also has arrived," the past participle *ā-ado* (Sanskrit *ā-gataḥ*) agrees with *mahārāo* "king" (Sanskrit *mahā-rājaḥ*) in number and gender. If the verb is transitive, the participle agrees with the direct object, and the agent is denoted by an instrumental form: in Jaina Māhārāṣṭrī, *teṇa vi savvaṃ siṭṭhaṃ* "He has told everything," *teṇa* "by him" denotes the agent, and *siṭṭhaṃ* "told" (Sanskrit *śiṣṭam*) agrees with the neuter singular form *savvaṃ* (Sanskrit *sarvam*). When no object is denoted, the verb is in the neuter singular. Old Indo-Aryan used both the participial construction and the finite verb; thus to Prākrit *so vi teṇa samaṃ gao* "He also went with him" could correspond Sanskrit *so'pi tena saha gataḥ* or *so'pi tena sahāgamat* (*saha agamat*). The Middle Indo-Aryan development eliminated the latter.

Alternations of the Sanskrit type *as-mi, s-mas* were eliminated in Middle Indo-Aryan; the predominant type of present tense was formed from an unchanging vowel stem (Pāli *e-ti, e-nti* "go[es]").

Nominal forms of the verb system are of the same types as Old Indo-Aryan; *e.g.,* the Pāli future passive participle *kātabba-* (Sanskrit *kartavya-*) "to be done," Śaurasenī *karaṇia;* Ardhamāgadhī, Jaina Māhārāṣṭrī, and Māhārāṣṭrī *karaṇijja-* "to be done." The infinitive is commonly formed on the present tense stem, not on the root, as in Old Indo-Aryan. Thus Pāli *pappotum* is formed on the present *pappoti;* Sanskrit *prāptum* is formed on the root *prāp,* present tense *prāpnoti.*

Middle Indo-Aryan shows evidence of dialectal differentiation. The earliest documents that allow one to determine roughly the dialect distribution are Aśoka's inscriptions. These represent three major dialect areas: east, as in the inscriptions of Jaugaḍa, Dhauli, and Kālsī; west, in Girnār; and northwest, in Mānsehrā and Shāhbāzgaṛhī. Characteristic of the east dialect area is final *-e,* corresponding to *-o* in the west and *-as* in Sanskrit; in the east dialect area *l* also regularly corresponds to *r* of the west and of Sanskrit. Moreover, in the east dialect area there is a tendency to insert a vowel within consonant clusters, while in the west and northwest one of the consonants is assimilated to the other without an intervening vowel. For example, to Sanskrit *rājñas* "of the king" corresponds Girnār *rañño,* Shāhbāzgaṛhī *raño,* Jaugaḍa *lājine.* Northwest stands apart in retaining three spirant sounds, *ś, ṣ, s,* which merge to *s* elsewhere. Aśoka's eastern dialect, from the Magadha country, shows an *s* sound for Old Indo-Aryan *ś, ṣ, s,* rather than the *ś* sound typical of literary Māgadhī. Grammatical features also show dialectal variation; *e.g.,* the Aśokan dative singular form is *-āya* in the western dialects (Girnār *atthāya* "for the purpose of") but *-āye* in the east (Kālsī, Dhauli *aṭṭhāye*).

Apabhraṃśa

As noted above, the most advanced development of Middle Indo-Aryan is seen in Apabhraṃśa. Sound changes that are typical of Apabhraṃśa include the replacement of the vowel sound *a* by *u* in final syllables; *e.g., karahu* "you do, make," corresponding to *karaha* (*karadha*) in other Prākrits. From stems in *-aya-* develop forms in *-au* and nasalized *-aũ* (nasalization is here indicated by a tilde): *bhaḍārau* "honored one, king" (Prākrit *bhaṭṭārayo*), *haũ* "I" (Aśokan *hakaṃ*). Nasalization also appears in environments in which earlier *m* occurred between vowels; *e.g., gāũ* "village" (from *gāma,* Sanskrit *grāma*). Numerous other sound changes are evident, among them the development of *-s*(*s*)- between vowels into *h: tahŏ* "of him" (from Prākrit *tassa,* Sanskrit *tasya*); *hohinti* "will be" (compare Pāli *hossati*). Apabhraṃśa contractions, such as *-aya-* changing to *-a* and *-iya* to *-ī,* foreshadow New Indo-Aryan, in which the development was extended; *e.g.,* Apabhraṃśa *pāṇiu* "water" (Old Indo-Aryan *pāniyam*), Gujarati *pāṇī,* Hindi *pānī.*

In other points Apabhraṃśa also presaged New Indo-Aryan. The interest of Apabhraṃśa lies in the fact that contracted forms presage the New Indo-Aryan opposition of masculine, neuter, and feminine nouns; thus, Apabhraṃśa *-au, -aũ, -ī,* Gujarati *-o, -ũ, -ī* (*gayo, gayũ, gaī* "went"), Hindi *-ā, -ī* (*gayā, gaī*). The case system of Apabhraṃśa is also at a more advanced level of disintegration than that of earlier Middle Indo-Aryan, with the instrumental and locative plurals being identical in form (*-ahĩ* or *-ehĩ* for *-a*-stems) and instrumental singular forms also being used as locatives.

In the Apabhraṃśa verb system, present tense stems in *-a* predominate. Apabhraṃśa verb endings differ from those of other Prākrits. Most interesting is the 3rd person plural type *kara-hĩ* "they do," which coexists with *karanti.* The form *kara-hĩ,* corresponding to the 3rd person singular *kara-i* "he does," is formed on the model of the pair *kara-ũ* (1st person singular, "I do") and *kara-hũ* (1st person plural, "we do"). Here again Apabhraṃśa comes close to New Indo-Aryan. Moreover, Apabhraṃśa has some causative formations that do not occur elsewhere in Middle Indo-Aryan but are known from New Indo-Aryan—*bham-āḍa-i* "causes to turn," Gujarati *bhamāṛe che* "causes to turn round," and *pais-āra-i* "causes to enter," Gujarati *pɛsāre che* "causes to enter, to penetrate."

Apabhraṃśa syntactic patterns

Also noteworthy are two syntactic usages that closely parallel those present in New Indo-Aryan. The present participle is used as a conditional; *e.g., jai haũ mi teṇa sahũ tau karantu to kiṃ asamāhie sahũ marantu* "Even if I had performed (*karantu*) ascetic acts with him, would I have died without mental concentration?" in which the participles *karantu* and *marantu* have the value of conditionals. In Sanskrit the conditionals *a-kar-iṣya-m* and *a-mar-iṣya-m* are used; but in speaking Gujarati a person would say *jo hũ . . . karat . . . to marat,* and Hindi would have the forms *kartā . . . martā.* The Apabhraṃśa gerundive in *-iv*(*v*)*a* or *-ev*(*v*)*a* can be used as an infinitive; *e.g., pi-evae laggā* "began to drink." This is the Gujarati construction *pi-vā lāgyo* "began to drink," in which *pi-vā* is an inflected form of *pi-vũ,* that is, a verbal noun (infinitive) corresponding etymologically to the Apabhraṃśa gerundive.

Influences on Middle Indo-Aryan. In the mid-2nd century BC, the grammarian Patañjali explained that to speak faultlessly the language now called Sanskrit (as described by Pāṇini) one should imitate the correct speakers (called *śiṣṭa* "learned, educated") of Āryāvarta ("Country of the Aryans"). Earlier, the grammarian Kātyāyana (*c.* 3rd–4th century BC) had noted that Pāṇini gave lists of verb roots in order that certain Middle Indo-Aryan forms not be accepted as having been correctly derived from a Sanskrit verb root. Moreover, Patañjali noted that one should study grammar in order to learn not to use incorrect words such as *helayaḥ* instead of *herayaḥ* (a phrase used in calling to people) or *gāvī* instead of *gauḥ* "cow"; *gāvī* is a Middle Indo-Aryan word. The observations of these grammarians are considered to lend support to the view that by the 6th or 5th century BC Sanskrit as a medium of learned conversation coexisted with Middle Indo-Aryan. Further, the Pāli canon records that the Buddha enjoined his followers to use the vernaculars in communicating his teachings, and the Jaina canon identifies Ardhamāgadhī as the language to be employed for communicating the teachings of Mahāvīra. Similarly, Aśoka used Middle Indo-Aryan, not Sanskrit, in the inscriptions he ordered written throughout his kingdom; Sanskrit does not appear on inscriptions until the early centuries AD (*e.g.,* Rudravarman's inscription at Junagarh, *c.* AD 150). The coexistence of Old Indo-Aryan and Middle Indo-Aryan is to be accepted even for the time when the earliest Old Indo-Aryan texts were put to writing.

Middle Indo-Aryan shows similar evidence of the influence of linguistically more advanced vernaculars on literary compositions. The Prākrits of elegant literary compositions must have been artificial, different in many respects from the vernaculars current at the time, though reflecting languages that were current at some former time. The Old Indo-Aryan and Middle Indo-Aryan stages, then, present a picture of concurrent vernaculars with dialects and literary languages influenced by the vernaculars; it is impossible to compartmentalize the different stages as beginning and ending at any definite date.

Sources of borrowing into Indo-Aryan

The literary languages borrowed words and suffixes from earlier languages. There are Prākritisms (*i.e.,* forms of earlier Prākrits) in Apabhraṃśa; *e.g.,* the genitive singular ending *-ssa* instead of *-hŏ* and 2nd person plural verb forms terminating in *-ha* instead of *-hu.* All the literary Prākrits had recourse to Sanskrit as a source for borrowing words. Words that were incorporated into the Prākrits from Sanskrit with no change in form are called *saṃskṛta-sama* "identical with Sanskrit" (or *tat-sama* "identical with that") and are contrasted with words termed *saṃskṛta-bhava (tad-bhava)* "whose origin is in Sanskrit"—that is, words that the grammarians can derive from Sanskrit by using certain rules. Another class of words, called *deśya* (or *deśī*) "belonging to the area, country," includes items that the grammarians cannot derive easily from Sanskrit and that are supposed to have been in use in particular areas from early times.

Many or most of the *deśya* words are indeed derivable from Sanskrit, but some are of Dravidian origin; *e.g., akka* "sister" (Telugu *akka*), *attā* "father's sister" (Telugu *atta*), *appa* "father" (Telugu *appa*), *ūra* "village" (Telugu *ūru*), *pulli* "tiger" (Telugu *puli*). Borrowing from Dravidian occurred also at earlier times; the Dravidians originally occupied territory much farther north than they did in Middle Indo-Aryan times. The Ṛgveda has such words as *kuṇḍa* "pitcher, pot," which is doubtless of Dravidian origin (Tamil *kuṭam* "pot"). Such borrowings become more numerous in later Sanskrit. It is not always certain that borrowing proceeded from Dravidian to Indo-Aryan, however, because Dravidian languages freely borrowed from Indo-Aryan. Thus, some scholars claim that Sanskrit *kaṭu* "sharp, pungent" is from Dravidian, but others claim that it is a Middle Indo-Aryan form deriving from an earlier **kṛt-u* "cutting" (root *kṛt*). (An asterisk [*] preceding a form indicates that it is not attested but has been reconstructed as a hypothetical form.) Whatever the judgment on any individual word, it is clear that Indo-Aryan did borrow from Dravidian, and this phenomenon is important in considering a group of sounds that sets Indo-Aryan apart from the rest of Indo-European—the retroflexes. Without doubt the influence of Dravidian is to be considered as contributing to the extension of these sounds beyond their limited occurrence in inherited Indo-European items such as *nīḍa* "nest" (from **ni-sd-o*), *iṣ-ṭa* "desired" (from **is-to*), and *stīr-ṇa* "spread out" (from **stṝ-no*). The Munda languages (or, more generally, the Austro-Asiatic languages) are also a source of some borrowing into Indo-Aryan; *e.g.,* Sanskrit *jambāla* "mud" (Santali *jo̱bo̱*).

In the 8th century AD, the philosopher Kumārila mentioned not only Dravidian but also Persian and Greek as sources of foreign words. Such borrowing can be traced back to early times. In the 6th century BC Darius counted Gandhāra as a province of his kingdom, and Alexander the Great penetrated into northern India in the 4th century BC. From Iranian come words such as that meaning "inscription, writing, script"; in the northwest inscriptions of Aśoka the word is *dipi* (Old Persian *dipi*) and Sanskrit has *lipi,* the form in other Aśokan versions and in Pāli. Also from Persian is Sanskrit *kṣatrapa* "satrap"—Old Persian *xšassa-pāvan-*. Of Greek origin are such mathematical and astronomical terms as Sanskrit *kendra* "centre" (Greek *kéntron*), *jāmitra* "diameter" (*diámetron*), and *horā* "hour" (*hora*). *Yavana* "foreigner," originally the Greek word for Ionian, is known from as early as the time of Pāṇini. Later, Arabic words such as *taślī* "trigon" came into Sanskrit.

Division of India into linguistic states

The modern Indo-Aryan stage. The division of the Indian subcontinent into linguistic states and even into countries (Pakistan, Bangladesh, and India) is a recent phenomenon. Even after independence from Britain was achieved and partition had taken place, Bombay state existed until it was split into Gujarāt and Mahārāshtra states in 1960. The division of Punjab into Punjab and Haryana states in 1966 occurred as a result of Punjabi agitation for a separate linguistic state. Before independence, under British rule (entrenched from the 18th century), there were princely states within dialect areas; under Mughal rule (16th–18th centuries), Persian was the language which was used by the court and by courts of justice and this practice continued in the latter function for a time under the British. Though Hindi–Urdu may have been a lingua franca, however, the great dialectal diversity of earlier times continued.

Some of the modern Indo-Aryan languages have literary traditions reaching back centuries, with enough textual continuity to distinguish Old, Middle, and Modern Bengali, Gujarati, and so on. Bengali can trace its literature back to Old Bengali *caryā-padas,* late Buddhist verses thought to date from the 10th century; Gujarati literature dates from the 12th century (Śālibhadra's *Bharateśvara-bāhubali-rāsa*) and to a period when the area of western Rājasthān and Gujarāt are believed to have had a literary language in common, called Old Western Rajasthani. Jñāneśvara's commentary on the *Bhagavadgītā* in Old Marathi dates from the 13th century and early Maithili from the 14th century (Jyotīśvara's *Varṇa-ratnākara*), while Assamese literary work dates from the 14th and 15th centuries (Mādhava Kandalī's translation of the *Rāmāyaṇa,* Śaṅkaradeva's Vaiṣṇaviṭe works). Also of the 14th century are the Kashmiri poems of Lallā (*Lallāvākyāṇi*), and Nepali works have also been assigned to this epoch. The work of Jagannāth Dās in Old Oriya dates from the 15th century.

Amīr Khosrow used the term *hindvī* in the 13th century, and he composed couplets that contained Hindi. In early times, however, other dialects were predominant in the midlands (Madhyadeśa) as literary media, especially Braj Bhasa (*e.g.,* Sūrdās' *Sūrsāgar,* 16th century) and Awadhi (*Rāmcaritmānas* of Tulsīdās, 16th century). In the south, in Golconda (Andhra, near Hyderābād), Urdu poetry was seriously cultivated in the 17th century, and Urdu poets later came north to Delhi and Lucknow. Punjabi was used in Sikh works as early as the 16th century, and Sindhi was used in Ṣūfī (Islāmic) poetry of the 17th–19th centuries. In addition, there is evidence in late Middle Indo-Aryan works for the use of early New Indo-Aryan; *e.g.,* provincial words and verses are cited.

The creation of linguistic states has reinforced the use of certain standard dialects for communication within a state in official transactions, teaching, and on the radio. In addition, attempts are being made to evolve standardized technical vocabularies in these languages. Dialectal diversity has not ceased, however, resulting in much bilingualism; for example, a native speaker of Braj Bhasa uses Hindi for communicating in large cities such as Delhi.

Attempt to establish one national language

Moreover, the attempt to establish a single national language other than English continues. This search has its origin in national and Hindu movements of the 19th century down to the time of Mahatma Gandhi, who promoted the use of a simplified Hindi–Urdu, called Hindustani. The constitution of India in 1947 stressed the use of Hindi, providing for it to be the official national language after a period of 15 years during which English would continue in use. When the time came, however, Hindi could not be declared the sole national language; English remains a co-official language. Though Hindi can claim to be the lingua franca of a large population in North India, other languages such as Bengali have long and great literary traditions—including the work of Nobel Prize winner Rabindranath Tagore—and equal status as intellectual languages, so that resistance to the imposition of Hindi exists. This resistance is even stronger in Dravidian-speaking southern India. The use of English as an official language entails problems, however, because with the use of state languages for education, the level of English competence is declining. Another danger faced is the agitation for more separate linguistic states, threatening India with linguistic fragmentation hearkening back to earlier days.

Characteristics of the modern Indo-Aryan languages. The trends noted in Middle Indo-Aryan continue in New Indo-Aryan. The Middle Indo-Aryan vowel sequences *ai* and *au* were changed to single vowels during the development of New Indo-Aryan, final vowels were shortened and deleted, and *ḍ* and *ḍh* sounds between vowels were replaced by the sounds *ṛ* and *ṛh.* The noun cases were further reduced, and the introduction of nominal (noun) forms into the verb system became more pronounced.

Literary languages tend to become somewhat removed from the usual standard colloquial. Literary, or High, Hindi, for example, tends to replace some of the Perso-Arabic vocabulary with Sanskritic items, whereas literary Urdu makes great use of Perso-Arabic words. The gap is formalized in Bengali, in which a distinction is made between the highly Sanskritic language Sadhu-Bhaṣa and the colloquial standard called Calit-Bhasa.

Phonology. [Note: The forms of the words given below reflect actual pronunciation, rather than being transliterated versions of the standard orthographies. For New Indo-Aryan the symbols *ə,* pronounced as the *a* in English "sofa," and *a* are used for the sounds earlier transcribed as *a* and *ā,* respectively; *e.g.,* Gujarati *karũ* "I do" and *māro* "beat" are now written *kərũ* and *maro.* This practice permits certain contrasts to be made among sounds that are significant in the description of dialectal features. In Kashmiri words, *a* is short, opposed to *ā.*]

Vowel changes in New Indo-Aryan

Vowels in sequence contracted in early New Indo-Aryan; *e.g.,* Old Indo-Aryan *aśīti* became Middle Indo-Aryan *asīi,* Hindi and Punjabi *əssī,* and Bengali *aši* "80." Further, *ai* and *au* sounds changed to *e* and *o,* and *aũ* to *ũ,* while *iu developed into ī.* The diphthongs *ai* and *au* were retained well into the New Indo-Aryan period and are still pronounced in some areas; *e.g.,* Braj Bhasa *kərəũ* "I do," *kərəi* "he does." Middle Indo-Aryan *-ḍ-* and *-ḍh* developed into the flaps *ṛ* and *ṛh; e.g.,* Prākrit *sāḍiā* "woman's garment," Kashmiri, Lahnda, Hindi, Gujarati, Bhojpuri, Bengali, Oriya *saṛī* "sari"; and Prākrit *paḍh-* "recite, read," Sindhi *pəṛh-əṇu,* Lahnda *pəṛh-əṇ,* Hindi, Punjabi *pəṛh-na,* Gujarati *pəṛh-vũ,* Marathi *pəṛh-ṇə* "study."

Stress is not generally contrastive in New Indo-Aryan as it is, for example, in English (*e.g.,* noun "éxport," verb "expórt"), though different areas have different rules for placing major emphasis on a given syllable. For example, in Hindi, in which vowel length is pertinent, *gilá* "swallowed" has major stress on the last syllable, *gī́la* "wet," on the first. In Gujarati, on the other hand, vowel length is not pertinent; the stress position depends on which vowels occur in contiguous syllables and on the structure of the syllables, whether open or closed; *e.g., júno* "old," but *dukán* "store." In Bengali each syllable of a word receives about equal stress.

The sounds that most clearly distinguish Indo-Aryan from the rest of Indo-European are the voiced aspirate stops (*gh* and the like, pronounced with an accompanying audible puff of breath) and the retroflexes (*ṭ* and so on, pronounced by curling the tongue upward toward the hard palate). In the outlying New Indo-Aryan areas, however, the sound system is reduced. Sinhalese has no aspirated stops, Assamese has no retroflexes, and Kashmiri has no voiced aspirates. The geographic position of these languages doubtless contributed to these losses: Sinhalese coexists with Tamil, Assamese is surrounded by Tibeto-Burman languages, and Kashmiri is on the border of the Iranian area.

Early New Indo-Aryan dialect distribution

New Indo-Aryan shows evidence of early dialect distribution; this is discernible by considering sound changes proper to each group. The eastern group (Assamese, Bengali, Oriya) has three important changes. Long and short *i* and *u* merged; *e.g.,* Assamese *nila,* Oriya *niḷɔ* (ɔ is similar to the *o* of "coffee" in some English dialects), Bengali *nil* "blue-black" but Sanskrit *nīla;* Assamese *dhuli,* Bengali *dhulo,* Oriya *dhuḷi* "dust" but Hindi *dhūl* and Sanskrit *dhūli.* The vowel sound *a* of Middle Indo-Aryan was replaced by ɔ in Bengali and Oriya and ɒ (similar to the *o* of "hot" in southern British English) in Assamese in initial position and open syllables; *e.g.,* Bengali *mɔron,* Oriya *mɔrɔn,* Assamese *mɒrɒn* "death"; Sindhi, *mərəno* "mortal, death," Sinhalese *mərəṇə,* Gujarati, Marathi *mərəṇ* (compare Sanskrit *maraṇa-*). Moreover, in this group a vowel is affected by the quality of the vowel in a following syllable. For example, in Bengali *ami kori* "I do," the verb root has *o* followed by *i* in the next syllable, but *tumi kɔro* "you do" has an ɔ sound; similarly, *ami kini* "I buy" but *tumi keno.* As a result of vowel assimilation also, Assamese has an ɔ sound instead of ɒ representing Middle Indo-Aryan *a:* Assamese *xɔhur,* Bengali *šošur* "husband's father" (compare Hindi *səsur,* Prākrit *sasura-,* Sanskrit *śvaśura-*).

Assamese and Bengali are set off from Oriya. In the former two, Middle Indo-Aryan *ḍ* and *ḍh* merge medially to *ḍ* (then *ṛ*) with a subsequent development to *r* in Assamese; *e.g.,* Oriya *daṛhi,* Bengali *daṛi,* Assamese *dari* "beard"; Hindi, Gujarati *daṛhī,* Prākrit *dāḍhiā.* Assamese is also distinguished from Bengali by several developments, among them the merger of Assamese retroflex sounds with dental sounds; *e.g.,* Assamese *ut* "camel" but Bengali *uṭ,* Oriya *oṭɔ,* Sindhi *uṭhu,* Lahnda, Pahari *uṭṭh,* and so on. Assamese also has *s* for earlier *c* and *ch* sounds and a *z* sound for *j* and *jh; e.g.,* Assamese *kas* "glass," Bengali *kac;* Assamese *azi* "today," Oriya *aji,* Bengali, Hindi *aj.* In addition, Assamese replaced an *s* sound initially by *x* and between vowels by *h—xɔhur.*

Particular sound changes also characterize languages of the northwest. In this group, an older voiceless stop (*e.g.,* ·*t*) became voiced (*e.g.,* became *d*) after a nasal sound; in other areas, the voiceless stop is retained: Kashmiri *dand,* Punjabi *dənd,* Sindhi *ḍəndu* "tooth" (the *ḍ* in Sindhi is an imploded stop; see below) but Assamese, Bengali, Hindi, Gujarati, Marathi *dãt,* Sinhalese *dətə* (Sanskrit *danta-*). Moreover, in the northwest group a voiced stop (*e.g., d*) preceded by a nasal was assimilated to the latter, resulting in two nasals, which were subsequently reduced to one in some areas; in the rest of New Indo-Aryan, the vowel preceding the nasal was nasalized. Thus, Kashmiri *don* "churning stick," Sindhi *ḍənu* "tribute," Punjabi *dənn* "fine," Lahnda *ḍənn* "force," Kumauni *dan* "roof" contrast with Assamese *dãr* "pole," Bengali *dãṛ* "oar," Hindi *dãḍ* "oppression, fine," and others; all forms derive from Old Indo-Aryan *daṇḍa-* "stick, staff, club, royal power, fine, punishment."

In the sequence of a short vowel followed by two consonants, Pahari differs from the rest of the northwest group and agrees with the rest of New Indo-Aryan. In the northwest this sequence either remained unchanged or the cluster was simplified without lengthening of the vowel; other languages generally simplified the cluster and lengthened the vowel: Punjabi *bhətt,* Sindhi *bhətu,* Lahnda *bhət,* Kashmiri *batɨ* "cooked rice, food" but Nepali, Kumauni, Hindi, Assamese, Bengali, Gujarati, Marathi *bhat.*

Sound changes in Dardic

Dardic occupies a special position. The sibilant sounds did not all merge here. For example, Kashmiri, a Dardic tongue, has *šurah* "16" with *š* rather than *s,* as in most other Indo-Aryan languages, and *sat* "7" with *s.* Further, voiced aspirated stops merged with unaspirated stops in Dardic; *e.g.,* Kashmiri *gur* "horse" but Hindi *ghoṛa;* Kashmiri *dɔd* "milk" but Hindi *dūdh.*

One major feature distinguishing Sindhi from the rest of the northwest group is the development of a series of imploded stops (also called suction stops and recursive stops), for *b, ḍ, j,* and *g.* Implosive stops also occur in the Sindhi vicinity; for example, Kacchi has imploded *b.* Another feature that distinguishes Sindhi from other northwest languages, including Kacchi, is the retention of the Middle Indo-Aryan final short vowels; *e.g.,* Sindhi *əkhi* "eye" but Hindi *ãkh* (Middle Indo-Aryan *akkhi-*).

Punjabi is distinguished from other members of the northwest group by its tonal system, having low (̀), mid (̄), and high (́) tones. Initial voiced aspirated stops of earlier Indo-Aryan appear in Punjabi as voiceless stops with low tone on the following vowel; *e.g.,* Punjabi *kòṛa* but Hindi *ghoṛa;* Punjabi *tàī* "2½" but Hindi *ḍhaī.* Non-initially, a voiced aspirate became unaspirated and the preceding vowel received high tone; thus, Punjabi *dū́d* "milk" but Hindi *dūdh,* and Punjabi *lā́b* "profit" but Hindi *labh.*

Gujarati, Marathi, and Konkani in the west and southwest differ from the languages of the midlands in that, as in the east, there is no contrast between long and short *i* and *u* vowels. The *i* of Gujarati and Marathi *vis* "20" is pronounced like the *ee* of English "teeth," the *i* of Gujarati *iccha* and Marathi *iččha* "wish" like the *i* of "pitch," but such a difference is not contrastive, as it is in Hindi (*gīla* "wet": *gila* "swallowed"). Gujarati has certain features that, in turn, set it apart from the other languages of this group. In addition to *e* and *o* sounds, it has the open vowels ɛ, ɔ; *e.g., cɔthũ* "fourth" (Middle Indo-Aryan *cauttha*), *bɛs-vũ* "to sit" (Middle Indo-Aryan *baisai* "sits"). Moreover, Gujarati has murmured vowels, generally developed from vowels followed by *h; e.g., kɛh che* "says" (*h* represents murmuring of the vowel), Old Gujarati *kahai chai.* Marathi and Konkani have two series of affricate sounds; *e.g., č* (pronounced as the *ch* in English "chat"; the equivalent of *c* in some other languages) and *c* (pronounced as the *ts* of "rats").

There was clearly mutual influence of Indo-Aryan languages at an early time, together with movement of groups of speakers (compare the position of Pahari). Thus, while Punjabi *səcc* "true" is the expected form comparable to Middle Indo-Aryan *sacca-* (Old Indo-Aryan *satya-*), Hindi *səc* "true" does not represent the expected outcome. The item *səc* must come from the Punjabi area.

Gender in New Indo-Aryan

Grammar. Like Middle Indo-Aryan, New Indo-Aryan distinguishes only two numbers—singular and plural. Unlike Middle Indo-Aryan, the New Indo-Aryan languages differ in the degree to which gender distinctions are made. Three genders are retained in the west and southwest (Gujarati, Marathi, Konkani), and this is true also of Sinhalese. Unlike Gujarati, Marathi, and Konkani, in which every noun, whether it denotes an animate being or not, has a particular gender that is unpredictable, Sinhalese restricts masculine and feminine gender to animates and neuter to inanimates. The eastern group (Assamese, Bengali, Oriya) has no grammatical gender distinctions, and two genders are distinguished elsewhere.

Over a large area of New Indo-Aryan the noun has only two cases—direct and oblique. A lack of distinction between direct and oblique cases in the plural is typical of several languages, including forms in Hindi, Gujarati, Marathi, and Bhojpuri. Direct forms are used independently, oblique forms before postpositions (words or word elements following a noun that function similarly to English prepositions) and other affixes; the combination of stem and postposition serves the function of inflected case forms of earlier Indo-Aryan. Thus, to denote an object (direct or indirect) Hindi uses the postposition *ko,* which occurs in direct object constructions normally only with nouns denoting animate beings; *e.g., ləṛke-ko dekhta hɛ* "He sees the boy," *ləṛke-ko miṭhaī do* "Give a sweet to the boy." Other postpositions are *mẽ* "in," *pər* "on," *se* "from, with, by means of." A large group of postpositions are linked to the noun with the affix *ka* (oblique form *ke,* feminine *kī*), which also is used to form adjectives (possessives); *e.g., ləṛke-ke sath gəya* "He went with the boy," *ləṛke-ke pas hɛ* "The boy has it" (literally, "It is by the boy"). Many such postpositions represent old nominal (noun) forms. Other New Indo-Aryan languages have systems similar to that of Hindi, though the forms of the postpositions differ.

Though the nominal (noun) system of Punjabi is very close to that of Hindi, it has separate ablative (indicating separation and source) and locative (indicating place) forms in the singular and plural, respectively, for nouns such as *koṭha* "house"; *e.g., koṭhiõ* "from the house," *koṭhī̃* "in the houses." Some languages have a fuller case system than that noted above; *e.g.,* Bengali has a genitive singular ending, a genitive plural ending, and a locative case. Similarly, Kashmiri has nominative, dative, ablative, and agentive cases. Not all such case forms are inherited from Middle Indo-Aryan. In addition to case endings, these languages also use postpositions; *e.g.,* Kashmiri *garājas-andar* "in the garage," with *-andar* after the dative ending *-as.*

Comparatives and superlatives

Adjectives behave generally in the same way as nouns but have a syntactic restriction. In Hindi the possessive is in the oblique (non-nominative) form, as is the noun after which it occurs; but in the plural, only the noun has the oblique form. Further, the formation of comparatives and superlatives with derivative affixes has been eliminated. To a Sanskrit sentence such as *ime amū-bhyaḥ āḍhya-tarāḥ* "These (people) are richer than those," in which the comparative *āḍhya-tara* occurs construed with the ablative form, corresponds a Hindi sentence *ye un-se əmīr hɛ̃,* in which no comparative affix is used—literally, "These are rich from (*i.e.,* in comparison with) those." Comparable constructions with a postposition meaning "from" occur elsewhere in New Indo-Aryan.

The pronominal system of New Indo-Aryan formally resembles the Middle Indo-Aryan stage more than its noun system. For example, Gujarati *hũ* "I," *mɛ̃* "I" (agentive), *əme* "we" (also agentive) are directly comparable to Apabhraṃśa *haũ, maĩ, amhaĩ.* The number distinctions of the Middle Indo-Aryan pronoun have been replaced, however, by distinctions of familiarity and politeness. For example, Hindi and Bengali have a three-way distinction—Hindi *ap,* Bengali *apni* "you" are polite or honorific forms; Hindi *tum,* Bengali *tumi* are informal forms; and Hindi *tũ,* Bengali *tui* are used only for inferiors and small children. (Hindi and Bengali differ, however, in the plural forms of these.) In Gujarati, on the other hand, *tũ* is a very familiar pronoun, whereas *təme* is used generally, covering the approximate domains of Hindi *ap* and *tum; ap,* if used, strikes the hearer as fawning. Marathi has a similar system. Southwestern languages also make a distinction in the 1st person plural between inclusive and exclusive, the exclusive excluding the person spoken to. In the form of the relative pronoun and the 3rd person pronoun, languages differ in the degree to which gender distinctions are made, thus contrasting with Old and Middle Indo-Aryan, in which these forms had three genders. For example, Marathi has masculine, feminine, and neuter for the relative pronoun, while Bengali has animate and inanimate.

New Indo-Aryan languages differ in the degree to which finite verb forms have been replaced by nominal (noun) forms. In Bengali a contrast is made between continuous or actual present (English "be . . . -ing") and noncontinuous or habitual present; *e.g., ami kaj kor-i* "I work" (literally, "I do work"), with the ending *-i,* contrasts with *ami kaj kor-ch-i* "I am working," in which *ch* intervenes between the root and the ending. Hindi has a similar contrast but uses nominal forms; *e.g., mɛ̃ kam kar-ta hũ* "I work," *mɛ̃ kam kər rəh-a hũ* "I am working." Both contain the finite form *hũ* of the auxiliary; but *kər-ta* and *rəh-a* are nominal forms, the latter the past of *rəh-*"stay." Gujarati has both types, the present tense using finite verb forms, the imperfect employing nominal forms; *e.g., hũ kam kərũ chũ* "I work, am working" and *hũ kam kər-to hə-to* "I was working, used to work." Even in areas in which finite forms are not used in the present, they occur in the imperative forms and what may be called the subjunctive; *e.g.,* Hindi *tum kam kər-o* "work," *mɛ̃ əndər aũ* "May I come in?"

Person–number system

The person–number system of the New Indo-Aryan verb accords with the use of pronouns. For example, the forms *ja-o, kər-o* in Gujarati *təme kyã jao cho* "Where are you going?" and *šũ kəro cho* "What are you doing?" are historically plurals but are used with reference to one person addressed by the pronoun *təme.* Similarly, in Hindi, in which a person distinction is not made in the plural, *ap kəhã ja rəhe hɛ̃, ap kya kər rəhe hɛ̃,* equivalent in meaning to the Gujarati sentences, have the plural form *rəhe hɛ̃.* Bengali has completely given up any number distinction in verb forms: *ami/amra kori* "I/we do." In the 3rd person a distinction is made between ordinary and honorific: *še* (ordinary)/*tini kɔren,* plural *tara/tãra kəren.* Other languages (*e.g.,* Hindi) also have honorific forms, for which the plural is used.

In the formation of the future there are again regional differences. Some retain the future in *-s-* (Gujarati *hũ kər-iš,* 3rd person *e kər-š-e*) or *-h-* (*e.g.,* eastern dialects of Braj Bhasa, *cəlihəɔ̃* "I will go"). Characteristic of the Eastern languages and of Bihari (including Bhojpuri, Magahi, Maithili) is the suffix *-b-; e.g.,* Bengali *jabe* "will go." All of these are finite forms. On the other hand, in Hindi and adjoining areas, the future is inflected for gender.

A similar contrast between the use of verbal and nominally inflected forms also appears in the past tense forms. The predominant pattern in New Indo-Aryan is that of Middle Indo-Aryan: forms are used that are etymologically participles.

The New Indo-Aryan languages retain the passive and causative forms. The causative is conservative in retaining both the affixes that appear in Middle Indo-Aryan and vowel alternation. The passive is also formed by affixation in some areas. But many languages also have a compound formation involving the verb *ja* "go" and an auxiliary (*hɛ̃*); *e.g.,* Hindi *yahã hindī bol-ī ja-t-ī hɛ̃* "Hindi is spoken here."

There are other auxiliaries, which, like *hɛ̃,* can occur with any verb in the language; *e.g.,* the verb "can," Hindi *sək-,* Gujarati *šək.* A characteristic feature of New Indo-Aryan, however, is the use of certain verbs, variously called vector verbs or compound verbs, in restricted contexts and with particular semantics. For example, one can say *mər gə-ya* "He died," *bhūl gə-ya* "He forgot," *bol uṭh-a* "He blurted out" in Hindi, using the verbs *ja* "go" (masculine singular past *gə-ya*), *uṭh* "stand up." This phenomenon is pan-Indo-Aryan and still requires investigation.

New Indo-Aryan word order

The examples cited above also illustrate the normal word order in New Indo-Aryan languages: subject (including agential forms), object (with attributive adjectives preceding), verb (together with auxiliaries). Adverbials can precede the full sentence or occur after the subject, with slight differences in emphasis; *e.g.,* Hindi *mɛ̃ kəl aũga,* or *kəl mɛ̃ aũga* "I will come tomorrow (*kəl*)." Relative clauses normally precede correlatives: Hindi *jo admī kəl tumhare ghər-mẽ tha vo kɔn hɛ* "Who (*kɔn*) is the man (*admī*) who (*jo*) was in your house yesterday?" A notable exception to the normal final position for verbs occurs in Kashmiri, in which the verb usually occurs in second position after the subject; thus, to Hindi *vo kha̦ rəha hɛ* "he is eating" corresponds Kashmiri *su chu kh́avān* with the auxiliary *chu* after the subject.

Vocabulary. The two most important sources of non-Indo-Aryan vocabulary in New Indo-Aryan are Persian (including Arabic items introduced through Persian), the court language of the Mughals, and English. The Perso-Arabic vocabulary permeates every aspect of New Indo-Aryan vocabulary, especially in the midlands (Uttar Pradesh through the Punjab). There are, of course, Hindi-Urdu words proper to Islām: Hindi *kuran* "Qur'ān," *ʿīd* (name of a holy day), *nəmaz* (certain prayers), *məsjid* "mosque," as well as the word for "religion," *məźhəb.* In addition, there are numerous Perso-Arabic military and administrative terms (*kila* "fort," *səvar* "horseman," *ədalət* "court of justice"); architectural and geographic terms (*imarət* "building," *məkan* "house," *məhəl* "palace," *duniya* "world," *ilaka* "province"); words having to do with learning and writing (*kələm* "pen," *kitab* "book," *ədəb* "literature, good manners") and with apparel (*jeb* "pocket," *moja* "socks," *rumal* "handkerchief") and anatomy (*khūn* "blood," *gərdən* "neck," *dil* "heart," *bazu* "arm," *sər* "head"). Indeed some of the most common vocabulary is of this origin: *tārīkh* "date," *vəkt* "time," *sal* "year," *həfta* "week," *umər* "age," *admī* "man," *ɔrət* "woman," and others. Even the grammatical apparatus of postpositions and conjunctions reflects Perso-Arabic influence; *e.g., -ke bad* "after," *əgər* "if," *məgər* "but," *ya* "or."

The colloquial language used by any Hindu or Muslim communicating in Hindi-Urdu will contain a large number of such words. There have been efforts to polarize the two, and at times champions of Indo-Aryan have tried to replace Perso-Arabic vocabulary with Sanskritic words. The style that tends toward eliminating all but the most common Perso-Arabic words may be called High Hindi, written in the Devanāgarī script, as opposed to High Urdu, which retains Perso-Arabic of long standing, uses Persian and Arabic for learned vocabulary and is written in the Perso-Arabic script.

Influence of English on vocabulary

The influence of English as a source of borrowing still continues, and it is rare to hear a conversation on any technical subject among speakers of any Indian language in which English words are not liberally used. Among loanwords from English are names of conveyances such as Hindi *rel-gaṛi* "railroad-train" and *ṭɛksī* "taxi"; profession names such as *injinīr* "engineer," *jəj* "judge," *ḍaktər* "Western doctor," *pulis* "police"; and terms of educational administration such as *kaləj* "college" and *yunivərsiṭī* "university." English words are susceptible to replacement in India by Sanskritic ones as are those of Perso-Arabic origin.

Of much lesser magnitude are New Indo-Aryan borrowings from other languages, among them Portuguese and Turkic. From the latter, the word *urdū* came to be used as the name of a language. From Portuguese come such Hindi words as *ənənnas* "pineapple," *paũ* "(Western style) bread," *kəmīz* "(Western) shirt," *kəmra* "room," and *girja* "(Christian) church."

Writing systems. Ancient India had two main scripts in which Indo-Aryan languages were written. Kharoṣṭi, used in the northwest, is of Aramaic origin and is written from right to left; Brāhmī, of North Semitic origin, is written from left to right and appears earliest on Aśokan inscriptions in areas other than the northwest. Most scripts of New Indo-Aryan are developments of the Brāhmī. The Devanāgarī (or simply Nāgarī), used for writing Sanskrit documents in North India, is the script of Hindi and Marathi as well as Nepali. Gujarati uses a more cursive derivative. Devanāgarī is also used, mainly among Hindus, for Kashmiri, which has, in addition, a traditional script called Sarada, which is not now in common use. The Perso-Arabic script is used instead. Also usually written in Perso-Arabic writing are Urdu and Sindhi (for which the Devanāgarī is also used in schools in India), whereas Punjabi employs it in Pakistan as well as a particular script of its own, known as Gurmukhi ("From the Teacher's Mouth") in the sacred writings of the Sikhs. In the east, the scripts used for Bengali and Assamese are closely related; and that of Oriya, related to the other two, is highly cursive like that of neighbouring Dravidian languages. Such is also the case with Sinhala.

The traditional alphabets are both over-explicit and not clear enough with regard to accurate representation of the spoken word. As systems in which a consonant symbol with no other accessory symbol accompanying it stands for the syllable consisting of the consonant followed by short *a,* they require previous knowledge of items for correct interpretation; Hindi *kərta* is written *ka-ra-tā* in the Devanāgarī, and one must know that the word has only two syllables. Though Bengali has only the spirant sound *š,* the alphabet has symbols for *ś, ṣ,* and *s,* as in Old Indo-Aryan; but verb forms such as *kori* and *kəren* are written *ka-ri* and *ka-re-na,* both with the same initial symbol. And, though syllabic *ṛ* was lost as early as Middle Indo-Aryan, the scripts have a separate symbol for this. Script reform has been suggested; it has even been proposed that all Indo-Aryan languages adopt a Latin (roman) alphabet with diacritics, but chances for this are poor. (Ge.Ca.)

THE IRANIAN LANGUAGES

Languages of the group. The various Iranian languages fall distinctly into three categories—Ancient, Middle, and Modern Iranian.

Avestan and Old Persian

Ancient (Old) Iranian. Of the ancient Iranian languages, only two are known from texts or inscriptions, Avestan and Old Persian, the oldest parts of which date from the 6th century BC. Avestan was probably spoken in northeastern Iran, and Old Persian is known to have been used in southwestern Iran. Other ancient Iranian languages must have existed, and indirect evidence is available concerning some of these. Thus, from the 5th-century-BC historian Herodotus, the Median word for "female dog" (*spaka*) is known, and a number of Median loanwords have been recognized in the Old Persian inscriptions. In addition, a number of Median personal names are attested in various sources. It is likely that all those languages that are known only from the Middle Iranian period were in fact spoken in a less developed form in the ancient period. The same observation may apply to some of those modern Iranian languages that are not attested in the earlier periods.

The degree of mutual intelligibility that existed among the ancient Iranian languages is not known with certainty. The differences in the nature of the surviving sources have to be borne in mind. On the one hand, there is the religious poetry of Zoroaster in the Avestan language and on the other, the official inscriptions of the Achaemenid rulers in Old Persian. Differences in the method of transmission present a further difficulty in the way of direct comparison. Nevertheless, it can safely be stated that the degree of mutual intelligibility must have been much greater between the ancient languages than between the Middle Iranian languages and that those languages geographically closer to each other probably were mutually understood better than those spoken in areas farther apart.

Avestan can hardly be said to be known beyond the ancient period, although only the earliest texts, the Gāthās, are as old as the 6th century BC, and the later texts represent the language of several subsequent centuries. Old Persian, on the other hand, itself spanning the 6th to the 4th century BC, was continued more or less directly by the various forms of Middle Persian. Even here, however, although both Old and Middle Persian represent the language of the royal court, there are considerable differences between them for which no satisfactory explanation has yet been given.

Middle Iranian. Middle Persian is known in three forms, not entirely homogeneous—inscriptional Middle Persian, Pahlavi (often more precisely called Book Pahlavi), and Manichaean Middle Persian. Middle Persian belongs to the period 300 BC to AD 950, and was, like Old Persian, the language of southwestern Iran. In the northeast and northwest the language spoken was Parthian, which is known from inscriptions and from Manichaean texts. There are no significant linguistic differences in the Parthian of these two sources. Most Parthian belongs to the first three centuries AD.

Eastern group of Middle Iranian languages

Middle Persian and Parthian were doubtlessly similar enough to be mutually intelligible, but they differ so greatly from the eastern group of Middle Iranian languages that these must have appeared to be almost foreign languages. The languages of the eastern group, moreover, cannot have been themselves mutually intelligible. The main known languages of this group are Khwārezmian (Chorasmian), Sogdian, and Saka. Less well-known are Old Ossetic (Scytho-Sarmatian) and Bactrian, but from what is known it would seem likely that these languages were equally distinctive. There was probably more than one dialect of each of the languages of the eastern group, although there is certainty only in the case of Saka, for which at least two dialects are clearly attested. The main Saka dialect is known as Khotanese, but a small amount of material survives in a closely related dialect called Tumshuq, formerly known as Maralbashi.

A few words are known in all of these eastern Iranian languages from as early as the 2nd to the 4th century AD, but substantial evidence begins for Sogdian in the 4th century, for Saka probably no earlier than the 7th century (though that for Tumshuq may be a few centuries older), and for Khwārezmian not until the 12th century and later. The principal evidence for Bactrian belongs to the 2nd century. To the same period belong the Scytho-Sarmatian names of the earliest inscriptions.

All the eastern Iranian languages of the Middle Iranian period were spoken in Central Asia, with the exception of the language of the Scytho-Sarmatian inscriptions from southern Russia, north of the Black Sea. More precisely, Bactrian was spoken in northern Afghanistan and in the adjacent parts of what is now Soviet Central Asia. Khwārezmian was the language of Khwārezm (Khiva), now an *oblast* in western Uzbekistan but formerly of greater extent. Sogdian was probably spoken over most of Soviet Central Asia, especially in eastern Uzbekistan, Tadzhikistan, and western Kirgiziya. There were also colonies of Sogdians in various cities along the trade routes to China; in fact, most Sogdian material comes from outside Sogdiana. The Saka dialects, Khotanese and Tumshuq, were spoken in Chinese Turkistan, modern Sinkiang; Tumshuq is the name of a small village in the extreme west of Sinkiang. Khotanese was spoken in Khotan near the modern city of Khotan (Chinese Ho-t'ien) on the southern route across the Takla Makan desert and within about 100 miles (160 kilometres) to the north and to the east of Khotan, where manuscripts have been found, mainly at the sites of former shrines and monasteries.

Modern Iranian. The discontinuity already observed between Old and Middle Iranian is even more striking between Middle and Modern Iranian. There are no modern counterparts at all to Khwārezmian, Bactrian, and Saka, and there is no direct continuity in the case of any of the other Middle Iranian languages. Even Modern Persian does not represent a straightforward continuation of Middle Persian but is rather a koine (a dialect or language of a small area that becomes a common or standard language of a larger area), based mainly on Middle Persian and Parthian but including elements from other languages and dialects. Although Sogdian is known in several forms, possibly representing different dialects, none of these can be considered the direct ancestor of modern Yaghnābī, spoken at present in the valley of the Yaghnob (Yagnob) River, a tributary of the Zeravshan. Yaghnābī, nevertheless, certainly belongs linguistically to the Sogdian family. Similarly, the languages of the Scytho-Sarmatian inscriptions may represent dialects of a language family of which Modern Ossetic is a continuation, but it does not simply represent the same language at an earlier date.

Modern Iranian state languages

Only four of the many modern Iranian languages are the official languages of the state in which they are spoken. The chief of these is Persian (known in Persian itself as Fārsī), the national language of Iran, which is spoken by about 18,000,000 people as a native language. It is recognized, moreover, as a second language in Afghanistan, where it is spoken in only a slightly different form. The national language of Afghanistan is the East Iranian language known as Pashto, of which there are about 15,000,000 speakers, many living in Pakistan. Tadzhik is spoken by at least 6,000,000 people widely spread throughout the Tadzhik Soviet Socialist Republic and the rest of Soviet Central Asia and is readily intelligible to speakers of Persian, to which it is very closely related, although it is in some respects more archaic. In addition to being the national language of Tadzhikistan, Tadzhik is important as the lingua franca of the Pamirs, a region where a remarkable variety of Iranian languages and dialects is spoken. Fewer than 500,000 people speak Ossetic. Most of the Ossetes live in two administrative divisions of the U.S.S.R., the Severo-Ossetian A.S.S.R. of the Russia S.F.S.R. and the Yugo-Ossetian autonomous *oblast* of the Georgian S.S.R. Although spoken in the heart of the Caucasus Mountains, Ossetic is an East Iranian language not mutually intelligible with any other Iranian language.

Two other Iranian languages, Kurdish and Baluchi, are spoken over a vast area, although they have not been officially accepted as the national language of an established state. Kurdish is spoken by about 10,000,000 people living in Iran, Iraq, Turkey, Syria, and Soviet Transcaucasia. More than 2,000,000 people speak Baluchi as their chief language; they are spread widely over parts of eastern Iran, Pakistan, Afghanistan, and southern Soviet Central Asia. In Iran, Baluchi speakers live mainly in the region of Baluchistan, a region in the southeast that now forms part of a province with Seistan. In Pakistan, Baluchi speakers live mainly in the southwestern province of Baluchistan; in Soviet Central Asia, they are found mainly around Merv in southern Turkmenistan; and in Afghanistan, they are widely scattered, mainly over the southwestern portion of the country. There is a sizable Baluchi colony in Oman, and many Baluchi merchants have settled in the sheikhdoms of southern Arabia and along the east coast of Africa as far south as Kenya. Linguistically, Baluchi and Kurdish are both West Iranian languages. Baluchi is thus much more closely related to Kurdish than it is to its close neighbour Pashto. According to the most likely theory, the present eastern location of Baluchi speakers is the result of migrations from the region of the Caspian Sea during the Middle Ages.

Dialects. The six modern Iranian languages discussed above are the only ones that have an established literary tradition. They are not, however, homogeneous, each having its own dialect divisions. No definitive dialect classification has yet been made, nor indeed has any attempt at systematic classification of the whole range of Iranian languages won wide acceptance. The usual practice, followed here, is simply to list the main languages in groups of varying size, arranged on a roughly geographical basis.

Ossetic dialects

There are two main dialects of Ossetic: the eastern, known as Iron, and the western, known as Digor (Digoron). Of these, Digor is the more archaic, Iron words being often a syllable shorter than their Digor counterparts; *e.g.,* Digor *madä,* Iron *mad* "mother." Iron is spoken by the majority of Ossetic speakers and is the basis of the literary language. Chosen in the 19th century for the translation of the Bible, it is still the official language today. Little is known of the other Ossetic dialects. A small amount of the Ossetic dialect of Tual in the south, which differs little from Iron, was published in Georgian script at the beginning of the 19th century.

Yaghnābī is still spoken by a small number of people southeast of Samarkand. It has two main dialects, eastern and western, which differ only slightly. The characteristic difference is between a western *t* sound and an eastern *s* sound from an older *θ* sound (as *th* in English "thin"); *e.g.,* western *mēt,* eastern *mēs* "day," beside Sogdian *mēθ* (Christian Sogdian *myθ*).

Dialects of the Shughnī group are spoken in the Pamirs. Closely related to this group is Yāzgulāmī. A period of a Yāzgulāmī–Shughnī common language (protolanguage) has been postulated by some scholars, after which it separated first into Yāzgulāmī and Common Shughnī; and then Common Shughnī gradually divided into Sarīkolī, Oroshorī-Bartangī, Rōshānī-Khufī, and Bajuvī-Shughnī. Sarīkolī, the easternmost of these dialects, is spoken in Chinese Sinkiang.

Speakers of Wakhī number 10,000 or so in the region of the upper Panj. Vākhān (Wākhān), the Persian name for the region in which Wakhī is spoken, is based on the local name Wux̌, a Wakhī development of *Waxšu, the old name of the Oxus (modern Amu Darya). (An asterisk denotes a hypothetical, unattested, reconstructed form or word.) The Wakhī language is remarkably distinct from its neighbours and has many archaic features.

Around the bend of the Amu Darya and in the valley of the Vardūj River to the southwest, a few people speak dialects of the Sanglēchī-Ishkāshmī group. This group is clearly distinguished from its neighbours but is closely related to the other languages of the Pamirs.

Scarcely more than 2,000 people speak dialects of the Yidghā-Munjī group. Monjān is a very remote valley located in northern Afghanistan, and it is separated by a mountain pass from the Sanglēchī-speaking region. Yidghā is spoken in the valley of the Lutkuh in Chitrāl, which is now in Pakistan. Yidghā-Munjī is most closely related to Pashto.

Pashto dialectal groups

The existence of two dialectal groups within Pashto has long been known. Thus, the word Pashto represents a southwestern dialect form (*paštō*), in contrast to a northeastern (*paxtō*). According to one hypothesis, Pashto literature, which exists certainly from the 17th century and possibly from the 11th, was created among the northeastern tribes. Two minor dialects, Wazīrī and Waṉētsī, have some features of special interest.

Although spoken in a few villages in Afghanistan, two languages have features closely associating them with Western Iranian. These are Parāchī, spoken in the Hindu Kush north of Kābul, and Ōrmuṛī, found in two dialects, one in the Lowgar Valley south of Kābul and the other in Kāniguram in Wazīristan.

Farther south is the wholly West Iranian language, Baluchi, mentioned above. Despite the vast area over which Baluchi is spoken, its numerous dialects are all mutually intelligible. The most recent study of the Baluchi dialects divides them into six groups: Eastern Hill dialects; Rākhshānī dialects including that of Merv; Sarawānī; Kechī; Loṭunī; and the coastal dialects. Of these, Rākhshānī is the most widely spoken and is used for broadcasting both in Pakistan and in Afghanistan, but the coastal dialects have the greatest prestige and the most extensive literature.

In the southeastern corner of Iran, Baluchi gradually gives way to the Bashkardī dialects.

In central Iran the influence of Modern Persian is everywhere strongly felt, and it is often difficult to distinguish between dialects of Modern Persian, Persian with dialectal traits, and closely related languages. In Yazd and Kerman the Parsis speak the old Gabrī dialect, whereas the Muslims speak Persian. Among other central dialects are Nātanzī, Sōī, Khunsārī, Gazī (near Isfahan), Sīvandī (northeast of Shīrāz), Vafsī, and Ashtiyānī, to name but a few.

Semnānī, spoken east of Tehrān, forms a transitional stage between the central dialects and the Caspian dialects. The latter are divided into two groups, Gīlakī and Māzandarānī (Tabarī). Also closely related is Tālishī, spoken on the west coast of the Caspian Sea on both sides of the border with the U.S.S.R. To this northwestern group belong the so-called southern Tātī dialects spoken south and southwest of Qazvīn, as well as the scarcely known dialects of Harzan and Galinqaya spoken northwest of Tabriz. The name Tātī is usually applied to the dialects spoken in Russian Dagestan and northeast Azerbaijan. They differ little from Modern Persian.

Of the dialects of Fars Province, only Larī, southeast of Shīrāz, is notably distinctive. Kumzarī in Oman and the Lur dialects of the southwest also differ little from Persian.

Kurdish dialects

There are many dialects of Kurdish, the widely spoken West Iranian language that is thought to occupy a dialectal position intermediate between Baluchi and Persian. Three main dialect groups can be distinguished—northern, central, and southern. A systematic study has been made of the dialects of Iraq, which include ʿAqrah (Akre), ʿAmādīyah, Dahūk, Shaykhān, and Zākhū in the northern group, and Irbīl (Arbīl), Bingird, Pishdar (Pizhdar), Sulaymānīyah (Suleimaniye), and Wārmāwah in the central group. The Central Mukrī dialect is spoken in the extreme west of Iran, south of Lake Urmia.

Gorānī is spoken in several dialects, mainly in the Zagros Mountains, and it is strongly influenced by the surrounding Kurdish dialects. The Gorānī dialect of Hawrāman, Hawrāmī, is notable for its many archaic features. Closely related to Gorānī is Zaza (Dimli), spoken west of Iran.

Historical survey of the Iranian languages. *The Iranian protolanguage and its development.* By the time Iranian begins to be attested in the 6th century BC, the language is already found differentiated into several distinct languages. Scholars have reconstructed the sound system and some of the grammatical features of Common Old Iranian, the protolanguage that preceded these dialects.

The phonological system that underlay Common Old Iranian was by and large maintained everywhere throughout the Iranian-speaking world. It consisted of the following distinctive consonant sounds:

k	g	x	[ɣ]	ŋ
č	ǰ			
p	b	f	[β]	m
t	d	θ	[ð]	n
š	ž			
y	r	l	w	h

Unfamiliar symbols are taken from the International Phonetic Alphabet, or are conventional transcriptions (*e.g.*, *š* for the *sh* sound in "ship," *ž* for the *z* sound in "azure," *č* for *ch* in "church," and *ǰ* for *j* in "jam"). The voiced fricatives (*i.e.*, the first three consonants represented in the fourth column—ɣ, β, and ð), which are produced with vibrating vocal cords and local friction, may be regarded as variants of the voiced stops (*e.g.*, *g*, *b*, *d*); but they are characteristic of Iranian languages generally and especially of the eastern Iranian languages. In addition to these sounds Old Persian had another sibilant sound, often transcribed as *ç* or *ss*, which developed from the cluster *θr* (pronounced as the *thr* in "three"). In Middle Persian it fell together with the *s* sound. The most noticeable alteration of the old sound system is the introduction in some languages of additional series of consonants under the influence of neighbouring languages. Thus, Ossetic has a series of ejective sounds (uttered with a simultaneous glottal stop) on the pattern of the unrelated Caucasian languages; and a number of Iranian languages have a retroflex series (produced with the tongue tip curled up toward the roof of the mouth) as a result of contact with Indo-Aryan languages.

Changes from Indo-European sounds to Iranian sounds

Some of the differences between Iranian languages arose as a result of different developments of the earlier sounds. Thus, the Indo-European sounds *ḱ*, *ǵ*, and *ǵh* resulted in Indo-Iranian *ś*, *ź*, and *źh*, which in turn became *s*, *z*, and *z*, respectively, in Avestan but *θ*, *d*, and *d* in Old Persian. Hence, Indo-European **ḱm̥tó-* "hundred" became Indo-Iranian **śatá-*, attested by Old Indo-Aryan *śatá-*, and then Avestan *sata-*, but Old Persian *θata-*. Nevertheless, *θ* and *d* as well as *s* and *z* belong to the basic pattern, the difference being merely distributional.

The main source of differentiation is in the variation of consonant cluster development and that of groups of consonants and semivowels. Here again it is mainly a question of distributional differences. Thus, the Indo-European group **ḱu̯* became Indo-Iranian **śu̯*, retained in Old Indo-Aryan in the spelling *śv* of the standard transcription. Indo-Iranian **śu̯* developed variously in Iranian: *s* in Old Persian, *sp* in Avestan and Median, *ś* (written *śś*) in Khotanese, and *š* in Wakhī. These developments can be seen in the following forms of the Indo-European word **eḱu̯o-* "horse": Old Indo-Aryan *áśva-*, Avestan and Median *aspa-*, Old Persian *asa-*, Khotanese *aśśa-*, and Wakhī *yaš*. Yet another development can be seen in Ossetic, in which the word for "mare," Avestan *aspā-*, appears as Digor *äfsä* and Iron *yäfs*.

The vowel system of Common Old Iranian consisted of short and long varieties of *a*, *i*, and *u*, and a neutral vowel *ə* (similar to the *a* in "sofa"). This analysis assumes that the Indo-Iranian vocalic *r* (r̥) had already developed to *ər* in Proto-Iranian, just as its long counterpart became *ar*. An early and general monophthongization of the diphthongs *ai* and *au* to *ē* and *ō*, respectively, must also be considered characteristic, although it should not be ascribed to Common Old Iranian as is sometimes done. This basic system was almost everywhere maintained, sometimes with the addition of one or two distinctive vowel sounds (phonemes).

For further details concerning the relationship of Iranian to Indo-European and Indo-Aryan, see the introduction of this article.

The Old Iranian stage. Old Persian was the language of the Achaemenid court. It is first attested in the inscriptions of Darius I (ruled 522–486 BC), of which the longest, earliest, and most important is that of Bīsitūn. At Bīsitūn are also inscribed versions of the same text in Elamite and Babylonian, and fragments of an Aramaic version on papyrus documents from Elephantine (modern Jazīrat Aswān) also exist. Old Persian words and names are also to be found in large numbers as loanwords in contemporary Elamite sources and in 5th-century-BC Aramaic documents.

As early as the time of Darius the Great's successor, Xerxes I (ruled 486–465 BC), the inscriptions show linguistic tendencies characteristic of the development from Old to Middle Persian. After Xerxes the production of original Old Persian inscriptions declined, probably as a result of the wider adoption of Aramaic and Elamite as the usual means of writing. With Artaxerxes III (ruled 359/358–338 BC), Old Persian inscriptions came to an end. The break is marked by Alexander's destruction of Persepolis in 330 BC.

Alternate terms for Avestan

By far the largest part of attested Old Iranian is written in the language now usually called Avestan, after the Avesta, the name given to the collection of works forming the scripture of the Zoroastrians. The name itself is Middle Persian. In former times, this language was called Zend, another Middle Persian word, which refers to the Middle Persian (Pahlavi) commentary on the Avesta. Because the homeland of the Avestan language was long thought to be in Bactria, it was often in the past called Bactrian. Bactrian is now used to designate a different Iranian language belonging to the Middle Iranian period.

Since the beginning of the 20th century it has been generally accepted that the homeland of the Avesta was Khwārezm, which in ancient times included both Merv and Herāt. Merv is now in Turkmenistan, Herāt in northwest Afghanistan.

The oldest part of the Avesta is known as the Gāthās, the poems composed by Zoroaster (Zaraθuštra), the founder of the Zoroastrian religion. His date is uncertain but is traditionally ascribed to the 7th to 6th century BC. The so-called *Khurda Avesta* ("Little Avesta") is a miscellany of texts of later date, the oldest parts of which may have been composed about 400 BC. The language of the *Khurda Avesta* is different in many details from that of the more archaic language of the Gāthās, and it may even represent a different dialect. Many uncertainties surround the detailed interpretation of the Avesta as a result of the method of transmission. The Avesta was not recorded until after the language had ceased to be used, except by Zoroastrian priests. The present manuscripts date from the 13th century and later, although they reflect the recording of the priestly tradition in the special Avestan script during the 6th century AD.

The Middle Iranian stage. Middle Persian was the official language of the Sāsānians (AD 224–651) and was used for their inscriptions. The most important of these is the 3rd-century inscription of Shāpūr I, which has parallel versions in Parthian and Greek. Middle Persian was also the language of the Manichaean and Zoroastrian books during the 3rd to the 10th century AD. The extant literature of the Zoroastrian books is much more extensive than that of the Manichaean texts, but the latter have the advantage of having been recorded in a clear and unambiguous script. Moreover, the Middle Persian of the Zoroastrian books, or Pahlavi, as it is usually called, does not simply represent the spoken language of the writers of the 9th-century Zoroastrian texts. It is probable that they spoke early Modern Persian and that their speech often impinged upon their writing but that they strove to write the Middle Persian of several centuries earlier as it was attested in the inscriptions of the early Sāsānian Empire when Middle Persian was the koine. By contrast, in the case of Manichaean Middle Persian, some texts survive unchanged from the 3rd century AD, the time of the Persian teacher Mani himself (AD 216–274).

Records of Parthian

Very little Parthian survives from the pre-Sāsānian period. A large number of Parthian ostraca (inscribed pottery fragments) from the 1st century BC were discovered at Nisa near modern Ashkhabad, but they are inscribed in ideographic Aramaic (*i.e.,* Aramaic writing that uses Aramaic words as symbols to represent Parthian words). Dating before the 3rd century are a document from Hawrāman, some coin legends, and a dated grave stele. The most copious and important material is the work of the Sāsānian kings of the 3rd century, who added a Parthian version to their inscriptions—Ḥājjīābād, Naqsh-e Rustam (Ka'be yi Zardusht), and Paikūla. A few decades later Parthian disappeared as a result of the rise of the Sāsānians and the predominance of their native tongue, Middle Persian. Manichaean Parthian of the 3rd century was preserved as a church language in Central Asia.

The oldest surviving Sogdian documents are the so-called Ancient Letters found in a watchtower on the Chinese Great Wall, west of Tun-huang, and dated at the beginning of the 4th century AD. Most of the religious literature written in Sogdian dates from the 9th and 10th centuries. The Manichaean, Buddhist, and Christian Sogdian texts come mainly from small communities of Sogdians in the Turfan oasis and in Tun-huang. From Sogdiana itself there is only a small collection of documents from Mt. Mugh in the Zarafshān region, mainly the business correspondence of a minor Sogdian king, Dewashtich, from the time of the Arab conquest around 700.

The relationship of the various forms of Sogdian to one another has not yet been sufficiently investigated, so that it is not clear whether different dialects are represented by the extant material or whether the differences can be accounted for by reference to other relevant factors, such as differences of script, period, subject, style, or social milieu. The importance of social milieu can be seen by comparing the elegant Manichaean literature directed to the court with the more vulgar language of the Christian literature directed to the lower classes.

Of the Saka dialect known as Tumshuq very little has survived, and despite its evidently close relationship to the much better known Khotanese dialect, full interpretation has proved difficult. Knowledge of Khotanese is more firmly based on a substantial corpus of material, including extensive bilingual texts. Although the chronological range of the extant Khotanese material is limited to only a few centuries, probably the 7th to the 10th, a rapid development of the language is apparent. At the phonological level, most noticeable is the loss of syllables between the older and later stages of the language. Thus, *hvatana-* "Khotanese" at the oldest stage is successively weakened to *hvatäna-, hvaṃna-, hvana-, hvaṃ.* At the morphological level, most striking is the tendency to simplify the case endings and even to replace them by analytical expressions, constructions of two or more words. Thus, Late Khotanese has *rakṣaysā hīya rāde* "kings of the *rākṣasas,*" whereas Old Khotanese would have *rakṣaysänu rrunde.* The Old Khotanese *-änu* ending is unmistakably genitive plural, but the Late Khotanese *-ā* is merely a general oblique plural ending and has been reinforced by *hīya* "own," used to mean "of."

Buddhist influence on Khotanese

Khotan was a great centre of Buddhism during the 1st millennium AD, and all the surviving literature in Khotanese is either Buddhist or coloured by Buddhism. Even in business documents and official letters the Buddhist background is usually not difficult to discern. It can scarcely be coincidental that the Buddhist literature of Khotan, flourishing so vigorously during the 10th century, ended abruptly with the Muslim conquest at the beginning of the 11th.

Little survives of Bactrian and Scytho-Sarmatian. Knowledge of Bactrian is based almost entirely on a single inscription of 25 lines from Āteshkadeh-ye Sorkh Kowtal in northern Afghanistan. Even less is known of Scytho-Sarmatian.

Little is also known of Old Khwārezmian; that is, Khwārezmian written in the indigenous Khwārezmian script. Apart from a few coin legends and inscriptions on silver vessels, the material that survives consists of inscriptions of the 2nd century AD from Topraq-qal'ah (Toprakkala) and of the 7th from Toqqal'ah, archaeological sites in Uzbek S.S.R. Much more is known of Late Khwārezmian, written in the Arabic script. This material is found mainly in two Arabic works, the 13th-century *fiqh* work of Mukhtār az-Zāhidī, called the *Qunyat almunyah,* and the Arabic dictionary *Muqaddimat al-Adab* of az-Zamakhshari (1075–1143/44), of which a manuscript glossed in Khwārezmian was found.

Modern Iranian. Of the modern Iranian languages, by far the most widely spoken is Persian, which, as already indicated, developed from Middle Persian and Parthian, with elements from other Iranian languages such as Sogdian, as early as the 9th century AD. Since then, it has changed little except for acquiring an increasing proportion of loanwords, mainly from Arabic. Persian has been a literary language since the 9th century, and there is an increasing awareness of the continuity of its literary tradition with the earlier periods. As the national language of Iran in succession to Middle Persian, it has for centuries strongly influenced the other Iranian languages, especially on Iranian territory. In fact, it seems likely that with the increase of modern methods of communication, Persian will eventually supplant entirely most of the other languages and dialects. Against this trend stand only Kurdish and Baluchi, the speakers of which tend to regard their languages as an expression of their particular identities. Nevertheless, even Kurdish and Baluchi have been and are strongly influenced by Persian.

Modern Iranian languages in Afghanistan and the Soviet Union

Outside Iran, the situation is rather different. In Afghanistan the first national language is Pashto, even though Persian is the official second language. Pashto became the official language by royal decree in 1936, and literary activity has been encouraged by the Pashto Ṭolana (Pashto Society) of Kābul. On Soviet territory both Ossetic and Tadzhik have received official encouragement; nevertheless, both languages will in time give way to the Russian language as the language of administration. Other languages also compete with Ossetic and Tadzhik. Ossetic became a literary language only in the second half of the 19th century, but the neighbouring Georgian has a still flourishing ancient literary tradition dating back to the 5th century AD and has many more speakers. Tadzhik, on the other hand, has a lifeline through its close connection with Persian, but it too has been retreating before Uzbek, an unrelated language of the Turkic group.

Characteristics of the Iranian languages. All Iranian languages show in their basic elements the characteristic features of an Indo-European language. Apart from the extensive borrowing of Arabic words in Modern Persian, the Iranian languages have scarcely been affected by unrelated languages, with the notable exception of Ossetic, which has been strongly influenced by the neighbouring Caucasian languages. Some dialects of Tadzhik have been very receptive to Uzbek elements. In the case of languages in contact with Indian civilization, the most noticeable non-Iranian feature often taken over is the Indo-Aryan series of retroflex sounds. These are foreign to Indo-Aryan itself, being a result of the influence of the Dravidian languages.

The elaborate phonological and morphological structure of the Indo-European parent language has been progressively simplified in the development of the Iranian languages. The basic phonological structure of Common Old Iranian has on the whole been maintained, but the morphological system has continued to be simplified. There has been a constant move in almost all Iranian languages toward an analytic structure; *i.e.,* the use of prepositions and word order rather than case endings to indicate grammatical relationships.

Development of Iranian fricative sounds

Phonology. The most characteristic features of the Iranian phonological system are those that distinguish it from the Indo-Aryan system. These are the development of various fricative sounds (indicated in phonetic symbols as x, f, *θ*, and later ɣ, β, ð), and of the voiced sibilant sounds *z* and *ž*. Even in Iranian, however, these sounds did not persist universally. In western Middle Iranian the *θ* sound was lost, and it is rare in the modern languages. In Pashto the inherited *f* sound has been discarded. Baluchi, except in the extreme east, is entirely without fricatives. Voiced bilabial and dental fricative sounds (β and ð) were recorded in some early manuscripts of Modern Persian, but they became *b* and *d* by the 13th century

Two negative features have also resulted in differentiation between Indo-Aryan and Iranian. One is the result of the coalescence in Proto-Iranian of aspirated and unaspirated voiced stops. Thus, Indo-European **b* and **bh* were maintained in contrast in Indo-Aryan as *b* and *bh,* but they fell together in Iranian as *b.* This resulted in an alteration of the phonological structure because the number of consonant contrasts (oppositions) was reduced. The other negative feature is the absence of the retroflex consonants from Iranian except as a later importation in contiguous regions.

Other divergences in development, such as the change of an *s* sound to *h* in Iranian, brought about a difference in distribution rather than in structure because *h* developed also in Indo-Aryan but from Indo-Iranian **źh* and **gh* before front vowels (*e.g., e* and *i*). The features discussed here are illustrated in Table 6.

In Old Iranian the stress lay on the next to the last syllable if it was heavy (*i.e.,* contained a long vowel or was closed by a consonant)—otherwise on the preceding syllable. With the loss of final unstressed vowels in the development of many Iranian languages, the stress often came to be on the final syllable. End stress is characteristic of Modern Persian.

Simplification of the case system

Grammar. In Old Persian the Indo-European inflectional system appears considerably simplified. In particular, the genitive and the dative coalesced into one case and the instrumental and ablative into another. Moreover, in the plural the nominative and accusative cases are not distinguished. This reduced system is still found in the Middle Iranian period in Old Khotanese and to a certain extent in Sogdian. Eastern Iranian is in this respect more conservative than western. By the Middle Iranian period, western Iranian had abandoned nominal (noun, adjective, pronoun) inflection altogether, as is the case with Middle and Modern Persian and with Parthian. In some languages, both western and eastern, two or, rarely, three cases survive. Ossetic is quite exceptional in maintaining an elaborate case system; it is partly a result of secondary, purely Ossetic developments.

Table 6: Phonetic Developments in Indo-Iranian Languages
key: NP—New Persian; Bal.—Baluchi; Yaghn.—Yaghnobi

Sanskrit	Avestan	Old Persian	modern Iranian	English translation
kratu-	xratu-	xratu-	NP xirad	"insight"
viś-	vis-	viθ-	Bal. gis	"house"
jānắti ("he knows")	zān(ā)	dān(ā)-	NP dān-	"know"
bandh-	band-	ba(n)d-	NP band-	"bind"
bhūta-	būta-		NP būd	"been"
sacā ("with, at the same time as. . .")	hačā	hačā	NP az	"from"
han-	ǰan-	ǰan-	Bal. ǰan-	"strike"
abhra-	aβra-		NP abr	"cloud"
mṛga- ("deer")	mərəɣa-		NP murɣ	"bird"
nir-ay-		nij-ay-*	Yaghn. nīž-	"go out"
pramāṇa- ("measure, authority")		framānā-	NP farmān	"command"
sthūṇā-	stunā-	stūnā-	NP sitūn	"pillar"

*In Old Persian *nij-ay-* "go out," *j* is written for *ž*, which was not represented in the script.

The elaborate conjugational system of the Indo-European verb followed a similar path to disintegration. In particular, the whole past tense system was given up by the Middle Iranian period. Only a few relics remain of the Indo-European system, such as the partial survival of the augment (a prefixed vowel or lengthening of the initial vowel) in the Sogdian imperfect tense. But a new past tense system developed, based on the old past participle, often combined with auxiliary verbs. Many languages distinguish between transitive and intransitive verbs in the past tense system; and in some, such as Khotanese and Pashto, even gender and number are distinguished.

The present tense system was far better preserved. The dual number was in retreat in Old Iranian and is not attested later. The middle voice, a form that indicates that a person or thing both performs and is affected by the action represented, was generally abandoned by the Middle Iranian period, although middle voice inflection is well represented in Khotanese. With these qualifications, the endings of the present indicative (active) have been generally well preserved. A variety of imperative, subjunctive, and optative forms, partly based on inherited forms and partly the result of innovation, is found especially in the eastern languages, including Ossetic.

Rigidity of word order is, on the whole, most characteristic of those languages, such as Persian, that have gone furthest in the reduction of the inherited morphological system.

Vocabulary. The Islāmic conquest of Iran during the 7th century entailed not only a change of religion but also a change of language. The sacred language of Islām was Arabic, and the proportion of Arabic words used in Persian rapidly increased until it reached something like the 40 to 50 percent of the present day. Before the introduction of the Arabic element, most loanwords were mainly from other Iranian languages. Most familiar is the extensive borrowing from Median found in Old Persian. In later periods, Modern Persian borrowed words extensively from Turkish and from European languages. Persian is itself the donor language in the case of the other Iranian languages, all of which have drawn upon its vocabulary.

Loanwords from languages of India

Buddhism was similarly responsible for the large proportion of Indo-Aryan words, both Sanskrit and Prākrit, found in Sogdian and especially in Khotanese. A considerable Indian element occurs in the vocabulary of those modern Iranian languages that have been or are in contact with modern Indo-Aryan languages in the northwest, such as Lahnda and Sindhi. There the Dardic languages have also been influential. Baluchi has also borrowed from Brahui, a Dravidian language spoken in Baluchistan in Pakistan.

Ossetic occupies an exceptional position. Most of its Persian and Arabic borrowings have come to it through Turkish, but more striking are the large number of words borrowed from the Caucasian languages, especially Georgian. In modern times, Ossetic continues to be influenced by Russian.

Writing systems. Iranian languages have been written in many different scripts during their long history, although various forms of Aramaic script have been predominant. Modern Persian is written in Arabic script, which is of Aramaic origin. For writing the Persian sounds *p, č, ž,* and *g,* four letters have been added by means of diacritical marks. By the addition of further letters, this Perso-Arabic script has been adapted to write not only the other main modern Iranian languages, Pashto, Kurdish, and Baluchi, but also those minor ones that are occasionally recorded. An advantage of the use of this consonantal script is that

Table 7: The Persian Alphabet

consonants					equivalents		approximate pronunciation
alone	initial	medial	final	name	EB preferred	alternatives	
ا	ا	ـا	ـا	alef	*		*
ب	بـ	ـبـ	ـب	be	b		*b*a*b*y
پ	پـ	ـپـ	ـپ	pe	p		*p*e*pp*er
ت	تـ	ـتـ	ـت	te	t		*t*ie
ث	ثـ	ـثـ	ـث	s̄e	s̄	s̲, t̲h̲	*s*and
ج	جـ	ـجـ	ـج	jīm	j	d̲j̲	*j*ob
چ	چـ	ـچـ	ـچ	che	ch	č	*ch*in
ح	حـ	ـحـ	ـح	ḥe hoti	ḥ	ḩ	*h*at
خ	خـ	ـخـ	ـخ	khe	kh	k̲h̲	Ger. Bu*ch*
د	د	ـد	ـد	dāl	d		*d*i*d*
ذ	ذ	ـذ	ـذ	z̄āl	z̄	z̲, d̲h̲	*z*one
ر	ر	ـر	ـر	re	r		*r*ip
ز	ز	ـز	ـز	ze	z		*z*one
ژ	ژ	ـژ	ـژ	zhe	zh	z̲h̲	a*z*ure
س	سـ	ـسـ	ـس	sīn	s		*s*and
ش	شـ	ـشـ	ـش	shīn	sh	s̲h̲	*sh*y
ص	صـ	ـصـ	ـص	ṣād	ṣ	ş	*s*and
ض	ضـ	ـضـ	ـض	ẕād	ẕ	z̤	*z*one
ط	طـ	ـطـ	ـط	ṭā	ṭ	ţ	*t*ime
ظ	ظـ	ـظـ	ـظ	ẓā	ẓ	z̧	*z*one
ع	عـ	ـعـ	ـع	ʿeyn	ʿ		†
غ	غـ	ـغـ	ـغ	gheyn	gh	g̲h̲, q	‡
ف	فـ	ـفـ	ـف	fe	f	fe	*f*i*f*ty
ق	قـ	ـقـ	ـق	qāf	q	ḳ	‡
ک	کـ	ـکـ	ـک	kāf	k		*k*in
گ	گـ	ـگـ	ـگ	gāf	g		*g*o
ل	لـ	ـلـ	ـل	lām	l		*l*i*l*y
م	مـ	ـمـ	ـم	mīm	m		*m*ai*m*
ن	نـ	ـنـ	ـن	nūn	n		*n*o
و	و	ـو	ـو	vāv	v	w	*v*an§
ه	هـ	ـهـ	ـه	he havaz	h		‖
ی	یـ	ـیـ	ـی	ye	y		*y*et¶

vowels, diphthongs, and special diacritical marks		equivalents		approximate pronunciation
letter or sign	name	EB preferred	alternatives	
آ	alef maddeh	ā	á	*a*rm
ىٰ	alef maqṣūreh	ā	á	*a*rm
ا	alef	*		*
و	vāv	ū		fo*o*d§
ی	ye	ī		bl*ee*d¶
ـَ	fatḥeh (or zebar)	a		m*a*p
ـِ	kasreh (or zīr)	e		b*e*t
ـُ	ẕammeh (or pīsh)	o		b*o*ne or *o*range
ی -(ـِی)	kasreh ye	ī		bl*ee*d
و -(ـُو)	ẕammeh vāv	ū		f*oo*d
یْ -(ـَی)	kasreh ye sâken	ey	ay, ai	f*a*de
وْ-(ـَو)	ẕammeh vāv sāken	ow	aw, au	b*o*ne
ـْ	sokūn (or jazm)	omit		♀
ة	tā marbūtah	eh, ah, or at		♁
ـّ	tashdīd	double consonant		me*dd*ling, etc.
ء	hamzeh	initial, omit; medial and final,ʾ		□
— ی، ـٔ	eẕāfeh	-e, -ye	-i, -yi	◇

*Initially *a* or *e*, pronounced m*a*p or b*e*t; medially and finally, *ā*, pronounced *a*rm. †A glottal stop, as in New York or Cockney "bottle." ‡A guttural *gh*; also medially and finally often softened as in French *r*ien. §As a consonant, *v*an; as a vowel, f*oo*d. ‖Generally silent in final position; otherwise *h*at. ¶As a consonant, yet; as a vowel, bl*ee*d. ♀Used to show that a consonant is not vocalized. ♁Arabic feminine ending usually not pronounced in Persian except as silent *h*. □A pause between two vowels, as in English "di-et," "qui-et." ◇A particle linking a qualifying noun or adjective with a noun; usually not written. Transliterate "-e" (with hyphen) after final consonants (except silent *h*), "-ye" (with hyphen) after silent *h* and vowels.

by not defining vowel qualities it is possible to include local dialect variations to a considerable extent.

Two modern Iranian languages spoken on Soviet territory are currently written in a modified version of the Russian alphabet: Tadzhik and Ossetic. Soviet scholars have, however, tended to use modified Latin alphabets to record the minor languages that have no literary tradition, such as some of the Pamir languages. Ossetic has also been written in the Georgian script.

Old Persian was written with a cuneiform syllabary, the origin of which is still hotly disputed. Middle Persian, Parthian, Sogdian, and Old Khwārezmian were recorded in various forms of Aramaic script. Two forms of this script as they developed for writing Sogdian were adopted by the Uighurs. In its cursive form this script spread even further, to the Mongols and Manchus. Three other scripts are important for the remaining Middle Iranian languages: Greek script for Bactrian, Arabic script for Late Khwārezmian, and varieties of Central Asian Brāhmī script of Indian origin for Khotanese and Tumshuq.

The Aramaic script was not systematically adapted to the writing of Middle Iranian; and despite the introduction of a variety of diacritical marks to differentiate letters, considerable ambiguity remained. Moreover, several letters tended to coalesce in form. In this respect, the Pahlavi script, used for writing the Middle Persian of the Zoroastrian books, developed furthest. In it, the original 22 letters of the Aramaic alphabet have been reduced to 14, which are further confused by the use of numerous ligatures (linked letters). It was the realization that this script was inadequate to record precisely the traditional pronunciation of the sacred text of the Avesta that led the Zoroastrian priests to devise the elaborate Avestan script, which, with its 48 distinct letters formed by differentiation out of the 14 used for Pahlavi, was well suited to the task. (R.E.E.)

Greek language

Greek is an Indo-European language whose history can be followed from the 14th century BC to the present day. Its documents cover a longer period of time (34 centuries) than those of any other Indo-European language. There is an Ancient phase, subdivided into a Mycenaean period (texts in syllabic script from the 14th to the 12th centuries BC) and Archaic and Classical periods (beginning with the adoption of the alphabet, from the 8th to the 4th centuries BC); a Hellenistic and Roman phase (4th century BC to 4th century AD); a Byzantine phase (5th–15th centuries AD); and a Modern phase.

Separate transliteration tables for Classical and Modern Greek accompany this article. Some differences in transliteration result from changes in pronunciation of the Greek language; others reflect convention, as for example the χ (*chi* or *khi*), which was transliterated by the Romans as *ch* (because they lacked the letter *k* in their usual alphabet). In Modern Greek, however, the standard transliteration for χ is *kh*. Another difference is the representation of β (*bēta* or *víta*); in Classical Greek it is transliterated as *b* in every instance, and in Modern Greek as *v*. The pronunciation of Ancient Greek vowels is indicated by the transliteration used by the Romans. Y (*upsilon*) was written as *y* by the Romans, indicating that the sound was not identical to the sound of their letter *i*. Modern Greek υ (*ípsilon*) is transliterated as *i*, indicating that the sound used today differs from the ancient υ. (See Tables 8 and 9 for transliterations of all the Greek letters.)

GENERAL CONSIDERATIONS

In the course of the 2nd millennium BC, groups of Greek-speaking Indo-Europeans established themselves by stages on the Greek peninsula, on most of the islands of the Aegean, and on the west coast of Anatolia; with few exceptions that is still the area occupied by the Greek language today. In the second quarter of the 1st millennium BC a vast "colonial" movement took place, resulting in establishments founded by various Greek cities all around the Mediterranean and the Black Sea, especially in southern Italy and Sicily. This extension of the linguistic area of Greek lasted only a few centuries; in the Roman period, Latin, more or less rapidly, took the place of Greek in most of these ancient colonies. "Colonial" Greek survived longest at Byzantium, as the official language of the Eastern empire.

Relationship of Greek to Indo-European. Ancient Greek is, next to Hittite, the Indo-European language with documents going furthest back into the past. At the time when it comes within view in the 2nd millennium BC, it has already acquired a completely distinct character from the parent Indo-European language. Its linguistic features place it in a central region on the dialect map that can be reconstructed for Common Indo-European; the ancient languages with which it has the most features in common are little known ones such as Phrygian or Macedonian. In the study of Indo-European dialectology, phonetic data are the most readily available and provide the most information. In this respect the position of Ancient Greek is as follows. The original Indo-European vowels of *a* and *o* quality, both short and long, remain distinct, whereas they are completely or partially confused in Hittite, Indo-Iranian, Baltic, Slavic, and Germanic. Greek is the only language that distinguishes by three different qualities (*ĕ, ă, ŏ*) the secondary short vowels resulting in certain positions from the three laryngeal sounds, $*H_1$, $*H_2$, $*H_3$, of Indo-European. (An asterisk preceding a sound or word indicates that it is not attested, but is a reconstructed, hypothetical form. For a discussion of these laryngeal sounds, see *Indo-European languages.*) Greek keeps the distinction between the original voiced stops and voiced aspirated stops of Indo-European (*e.g.,* Indo-European **d* becomes Greek *d,* and Indo-European **dh* becomes Greek *th*), whereas Iranian, Slavic, Baltic, and Celtic confuse them. Greek avoids the general shifts of stop consonants that are displayed, independently, by Armenian and Germanic, as well as the palatalization that affects guttural stops in Indo-Iranian, Armenian, Baltic, and Slavic. In these respects, Ancient Greek is conservative, as are, generally speaking, the western Indo-European languages (Italic and Celtic). On the other hand, it does show innovations. One of these, the devoicing of the original voiced stops, is shared with Italic, although it is realized in different ways (**dh*-yields Greek *th-,* Latin *f-,* Osco-Umbrian *f-*); but others are foreign to Italic: for example, the weakening of spirants and semivowels at the beginning of words before a vowel, the evolution of **s-* to *h-* (pre-Mycenaean), and **y-* to *h-* (contemporary with Mycenaean).

Comparison of the sounds of Ancient Greek with other Indo-European tongues

Morphological criteria must, of course, be taken into account in defining the position of a language. It should be noted that there are few grammatical innovations shared by Greek and Italic, apart from the extension to nouns of the pronominal ending of the genitive feminine plural *-*āsōm* (Greek *-āōn;* Latin *-ārum,* Umbrian *-aru,* Oscan *-azum*) and of the pronominal ending of the nominative masculine plural *-*oi* (Greek *-oi;* Latin *-ī*). The last innovation, however, is not shared with Osco-Umbrian, but is found instead in Germanic (in the strong declension of adjectives) and partly in Celtic. The dialectal individuality of Greek is very clearly marked in the organization of the verb (see below), which is without parallel except for an approximation in Indo-Iranian.

Greek syllabaries. Starting from a foreign script known as Linear A (used in Crete to record a native language known as Minoan), the Greeks devised, toward 1400 BC at the latest, a syllabic script to record their own language. Known as Linear B, this script was deciphered in 1952 by the British architect Michael Ventris and the British classicist John Chadwick. At present just over 100 very short Linear B inscriptions painted on vases have been found at Khaniá, Knossos, Tiryns, Mycenae, Eleusis, Orchomenus, and especially at Thebes. The major source of Linear B inscriptions are the 3,000 to 4,000 unbaked clay tablets found at Knossos (1400–1350 BC—this date has been questioned), at Thebes (probably 1350–1300 BC), and at Mycenae and Pylos (1250–1150 BC). There are no literary texts, and hardly any continuous texts (there is only a small number of real sentences); all that is currently known, and that only in part, is the accounts of the great

Mycenaean palaces and their dependencies, compiled in the Greek language.

Signs of Linear B

The Linear B syllabary consists of about 90 signs. There are signs for the vowels *a, e, i, o, u,* but these are hardly used except for initial vowels of words. There are no signs noting consonants in isolation, only signs noting consonant + vowel combinations; thus there is no sign for *t-,* but five different signs for *ta, te, ti, to, tu.* The script does not distinguish *r-* and *l-;* with the exception of *d-,* it does not distinguish between unvoiced, aspirated, and voiced stops (so the sign *ka* can be read in Greek as *ka, cha,* or *ga*). In addition, the scribes used a shorthand spelling and saved time by omissions, mainly of certain consonants (in particular, those that end syllables or words). Consequently, the spellings are often clumsy and ambiguous, such as *ka-po* for *karpos, a-re-ku-tu-ru-wo* for *alektryōn, ka-sa-to* for *xanthōi.* This inconvenient script and the nature of the documents make Mycenaean inscriptions harder to use and less rich in data than the later alphabetic inscriptions; but the information that can be gathered on the state of Greek five centuries before Homer, incomplete as it may be, is of capital importance.

Another syllabary, distantly related to Linear B, was in use in Cyprus much later (7th–3rd centuries BC) to record a native language of the island (Eteocypriot) as well as Greek.

The Greek alphabet. The Mycenaean script dropped out of use in the 12th century when the Mycenaean civilization was destroyed by the Dorian invasions. For nearly four centuries the Greeks seem to have been illiterate.

In the 8th century at the latest, starting from a Semitic model (which had separate signs for the consonants, but none for the vowels), a new system of writing was created by Greeks—the alphabet. For this purpose the list of Semitic consonants was adapted to the needs of Greek phonology, but the major innovation was the invention of five letters with the value of vowels—α(*a*), ε(*e*), ι(*i*), ο(*o*), υ(*u*). The use of the alphabet spread very quickly from east to west across the Greek world. The earliest datable inscriptions, both from around 725 BC, come from Athens (the Dipylon vase) and the colony of Ischia in the Tyrrhenian Sea (the so-called Nestor's cup).

During the period from the 8th to the 5th centuries BC, local differences caused certain details in the forms of the letters to vary from one city to another. Moreover, the primitive Greek alphabet underwent various reforms—the creation of new letters, first φ(*ph*), χ(*ch*), ξ(*ks*), and ψ(*ps*), and η(*ē*) and ω(*ō*). From the 4th century BC on, the alphabet became uniform throughout the Greek world as the result of the general adoption of the form it had taken in Asiatic Ionia.

Greek alphabetic inscriptions

Greek alphabetic inscriptions are numbered in tens of thousands: dedications, epitaphs, decrees, laws, treaties, religious rules, judicial decisions, and so forth. The majority are of Hellenistic or Roman date. The less numerous Archaic inscriptions (8th–5th centuries BC) are of particular interest for their contribution to the knowledge of the dialects (see below). It is only in Hellenistic papyri, and later in Byzantine manuscripts, that the great works of ancient literature (the originals of which have disappeared) come into view in the form of copies, some further and some less far removed from the originals.

The Greek alphabet, still in use today in Greece in the form it reached in the Hellenistic period, has enjoyed an extraordinary success as a direct or indirect model for other alphabets (notably the Latin alphabet); on it are based the writing systems employed in a great part of the modern world.

ANCIENT GREEK

History and development. Only from the 4th century BC, in the Hellenistic period, did Greek approach great unity throughout the area it covered (see below *Koine*). In the preceding ten centuries there were numerous Greek dialects, which differed in phonetic and morphological details, but which were mutually intelligible. The features shared by the local speech of different regions allow the delineation of dialect groups, of which the Greeks themselves were aware. The classifications of modern scholars modify in various ways the classifications made by the ancients, but still retain these as their basis. Among the dialects there are a West group, an Aeolic group, an Ionic-Attic group, and an Arcado-Cypriot group (the last group was neglected by the ancient Greeks because neither Arcadian nor Cypriot gave rise to a literary language). Modern scholars have never questioned the isolation of the West group, but they have tried in various ways to combine the other three into two divisions (*e.g.,* by considering Aeolic and Arcado-Cypriot as varieties of "central" Greek, or by considering Arcado-Cypriot and Ionic-Attic as varieties of "southern" Greek).

In regard to the dialects, two very different situations must be distinguished: that established for the period between the 14th and the 12th centuries BC and that for the period between the 8th and the 4th centuries BC.

In Mycenaean times the carriers of "West Greek" had not yet reached Greece; they did not irrupt into it until the end of the 2nd millennium. In continental Greece (north and south of the Isthmus of Corinth) and on certain Aegean islands (notably Crete), only varieties of Greek other than West Greek were spoken. The tablets reveal a somewhat artificial chancellery language current in the palace offices and taught as a written language in scribal schools. Based essentially on a dialect of the type that was eventually called Arcado-Cypriot, it shows great uniformity in time (during the two centuries or thereabouts covered by documents) and in space (from Knossos to Thebes). Certain fluctuations in details, however, which can be shown to vary between scribes (even at the same site and at the same date), permit the assumption that, behind this official written form of Greek, there must have been various forms of spoken Greek. The problem of the genesis of the dialects other than West Greek does not now present any provable solution.

Influence of Dorian invasions and Greek colonization

There followed two great events that upset the dialectal distribution within the Greek world. First, the Dorian invasions brought speakers of West Greek into northern Greece, then into the Peloponnese, and finally into the Aegean. Some pre-Dorian Greek populations were expelled from their homes and emigrated eastward to the west coast of Anatolia and to Cyprus. Others, who remained where they were, became more or less thoroughly Dorian in speech. It has long been thought that some of the features that Thessalian and, even more, Boeotian (both of which are Aeolic) shared in the 1st millennium with West Greek can be attributed to "recent" influences; on the other hand, some Doric dialects of the 1st millennium (*e.g.,* in Crete) show sporadic traces of features attributable to an Arcado-Cypriot substratum. The other subsequent event, which is of a different sort, was the great colonization movement that began in the 8th century BC. Each group of emigrants took with them the speech of their mother city and planted it in the new foundation. Thus there developed on the shores of southern Italy a totally new grouping of Greek dialects, side by side—Asiatic Ionic at Siris (later called Heraclea); Euboean Ionic at Rhegium and Cumae; Laconian Doric at Tarentum and Heraclea; Achaean at Sybaris, Croton, and Metapontum; Locrian at Locri Epizephyrii; and so on.

Toward the middle of the 1st millennium BC the geographical distribution of the dialects (insofar as they are known directly through inscriptions) is briefly as follows:

West Group: (1) Doric proper: Messenia, Laconia (colonies—Tarentum, Heraclea); the Argolid; the territory of Corinth (colonies—Corcyra, Anactorium, Syracuse); the Megarid (colonies—Megara Hyblaea, Selinus, Byzantium); the Sporades (colony—Cyrene); Crete; Rhodes (colonies—Gela, Acragas). (2) North-West Greek: Elis, Achaea (colonies—Ithaca, Sybaris, Croton, Metapontum); Aetolia; Phocis; Locris (colony—Locri Epizephyrii).

Aeolic Group: (1) Boeotia; (2) Thessaly; (3) Lesbos and Asiatic Aeolis.

Ionic-Attic Group: (1) Attica; (2) Euboea (colonies—Catana, Zancle, Rhegium, Cumae); (3) Cyclades; (4) Asiatic Ionia (colonies—Siris, Phocaea, foundations in Pontus [Black Sea]).

Arcado-Cypriot Group: (1) Arcadia; (2) Cyprus; (3) Pamphylia.

This linguistic situation in the first half of the 1st millennium BC resulted in literature developing on a dialect basis. The Homeric epic in the state in which it became

Table 8: Classical Greek Alphabet and Numerals

letters			name	equivalent		approximate
capital	lowercase	combinations		EB preferred	alternatives	pronunciation
Α	α, α*		alpha	a		b*o*ther
		αι		ae in proper nouns, ai in common words	e	*i*ce
		αυ		au		n*ow*
Β	β		beta	b		*b*a*b*y
Γ	γ		gamma	g		*g*o
		γγ		ng		a*ng*le
		γκ		nk	nc	i*nk*
		γξ		nx		tha*nks*
		γχ		nch	nkh	Ger. Mü*nch*en
Δ	δ, ∂*		delta	d		*d*og
Ε	ϵ		epsilon	e		b*e*t
		ϵι		ei	e or i	d*ay*
		ϵυ		eu		f*u*ry
Ζ	ζ		zeta	z		a*dz*
Η	η		eta	ē	e	d*ay*
		ηυ		ēu	eu	*you*th
Θ	θ, ϑ*		theta	th		*th*in
Ι	ι		iota	i		*e*ven or p*i*n
Κ	κ		kappa	c in proper nouns, k in common words		*k*in
Λ	λ		lambda	l		*l*i*l*y
Μ	μ		mu	m		*m*ai*m*
Ν	ν		nu	n		*n*ot
Ξ	ξ		xi	x		a*x*
Ο	ο		omicron	o		*o*bey
		οι		oe in proper nouns, oi in common words	oe	b*oy*
		ου		ou		f*oo*d

letters			name	equivalent		approximate
capital	lowercase	combinations		EB preferred	alternatives	pronunciation
Π	π		pi	p		*p*in
Ρ	ρ		rho	initial, rh; medial, r		*r*ose
		ρρ		rrh		a*rr*ow
Σ	σ‡		sigma	s		*s*and
Τ	τ		tau	t		*t*ie
Υ	υ		upsilon	y	u	Fr. r*ue*
		υι		ui		*w*e
Φ	φ, ϕ*		phi	ph		*f*i*f*ty
Χ	χ		chi	ch	kh	Ger. Bu*ch*
Ψ	ψ		psi	ps		perha*ps*
Ω	ω		omega	ō	o	b*o*ne

numerals

Greek	Arabic	Greek	Arabic	Greek	Arabic
α′	1	ιε′	15	ο′	70
β′	2	ις′	16	π′	80
γ′	3	ιζ′	17	ϙ′†	90
δ′	4	ιη′	18	ρ′	100
ϵ′	5	ιθ′	19	σ′	200
ϛ′†	6	κ′	20	τ′	300
ζ′	7	κα′	21	υ′	400
η′	8	κβ′	22	φ′	500
θ′	9	κγ′	23	χ′	600
ι′	10	κδ′	24	ψ′	700
ια′	11	λ′	30	ω′	800
ιβ′	12	μ′	40	ϡ′†	900
ιγ′	13	ν′	50	,α	1,000
ιδ′	14	ξ′	60		

*Old-style character. †Special character. ‡Final, ς.

Literature and dialects

fixed by writing displays a mixture of Aeolic and Ionic features. Choral lyric is especially Doric in colouring. Prose developed first in Ionic surroundings (Herodotus, Hippocrates), then in Attica (Thucydides, Plato). Attic is the language of dialogue in tragedy, but alongside Attic comedy there also developed in Sicily a Doric comedy. Personal poetry employs, depending on the author, Ionic (Hipponax), Lesbian (Alcaeus, Sappho), Boeotian (Corinna), and other dialects. It was only in the Hellenistic and Roman periods that Ionic-Attic became clearly dominant, though in poetry of the later periods there were artificial imitations of the early genres.

Within the alphabetical period of Ancient Greek (8th–4th centuries BC), previous to Koine, there is no break between what is termed Archaic Greek (8th–6th centuries) and what is termed Classical Greek (5th–4th centuries). The Classical period is that in which the progressive elaboration of Archaic data brought Greek literature, as well as Greek art, to perfection.

In the linguistic subdivision of Ancient Greek the effects of substratum languages played only a minor part. In their penetration into Greece toward the beginning of the 2nd millennium BC, the Hellenic peoples found earlier populations established there, about whom Greek tradition gives only vague hints, and whose language or languages are unknown. From this "pre-Hellenic" stratum, Greek vocabulary made numerous borrowings (*kyparissos* "cypress," *pyrgos* "tower," and so on), and it received from it a number of place names (*e.g.,* Korinthos); but there is no reason to think that the divergent characters of the Greek dialects (in phonetics and morphology) could be connected with different substrata. The native "barbarian" languages also had little effect on colonial Greek in the 1st millennium, and these contacts show up only in a few local borrowings.

On the other hand, there is a connection between the facts of civilization (in the political and cultural fields) and the evolution of the language. In the Mycenaean period an evident unity of civilization and the organization in the palaces of record offices and scribal schools allowed the use of a stable and uniform chancellery language. In the first half of the 1st millennium, political subdivision and rivalry between cities allowed dialectal peculiarities to strike deep roots. The special conditions in which epic developed, however, resulted in presenting the "noble" literary genres from their inception with a model of dialect mixture. From the 5th century BC onward, the prestige of Ionic and Attic literature and the political authority of Athens opened the way to Ionic-Attic predominance, and this was eventually imposed by the Macedonian conquest.

Linguistic characteristics. *Phonology.* The phonological systems of Ancient Greek differ noticeably from one period to another and from one dialect to another. The system that has been chosen to serve as an example here is that which may be attributed to Old Attic of the 7th–6th centuries BC.

In Old Attic, there are seven vowel qualities: *i,* open and closed *e, a,* open and closed *o,* and *u,* each of which has a long and a short form, except open *e* and open *o,* which have only the long form. Diphthongs originally included *ei, ai, oi,* and *eu, au, ou,* but very soon *ei* began to evolve toward long closed *ē* and *ou* toward long closed *ō.* In addition, there is a rare diphthong *ui,* and usually at the end of words the diphthongs *-ēi, -āi, -ōi,* with preponderant first elements, which later were reduced respectively to long open *ē,* long *ā,* and long open *ō.*

The consonantal structure is characterized by relative richness in stops (sounds produced by momentary complete closure at some point in the vocal tract): unvoiced *p, t, k,* aspirated *ph, th, ch,* voiced *b, d, g;* and by few spirants: only *s* and *h* sounds (*h* restricted to initial position before a vowel). There is also a voiced affricate sound, *dz;* two liquid sounds, *l* and *r;* and two nasals, *m* and *n.* The guttural nasal is not distinctive, but is only a variant of the sound *n* in front of a guttural stop. Neither *y* nor *w* occur as distinctive sounds. All of the consonants except *h* and *dz* can be doubled between vowels. The only consonant sounds normally allowed at the end of the word are *-s, -n,* and *-r.*

Word accent

Apart from some unaccented monosyllabic or disyllabic terms of minor importance, each word is marked by a rise in the musical pitch of the voice (accent) on one of the vowels (one of the last three vowels, if the word has more than three syllables). Short vowels, if they carry the accent, have only a rising tone (noted from the Alexandrian grammarians onward by the sign of the acute accent); long vowels or diphthongs may have either a rising tone (noted by the acute accent) or a rising tone followed by a falling tone (noted by the circumflex accent). Within a phrase the vowel of a final syllable with a normally rising tone is weakened in accent (noted by the grave accent). The position and nature of the accent in the word are governed by rules so strict that they do not usually permit variations that would serve to differentiate two otherwise identical forms. There are, however, examples of such a differentiation: *oîkoi* ("houses") is a nominative plural form, and *oíkoi* ("at home"), an adverb of place; *tómos* means "a cut" and *tomós* "cutting"; and so on.

The accent (which is not associated with stress) does not play any part in the rhythm of the language. This rhythm (and that of poetry, which is a stylized form of it) is based upon the distribution in the sentence (and in the verse) of short and long syllables. For a syllable to be short, it must end in a short vowel; syllables ending in a long vowel, or closed syllables (*i.e.,* those ending in a consonant), are long. The rhythm of Ancient Greek is therefore quantitative.

Morphology. Every nominal (noun) or verbal form combines a "root" that carries the sense of the word and a certain number of grammatical markers that serve principally to define the function of the word in the phrase.

The category of gender, which differentiates masculine, feminine, and neuter, is expressed only in the substantive (noun), adjective, and pronoun. The category of person (1st, 2nd, and 3rd person) is restricted to the personal pronoun and the verb. There are three numbers—singular, dual, and plural—that are distinguished in both the noun and the verb. The survival of the dual is an archaism; although a living form in the Mycenaean period, it tends to be replaced by the plural in the 1st millennium. Attic is one of the dialects in which it is best preserved down to the threshold of the Hellenistic period.

Not counting the vocative case, the Greek declension in the Mycenaean period still contained at least six cases: nominative, accusative, genitive, dative, locative, and instrumental. Between the Mycenaean period and the 8th century the locative and the instrumental ceased to exist as living cases, their functions having been taken over by the dative.

Role of aspect in the Greek verb

The most original feature of Greek morphology is the structure of the verb system, which is determined fundamentally by the category of aspect. It is organized around three principal themes (or tenses) for each verb: the "present" theme for the durative aspect, the aorist theme for the punctual aspect, and the perfect theme for the aspect of completion. Each of these themes provides an indicative and, apart from the imperative, two nonassertive moods (subjunctive and optative). The expression of time exists only in the indicative; there is, on the one hand, a future theme, and on the other, an aorist indicative theme (which always represents past time). There are also past tenses attached respectively to the present indicative (imperfect) and to the perfect indicative (pluperfect). The past tense has as its distinguishing marks a certain set of endings, called secondary (which it shares with the optative), and the presence of a special preverb called the augment. The Greek verb has two voices (active and mediopassive), which are expressed (leaving aside the aorist passive) by the opposition of two series of endings. Finally, in each voice, complete series of participles and infinitives were established corresponding to the present, future, aorist, and perfect themes.

Syntax. A relatively free word order occurs in Greek. Above all, the creation of the definite article (post-Mycenaean and post-Homeric), and the various ways in which the nominal forms of the verb (participles and infinitives) come into play, confer on the Greek sentence a suppleness unmatched in other languages.

Vocabulary. If one considers the roots of words, it seems that, although the essential basis of the vocabulary is of Indo-European origin, a fairly considerable number of terms are borrowings. Most of these loans were taken from the idioms of the populations living in Greece prior to the arrival of the Greeks; many such words had already penetrated into Greek in the 2nd millennium, for there are forms found in Mycenaean that correspond to plant names such as *elaia* "olive," *pyxos* "box tree," and *selinon* "celery"; animal names such as *leōn* "lion" and *onos* "ass"; names for objects such as *asaminthos* "bathing tub," *depas* "goblet," and *xiphos* "sword"; and names of materials such as *elephās* "ivory," *chrysos* "gold," and *kyanos* "dark blue enamel."

The most important fact is that from the verbal and nominal roots (of whatever origin) the language extracted a vocabulary full of nuances and of great scope (by using preverbs, and by forming compounds and derived words). At all periods the lexical creativity of Greek has been very active, thus giving it a vocabulary of extraordinary richness. (M.Le.)

KOINE

The fairly uniform variety of spoken Greek that gradually replaced the local dialects after the breakdown of old political barriers and the establishment of Alexander's empire in the 4th century BC is known as the Koine (*hē koinē dialektos* "the common language"), or "Hellenistic Greek." Attic, by virtue of the undiminished cultural and commercial predominance of Athens, provided its basis; but as the medium of communication throughout the new urban centres of Egypt, Syria, and Asia Minor, it absorbed numerous non-Attic elements and underwent some degree of grammatical simplification. Numerous inscriptions enable scholars to trace its triumphant progress at the expense of the old dialects, at least as the language of business and administration, although some rural dialects are reported to have survived as late as the 2nd century AD. Other sources of information for the Koine are the translation of the Septuagint made in the 3rd century BC for the use of the Hellenized Jewish community of Alexandria, the New Testament, and the writings of a few people (*e.g.,* the historian Polybius and the philosopher Epictetus) who favoured it over Attic. As the everday colloquial language of urban Egypt it may be studied in papyri going back to the 4th century BC. The Koine may be dated very crudely from the period of Alexander's conquests in the 4th century BC to approximately the reign of Justinian in the 6th century AD.

Sound changes from Ancient Greek to the Koine

The Koine replaced the Attic sound *tt* by the *ss* characteristic of Ionic and other dialects (*e.g., glōssa* for *glōtta* "tongue") at an early date, but its main phonological significance lies in its gradual simplification of the rich vowel system of Classical Greek. Ancient *ei* (ει), *i* (ι), and *ē* (η) sounds merged as *i,* and *ai* (αι) was monophthongized to *e; oi* (οι) became pronounced as the sound symbolized by *ü,* thus merging with simple *y* (pronounced as in French *tu*). The second element of *au* (αυ) and *eu* (ευ) was changed to *v* or *f* (compare ancient *autós* to modern *aftós* "he"). These shifts, combined with the loss of length distinctions, led to a new six-term system of vowel sounds: *i, u, ü, e, o, a.* The loss of *h* also belongs to the Koine period, and there is evidence that the change of the ancient aspirates and voiced stops to fricative (spirant) sounds was well under way. As a result of this latter process, Classical *ph, th, ch* (pronounced as in English "pin," "tin," "kin") acquired the fricative articulations of "fin," "thin," and the final element of Scottish "loch," (or German *Buch*); *b, d, g* became the voiced fricative sounds *v, dh* (as in "that"), and *gh* (as in Spanish *fuego*).

Grammar too began to move in the direction of Modern Greek in this period. Nouns in consonant stems acquired the endings of the *-a* declension; *e.g., thygatēr,* "daughter," accusative *thygatera,* was remodelled after items such as *khōra, khōran* "country." The dual number was lost in nouns, verbs, and adjectives, as was the optative mood of verbs. Confusion arose between the perfect and aorist tense forms, leading to the loss of one or the other (the former in most verbs).

In vocabulary there were numerous borrowings from non-Attic dialects, and Attic words acquired new meanings; thus *opsaria* "fish" and *brechei* "it rains" for Classical Greek *ichthyes* and *hyei* both occur in the New Testament (*cf.* Modern Greek *psárya, vrékhi*).

This gradual divergence from the language of Plato and Demosthenes was viewed as a species of linguistic decadence by an influential school known as the Atticists, who unceasingly castigated the use of Koine forms by writers. It was thus that there developed a rift between the everyday spoken language and an archaizing, specifically written language. It became fashionable to publish manuals of "good usage" in which the Attic equivalents of Koine innovations were recommended for the student's imitation.

BYZANTINE GREEK

During the period of the Byzantine Empire (*i.e.,* until the fall of Constantinople in 1453) the language of administration and of most writing was firmly rooted in the Atticist tradition; it is this archaizing style that is often referred to as "Byzantine Greek." The spoken language continued to develop apace, however, and its course can be followed to some extent in the writings of the less educated chroniclers (such as Malalas, 6th century) and hagiographers. Furthermore, the increasing political and military disintegration characteristic of the last few centuries preceding the fall of Constantinople brought with it a general decline in educational level, and works appeared that reflect quite closely the colloquial language of the time, although learned and pseudolearned elements are never absent. While the differences between the *Chronicle of the Morea* (13th century), for example, and present-day spoken Greek are quite minor, Byzantium failed to produce a writer of the stature of Dante, capable of establishing once and for all the living vernacular as a worthy vehicle for great literature.

Changes occurring in the Koine

Most of the phonological and grammatical developments that separate present-day Greek from the Koine occurred during this period. The frequent misuse of the dative case of nouns shows that it went out of use in the spoken language, and the infinitive was replaced by various periphrastic constructions. (Periphrastic constructions involve the use of function words and auxiliaries.) In the early period numerous words (mostly Latin) were imported: the chronicler Malalas has (in their modern form) *pórta* "door," *kámbos* "plain," *saíta* "arrow," *paláti* "palace," *spíti* "house" (from *hospitium*), and hundreds of other borrowings, not all of which have survived. The later period is characterized by the richness of its compound words, usually from native roots. Some of these continued ancient patterns, such as that in which a modifying noun is linked to its head noun by *-o-(thalassovrákhi* "sea rock," *vunópulo* "mountain lad"*)*; but coordinative compounds of the type common today are also found (*e.g.,* Modern Greek *andróyino* "man and wife," *makheropíruna* "knives and forks"). Semantic shift was another source of innovation: *álogho* "horse," previously meant "irrational"; *skiázome* "I fear," earlier was "I am in shadow"; and (*u*)*dhén* "not," was, in Classical Greek, "nothing."

MODERN GREEK

History and development. Modern Greek derives from the Koine via the local varieties that presumably arose during the Byzantine period, and is the mother tongue of the inhabitants of the Kingdom of Greece and of the Greek majority in the Republic of Cyprus. Before the exchange of populations (1923) there were Greek-speaking communities in Turkey (Pontus and Cappadocia), and

Table 9: Modern Greek Alphabet

Greek letters			name	equivalents	approximate pronunciation
capital	lower case	combinations			
Α	α, *α**		álfa	a	b*o*ther
		αι		ai	b*e*d
		αυ		av	Sl*av*, l*augh*†
Β	β		víta	v	*v*an
Γ	γ		ghámma	g before α, ο, ου, ω and consonants other than γ, ξ, and χ; y before αι, ε, ει, η, ι, οι, υ, and υι; n before γ, ξ, and χ	*w*it, *y*et, si*ng*
		γκ		initial, g; medial, ng	*g*o, fi*ng*er
Δ	δ, ∂*		dhélta	dh; d between ν and ρ	*th*en, won*d*rous
Ε	ε		épsilon	e	b*e*t
		ει		i	*e*ven
		εϊ		eï	d*ay*
		ευ		ev	l*ef*t or r*ev*el
Ζ	ζ		zíta	z	*z*one
Η	η		íta	i	f*i*g
		ηυ		iv	*e*ven, l*eaf*
Θ	θ, ϑ*		thíta	th	*th*in
Ι	ι		ióta	i	*e*ven
Κ	κ		káppa	k	*k*in, coo*k*
Λ	λ		lámbdha	l	*l*i*l*y
Μ	μ		mi	m	*m*ai*m*
		μπ		initial, b; medial, mb	*b*aby, a*mb*ush
Ν	ν		ni	n	*n*ot
		ντ		initial, d; medial, nd	*d*og, fe*nd*er
		ντζ		ntz	chi*ntz*
Ξ	ξ		xi	x	a*x*
Ο	ο		ómikron	o	s*aw*
		οι		oi	*e*ven
		ου		ou	f*oo*d
Π	π		pi	p	*p*in
Ρ	ρ		ro	r	*r*ose
Σ	σ‡		sígma	s	*s*and
Τ	τ		tav	t	*t*ie
Υ	υ		ípsilon	i initially and between consonants	*e*ven
		υι		i	*e*ven
Φ	φ, *φ**		fi	f	*f*i*f*ty
Χ	χ		khi	kh	Ger. Bu*ch*
Ψ	ψ		psi	ps	perha*ps*
Ω	ω		oméga	o	b*o*ne

*Old-style character. †Pronounced with long *a*. ‡Final, ς.

it remains the language of the Greek community of Istanbul. Certain villages in Calabria in southern Italy are also Greek-speaking. Three main varieties may be distinguished: (1) the local dialects, which may differ from one another virtually to the point of mutual unintelligibility, (2) the standard colloquial Greek spoken in all the urban centres of Greece, known as Demotic, and (3) Katharevusa (from *katharós* "pure"), a strictly literary language.

Local dialects. Of the local dialects, Tsakonian, spoken in certain mountain villages in eastern Peloponnese, is quite aberrant and shows evidence of descent from the ancient Doric dialect (*e.g.,* it often has an *a* sound for the early Greek *ā* that went to *ē* in Attic, later to *i*). The Asia Minor dialects also display archaic features (*e.g.,* Pontic *e* for ancient *ē* in certain word elements). It is not certain whether southern Italian Greek represents a survival from ancient times or was reimported there during the Byzantine period. Apart from these peripheral varieties, the modern dialects may be grouped for practical purposes as follows:

Divisions of the modern dialects

1. Peloponnesian, differing but slightly from the dialects of the Ionian isles, forms the basis of standard demotic. It shows very few specifically local innovations in its phonology, although its verb morphology is less conservative than that of the island dialects.

2. Northern dialects, spoken on the mainland north of Attica, in northern Euboea, and on the islands of the northern Aegean, are characterized by their loss of unstressed *i* and *u* and the raising of unstressed *e* and *o* sounds to *i* and *u.* Thus, standard *kotópulo* "chicken" becomes *kutóplu, émine* "he stayed" becomes *émni.* They also mark certain 1st and 2nd person plural past tense verb forms with *-an* (*ímastan* "we were," Athenian *ímaste*) and use the accusative for indirect object pronouns in instances in which the southern dialects have the genitive (*na se pó* "let me tell you," standard *na sou pó*).

3. Old Athenian was spoken in Athens itself until it became the capital of the modern state (1833), and on Aegina until early in the 20th century; it survives in Megara and in the Kími district of central Euboea. Its salient feature is the replacement of the Byzantine *ü* sound (from ancient *y, oi*) by *u* rather than normal *i;* it changes the *k* sound before the vowels *e* and *i* to *ts* and fails to contract the vowels *i* and *e* to a *y* sound before vowels (ancient *sykéa* becomes *sutséa* "fig tree," standard *sikyá*).

4. Cretan softens *k* to a *č* sound (as in "church"), *kh* to *š* (as in "she") before *i* and *e,* and *y* to *ž* (as the *s* in "pleasure")—*e.g., če* "and," *šéri* "hand," *žéros* "old man," standard *ke, khéri, yéros.*

5. The southeastern dialects of Cyprus, Rhodes, Chios, and other islands in the area soften *k* to *č* as in Crete, drop voiced fricative consonants between vowels, and retain the ancient final *-n* (*láin* "oil," standard *ládhi*). They also retain the contrast between long and short consonants (*fíla* "kiss!" but *fílla* "leaves"). As is done in Cretan and Old Athenian, they add *gh* to the *-ev-* that occurs at the end of many verb stems (*dhulévgho* "I work").

Demotic. Demotic is the language used for everyday conversation in the towns of mainland Greece, and is understood without difficulty by all speakers. Differences within the Demotic form as it is spoken in the various parts of the country are so minor that it may be regarded as the standard spoken language. In all essential respects its phonology and grammar follow average Peloponnesian practice, but it has absorbed a vast number of vocabulary items from learned sources. It is also used as the vehicle of poetry and, since the beginning of the 20th century, of fiction, although, except for certain genres such as drama and detective novels, the spoken language is not necessarily reproduced with any fidelity. Indeed, one may speak of a specifically literary demotic that differs from the spoken language in its extensive use of words culled from local dialects, its fondness for innovation in compound formation, and its somewhat eclectic verb morphology.

Use of Katharevusa

Katharevusa. Katharevusa is the purist, archaizing written language of administration; it is also used in technical publications, newspapers, and public notices. Its role in education has varied with the policies of successive governments; since 1967 it has been the official language of education beyond the elementary school. Although the concept of a distinctive written language based on earlier usage goes back to the Atticists, Katharevusa originated in the 19th century as a result of the effort to purify the local dialects of foreign elements and to systematize their morphology (inevitably on the Classical Greek model). Thus, Katharevusa is characterized by its exclusive use of Ancient Greek roots and much Classical inflection, while its syntax and idiom differ but slightly from those of Demotic (this is true, at least, of the "simple Katharevusa" current in today's newspapers and periodicals). Katharevusa elements abound in Demotic often with a specialized role; for example, *édhra* (from the Ancient Greek word for "chair") means "professorial chair," the Demotic word for "chair," *karḗkla,* being the term for the article of furniture. Many Katharevusa terms appear almost exclusively in print: *zíthos* "beer" and *ínos* "wine" are common in advertisements, but everyone says *bíra* and *krasí.* Words of Katharevusa origin often show ancient inflections: *kathiyitís* "professor" has the vocative form *kathiyitá, fititís* "student" is in the plural *fitité (cf.* Demotic *traghudhistís* "singer," plural *traghudhistés, traghudhistádhes).* Because of its use in newspapers and news bulletins, most Greeks have a good passive knowledge of Katharevusa and it is easily accessible to foreigners with reasonable competence in the classical language.

Linguistic characteristics. *Phonology.* Modern Greek has five distinct vowel sounds (*i, e, a, o, u*) and the glide *y,* most of which are indicated in Greek orthography in more than one way. The consonant sounds are:

Voiceless stops	*p*	*t*		*k*
Voiced stops	*b*	*d*		*g*
Voiceless fricatives	*f*	*th*	*s*	*kh*
Voiced fricatives	*v*	*dh*	*z*	*gh*
Nasals	*m*	*n*		
Liquids	*l*	*r*		

The sounds *f, th,* and *kh* derive from ancient aspirated consonants, and the voiced fricative sounds *v, dh,* and *gh* from *b, d,* and *g.* Modern *b, d,* and *g* are usually created by the voicing of *p, t,* and *k* after nasals; thus Ancient Greek *pénte* "five" becomes *pénde.* These also occur at the beginning of words in place of ancient nasal + stop sequences (*boró* "I am able" from *emporó*). Other important combinatory changes linking Ancient and Modern Greek include the following: (1) Ancient stop clusters and aspirate clusters both become fricative + stop; *e.g., hepta* "seven," *oktō* "eight," *ophthalmos* "eye" become *eftá, okhtó, ftarmós* ("evil eye"). (2) Double consonants are simplified except in the southeast, thus *thalassa* becomes *thálasa* "sea." (3) Nasal sounds assimilate to following fricative sounds; thus *nymphē* becomes *níffi* and then (except in the southeast dialects) it changes to *nífi* "bride." (4) The sound *l* is replaced by *r* before consonants; *e.g., adelphos* becomes *adherfós* "brother." (5) Before a vowel, *i* and *e* change to *y;* thus *paidia* becomes *pedhyá* "boys," *mēlea* becomes *milyá* "apple tree." Except for the simplification of double consonants, these statements do not apply to words of Katharevusa origin. Thus in *simfonía,* meaning "symphony" or "agreement," statements (3) and (5) are violated (the true Demotic form would be *sifonyá*).

Stress in Modern Greek

Modern Greek has dynamic stress (as in English) and not the ancient musical accent, so that while words may be distinguished by stress placement (*fíli* "friends," *filí* "kiss"), the old distinction between grave, acute, and circumflex is lost (but still represented in written accents).

Grammar. Much of the inflectional apparatus of the ancient language is retained in Modern Greek. Nouns may be singular or plural—the dual is lost—and all dialects distinguish a nominative (subject) case and accusative (object) case. A noun modifying a second noun is expressed by the genitive case except in the north, where a prepositional phrase usually replaces this. The indirect object is also expressed by the genitive case (or by the preposition *se* "to," which governs the accusative, as do all prepositions). Thus:

o yatrós	*ípe tin istoría*	*s ton adherjó tiz dhasḱlas*
"The doctor	told the story	to the brother of the teacher"
(nominative)	(accusative)	(genitive)

The ancient categorization of nouns into masculine, feminine, and neuter survives intact, and adjectives agree in gender, number, and case with their nouns, as do the articles (*o* "the," *enas* "a"). In general, pronouns exhibit the same categories as nouns, but the relative pronoun *pu* is invariant, its relation to its own clause being expressed when necessary by a personal pronoun in the appropriate case: *i yinéka pu tin ídhe to korítsi* "the woman *pu* her saw the girl" (*i.e.,* "the woman whom the girl saw").

The verb is inflected for mood (indicative, subjunctive, imperative), aspect (perfective, imperfective), voice (active, passive), tense (present, past), and person (1st, 2nd, 3rd, singular and plural). The future is expressed by a particle *tha* (from earlier *thé[o] na* "[I] want to") followed by the subjunctive. There are also two participles, an indeclinable present active one in *-ondas,* which is confined to certain individual usages (*tróghondas érkhete i óreksi* "in eating comes the appetite"), and a past passive one in *-ménos (kurazménos* "tired"*).* Formally, the finite forms of the verb (those with personal endings) consist of a stem + (optionally) the perfective aspect maker (*-s-* in active, *-th-* in passive) + personal ending (indicating person, tense, mood, voice). Past forms are prefixed by *e-* (the "augment"), although this is usually lost in mainland dialects when unstressed. In the active, the present endings are *-o, -is, -i, -ume, -ete, -un*(e) (*-usi* in southeast dialects), and the past endings are *-a, -es, -e, -ame, -ate* (Peloponnesian, standard, elsewhere *-ete*), *-an(e).* In both active and passive paradigms only six tenses are distinguishable in pronunciation. The active forms of "you (singular) write," for example, are:

	Present	Past	Imperative
Imperfective	(1) *ghráfis*	(3) *éghrafes*	(5) *ghráfe*
Perfective	(2) *ghrápsis*	(4) *éghrapses*	(6) *ghrápse*

Ghráfis "you write" represents both the present indicative and imperfective subjunctive; *ghrápsis* "you (may) write" is the perfective subjunctive, and *éghrafes* and *éghrapses* are the imperfect indicative "you were writing" and the preterite "you wrote," respectively.

Aspect in Modern Greek

Aspectual differences play a crucial role. Roughly, the perfective marker indicates completed, momentary action; its absence signifies an action viewed as incomplete, continuous, or repeated. Thus the imperfective imperative *ghráphe* might mean "start writing!" or "write regularly!" while *ghrápse* means rather "write down!" (on a particular occasion). Compare also *tha ghrápho* "I'll be writing" but *tha ghrápso* "I'll write" (once). The difference is sometimes represented lexically in English: *ákuye* "he listened" and *ákuse* "he heard." The passive forms are largely confined to certain verbs active in meaning like *érkhume* "I come," *fováme* "I am afraid," and reciprocal usages (*filyóndusan* "they were kissing"). There are also phrasal constructions representing completed action: *ékho ghrápsi* "I have written" (standard), *ékho ghramméno* (in most dialects). These are, however, much rarer than the corresponding English perfect forms. There is no infinitive; ancient constructions involving it are usually replaced by *na* (from ancient *hína* "so that") + subjunctive. Thus *thélo na ghrápso* "I want to write," *borí na ghrápsi* "he can write." Indirect statement is introduced by *oti* or *pos* (*léi oti théli* "he says that he wants").

Vocabulary. The vast majority of Demotic words are inherited from Ancient Greek, although quite often with changed meaning; *e.g., filó* "I kiss" (originally "love"), *trógho* "I eat" (from "nibble"), *kóri* "daughter" (from "girl"). Many others represent unattested combinations of ancient roots and affixes; others enter Demotic via Katharevusa: *musío* "museum," *stikhío* "element" (but inherited *stikhyó* "ghost"), *ekteló* "I execute." In addition, there are over 2,000 words in common use drawn from Italian and Turkish (accounting for about a third each), and from Latin, French, and, increasingly, English. The Latin, Italian, and Turkish elements (mostly nouns) acquire Greek inflections (from Italian *síghuros* "sure," *servitóros* "servant," from Turkish *zóri* "force," *khasápis* "butcher"), while more recent loans from French and English remain unintegrated (*spór* "sport," *bár* "bar," *asansér* "elevator," *futból* "football"). (B.E.N.)

Italic languages

Italic languages, in a broad sense, are certain Indo-European languages that were once spoken in the Apennine Peninsula (modern Italy) and in the eastern part of the Po Valley. These include the Latin, Faliscan, Osco-Umbrian, and Venetic languages, which have in common a considerable number of features that separate them from the other languages of the same area—*e.g.,* from Greek and Etruscan. (In a more narrow sense, the term Italic languages excludes Latin and denotes only Oscan, Umbrian, Faliscan, and Venetic.)

For a long time the Italic languages have been considered to be an Indo-European subfamily like Celtic, Germanic, or Slavic. Today, some scholars are inclined to distinguish within the so-called Italic branch at least three independent members of the Indo-European family: Latin (perhaps with Faliscan), Osco-Umbrian, and Venetic. They attribute the similarities—*i.e.,* the unifying phenomena in the division—to a convergence that took place when the speakers of these different idioms were integrated into the "Italic" civilization of the early first millennium BC. The culture that resulted is known as the "Etruscan koine." Figure 4 shows the assumed distribution of languages in ancient Italy; the solid line marks the Italic languages.

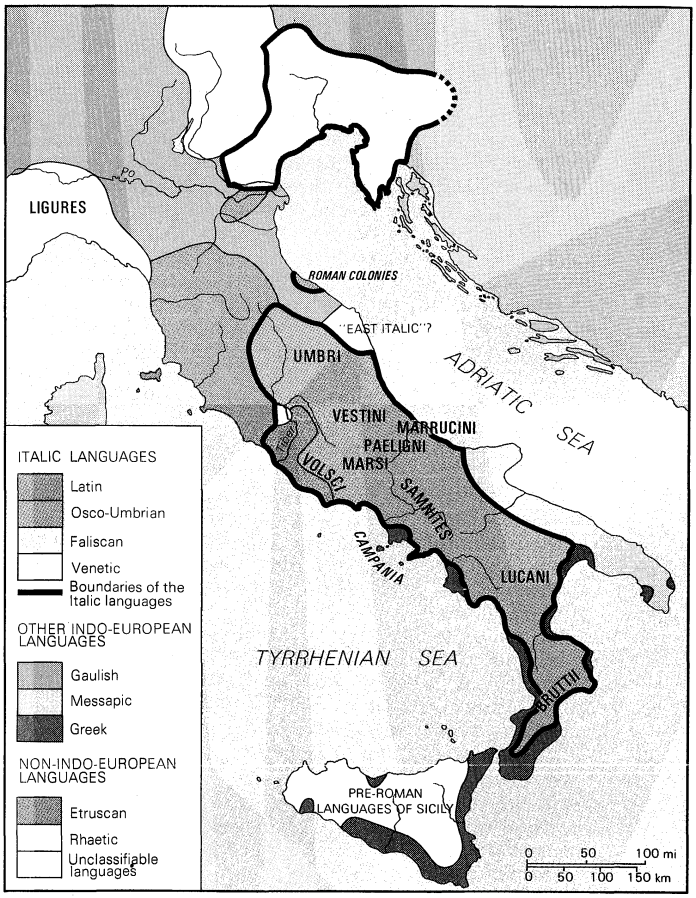


Figure 4: Supposed language areas of the Italic and neighbouring languages about 250 BC.

Languages of the group. *Latin.* Latin is the language of Latium and of Rome; its earliest known documents date from the 6th century BC. Rich epigraphical evidence and an extensive literature begin at the end of the 3rd century BC, at the time when Roman Latin was emerging as the predominant language of Italy. By AD 100 at the latest, Latin had effaced all the other dialects between Sicily and the Alps, with the exception of Greek in the colonies of Magna Graecia. (For more information about Latin and about the languages that derive from it, see below *Romance languages.*)

Spread of Latin on the Italian peninsula

The other Italic languages, Italic languages in the narrow sense, are known through local and personal names transmitted by Greek and Roman sources, and especially from inscriptions.

Oscan. Before Latin spread out, Oscan was the most widely spoken group of dialects of the Apennine Peninsula.

It was used by the Samnites in Samnium and Campania; by the inhabitants of Lucania and Bruttium; and, with slight variations, by smaller tribes between Latium and the Adriatic coast: the Volsci, Marsi, Paeligni, Vestini, and Marrucini. The legendary Sabines, who shared the earliest history of Rome, probably also spoke an Oscan dialect. The most important Oscan texts come from Campanian cities. The largest text, a treaty between Nola and Abella, is carved on a stone slab, called the Cippus Abellanus. In Bantia, a nearly unknown town of Lucania, the Tabula Bantina is preserved, the most extensive Oscan inscription. It is a bronze tablet with penal laws concerning municipal administration, written in Latin letters not earlier than 80 BC.

Umbrian. The Umbrian idiom, closely related to Oscan, is known from a few small inscriptions and from the Tabulae Iguvinae (Iguvine Tables), which consist of seven bronze tablets found at Gubbio (the ancient Iguvium). Constituting one of the largest and most important epigraphical documents of antiquity, the tablets contain ritual regulations of a sacred brotherhood to which a considerable part of the public cults of Iguvium was delegated. The Tabulae Iguvinae were incised, partly in the Umbrian alphabet and partly in Latin letters, within the last two centuries before Christ, but the text itself may result from a far more remote oral tradition.

Faliscan. Faliscan inscriptions appear only in the immediate surroundings of Falerii (the present Cività Castellana in central Italy), which, except for its dialect, seems to have been a completely Etruscan city.

Venetic. The language represented by inscriptions from the territory of the Veneti—between the Po River, the Carnic (Carniche) Alps, and Istria—is called Venetic. The majority of discoveries come from sanctuaries at Este and Làgole di Calalzo.

Alphabets

The alphabets used for writing these languages include the Greek one in Bruttium and Lucania and the Latin alphabet and various derivations of the Etruscan alphabet in the other regions. Four "national" scripts are distinguished: Oscan, Umbrian, Faliscan, and Venetic (see Figures 5–8).

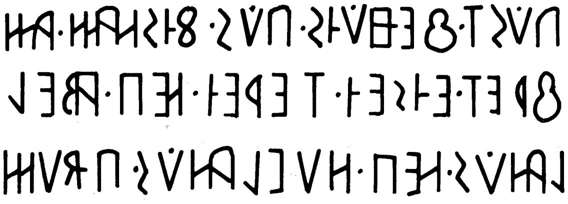


Figure 5: *Oscan.*
Inscription from the Cippus Abellanus: *púst feihúis.pús. fisnam.am/fret.eísеí.tereí.nep.abel/lanús.nep.núvlanús. pidum/ . . .* [*ú represents o*]. (Latin *Post muros qui fanum circumdant, in eo territorio neque Abellani neque Nolani quicquam* [*aedificaverint*].) "Behind the walls which go around the sanctuary,—in this area neither the inhabitants of Abella nor the inhabitants of Nola [are allowed to construct] anything."

Origin of the Italic languages. The Italic languages must have been brought from the original area of the Indo-European languages, probably in eastern parts of central Europe, when their speakers crossed the Alps. This is attested to by a stratum of very old placenames of non-Indo-European origin—*e.g.,* Tarracina, Capua, Tarentum—that covers not only the Apennine Peninsula but also Greece and Anatolia. This stratum is ascribed to a "Mediterranean" language believed to have dominated large parts of the ancient world before the arrival of the Indo-European peoples. Nothing is known about the date, the path, and the circumstances of the above-mentioned immigration, and none of the many attempts to combine archaeological evidence with linguistic prehistory has led to convincing results. Thus, the only resources available for studying the Italic languages are exclusively linguistic methods of comparative philology.

Phonology. Many of the phonetic processes that make the reconstructed Indo-European language differ from the attested Italic languages seem to have occurred rather late in time. The only one that can confidently be placed outside of Italy—that is, before the immigration over the Alps—is the change to *ss* of the combinations of the dental occlusive (stop) plus *t*. This is a common feature of Celtic, Germanic, and Latin. For example, Latin *visus* comes from the older, reconstructed form **wissos* "seen"; this is cognate with High German *gi-wiss* "surely known" and the Indo-European term with $d + t$, **widto-s.* Similarly, Oscan *nessimo-* "next" is the form equivalent to Welsh *nessaf* and Indo-European **nedh-t*(*e*)*-mo-*(An asterisk [*] before a word means that it is not attested, but reconstructed.)

The representation of the Indo-European labiovelar stop k^w is more complex. (A labiovelar stop is a sound pronounced with simultaneous articulation—movement—of the lips and the velum, the soft palate.) From this sound there has resulted a *qu* in Latin, *p* in Osco-Umbrian, *c* in Irish, and *p* in Brythonic Celtic; *e.g.,* Latin *quis* "who" is cognate with Oscan *pis* and with Indo-European *k^w *is;* and Irish *cia* is related to Welsh *pwy,* "who," which is cognate with Indo-European *k^w*ei.* Some scholars have tried to trace this development back to an Italo-Celtic unity, but the change of Brythonic k^w to *p* is surely later than the dropping of the *p* in Common Celtic. It is sounder, therefore, to assume independent processes in the different languages.

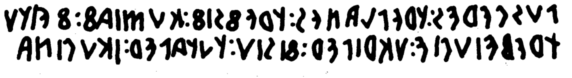


Figure 6: *Umbrian.*
Passage from the Tabulae Iguvinae: *pus veres treplanes tref sif kumiaf feitu/trebe iuvie ukriper fisiu tutaper ikuvina* [*u* represents *u* and *o*, *k* represents *k* and *g*]. (Latin *Post portam Trebulanam tres sues gravidas facito Trebo Iovio pro arce Fisia, pro civitate Iguvina.*) "Behind the Trebulan gate he shall sacrifice three pregnant sows to [the god] Trebus Iovius, for the Fisian citadel, for the state of Iguvium."

Other features developed in Italy itself—*e.g.,* the use of the voiceless dental spirant (fricative) *f* that is shared with Etruscan and is lacking in marginal districts of Venetic. In all Italic languages this *f* sound replaced the Indo-European voiced aspirated sounds in initial position. The latter are represented as *bh, dh,* g^wh and are pronounced with a small puff of air after the *b, d,* g^w. Examples of the use of *f* in Italic are as follows: Latin *frater* "brother" = Umbrian *frater* = Indo-European **bhrātēr;* Latin *facio* "I do, make" is related to Oscan *fakiiad* "he should do," to Venetic *fagsto* "he made," and to Indo-European **dhə-k-.* A more recent common process in Latin and Osco-Umbrian is the use of the full system of five short vowels in initial syllables only; short vowels of noninitial syllables in Latin became less open—*e.g., facio* "I do, make," but *in-ficio,* the compound of *in + facio.* In Osco-Umbrian these vowels tend to be lost completely—*e.g., benust* "he will have come," but *cebnust* "he will have come near." Some differences between Latin and Osco-Umbrian probably arose during the last centuries before Christ—*e.g.,* Osco-Umbrian *ō* changed to *u* (*duunated,* Latin *dōnavit* "he gave, has given"), *ē* became *i* (*ligud,* Latin *lēge* "law" in the ablative singular), and final *ā* developed into *o* (*viú* [ú in the Oscan national alphabet = o], Latin *via* "way"). Indo-European voiced aspirated sounds (*bh, dh,* g^wh) in internal position probably first became voiced spirants (*e.g.,* sounds such as *v*) in all Italic languages and, later, voiced stops in Latin and Venetic and the voiceless spirant *f* in Osco-Umbrian and Faliscan. Examples of these changes

The *f* sound

Figure 7: *Faliscan.*
Inscription on a bowl: *foied.uino.pipato.cra.carefo.* (Latin *Hodie vinum bibam, cras carebo.*) "Today I shall drink wine, tomorrow I shall have nothing."

are the voiced stop *b* in Latin *liberi* "(free) children" and Venetic *louderobos* "children" (in the dative plural) and the voiceless spirant *f* in Oscan *loufro-* "free" and Faliscan *loferta* "freed woman." The Oscan development is shown by early coins: the Greek form *allibanōn* was used for the inhabitants of the town that later was called Allifae, thus the sound *b* (later *v*), written as Greek beta, corresponding to roman *b*, became *f.*

Morphology. In contrast to the phonology, which shows so many correlations among the Italic languages, there are few definite connections between these tongues in their grammars. An innovation, probably to be ascribed to relatively recent contact between Latin and Osco-Umbrian, is the extension of the ablative singular case from *o*-stems and pronouns, where it occurred originally, to other declension classes: Latin *praidad* "with the plunder," later *praeda, meretod* "by merit," Oscan *toutad* "by the people," *slaagid* "of the border." Many of the morphological features common to Osco-Umbrian and Latin are shared by other Indo-European languages; that is, they are not Italic in a specific sense. For example, the *a*-subjunctive—*e.g.,* Latin *faciat* "may he do" and Oscan *fakiiad*—is also Celtic; passive endings in *-r*—*e.g.,* Oscan *vincter* and Latin *vincitur* "he is conquered"—are found in Celtic, Hittite, and Tocharian as well. More important are the discrepancies. For example, the genitive singular of *o*-stems shows *-ī* in Latin, Faliscan (perhaps also in Venetic), and in the Celtic languages, but *-eis* in Osco-Umbrian; the nominative plural of the same class is marked by *-oi* in early Latin, Celtic, and Greek but by *-ōs* in Osco-Umbrian, Germanic, Sanskrit, and other languages. In addition, the perfect stems of secondary verbs (verbs derived from nouns or from other verbs) are formed by *-u-* or *-v-* in Latin, by *-t(t)-* in Oscan, and by *-s-* in Venetic; *e.g.,* Latin *donavit* "he has given" = Oscan *duunated* = Venetic *donasto.*

From G.B. Pellegrini and A.L. Prosdocimi, *La Lingua Venetica* (1967)

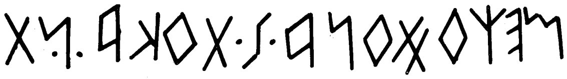


Figure 8: *Venetic.*
Inscription on a capital serving as pedestal of a votive statue found at Este: *mego donasto kanta.* (Latin *Me donavit Canta.*) "Canta gave me" ("to the goddess" is understood).

Vocabulary. Lexical comparison leads to more specific data about the history of the Italic languages. There are linguistic boundaries called isoglosses that may date back to pre-Italic history: *e.g.,* Oscan *humuns* "men" derives from the same word as *homines* and Gothic *gumans;* and Oscan *anamum* "mind" in the accusative singular form is directly related to Latin *animus* "mind, soul" and Irish *anam* "soul." There are many old differences between Latin and Osco-Umbrian. Latin *ignis* "fire" = Sanskrit *agni,* but Umbrian *pir* "fire" = Greek *pŷr* = Old English *fyr;* Latin *aqua* "water" = Gothic *ahwa,* but Umbrian *utur* "water" = Greek *hydōr* = Old English *wæter;* Latin *filius, filia* "son, daughter," but Oscan *puklo* "son" = Sanskrit *putra,* and Oscan *futir* "daughter" = Greek *thygatēr* = Gothic *dauhtar.* Adjectives of totality in Latin are *omnis, cunctus, totus,* in Osco-Umbrian *sollo-, sevo-, allo-* (cognate to English "all").

Certain lexical fields that reflect the acquisition of the Mediterranean culture show an independent terminology. The following forms strongly suggest that Latin and Osco-Umbrian speakers were not in contact with each other when they began to build cities: Latin *porta* "gate," Oscan *veru* "gate"; Latin *arx* "citadel," Umbrian *ocar* "citadel, castle"; Latin *moenia* "walls, ramparts," *murus* "wall," Oscan *feihúss* (accusative plural) "walls." On the other hand, Latins and Osco-Umbrians adopted the same terms for "write" and "read"; Latin *scribere* "to write," Oscan *scriftas* "written"; Latin *legere* "to read," Paelignian (an Oscan dialect) *lexe* "you will read." It is known that the Latin and Osco-Umbrian alphabets are derived from the Etruscan one; the spread of these terms can, therefore, be attributed to a period of Etruscan predominance. Etruscan features are obvious in archaic Italic religion; Osco-Umbrians and Veneti adopted even the Etruscan word for "god"—*ais.* Perhaps it is not by chance that many religious terms show a close community among Italic peoples; *e.g.,* the Latin forms *pius* "pious, obedient" and *piare* "to honour with religious rights" are equivalent to Volscian *pihom* (neuter singular) and Umbrian *pihatu* (imperative); Latin *feriae* "religious days" is related to Oscan *fiisiais* (ablative plural); and Latin *sacer* "sacred," *sacrare* "to consecrate, dedicate," *sanctus* "consecrated" are cognates with Oscan *sakrid* (ablative singular), *sakrafir* (subjunctive passive), *saahtum* (neuter singular).

Etruscan influence

The Etruscan supremacy ended with the foundation of local republics in Rome and in other cities of Italy in about 500 BC; when that occurred, the unifying force of Etruscan culture lost its influence. Early republican terminology developed independently; *e.g.,* Latin *consul* "consul," but Oscan *meddix* designate the first magistrate; to Latin *senatus* "senate" corresponds Oscan *kumparakineis* (genitive singular), and to Latin *comitia* "assembly," the Oscan forms *comono* or *kumbennieis.* The last period of Italic language history is characterized by an increasing influence of Roman models. For example, the title *censor* "censor" seems to have been borrowed by the Samnites in the 3rd century BC; Oscan *ceus* "citizen" is a Latin loanword that stems from a form **ceuis,* which existed in about 200 BC and was intermediate between *ceivis* and *civis;* Oscan *aidil* and *kvaisstur* imitate Latin *aedilis* and *quaestor,* terms for offices in the Roman government; and the Veneti adopted the Roman word for "freed man," *libertus.* In addition, the Tabula Bantina slavishly copied the juridical style and terminology of the Romans.

(J.U.)

Romance languages

The Romance languages, all derived from Latin within historical times, form a subgroup of the Italic branch of the Indo-European language family (see above *Italic languages*). The major languages of the family include French, Italian, Spanish, Portuguese, and Romanian; among the Romance languages that now have less political or literary significance or both are the Occitan and Rhaetian dialects, Catalan, Sardinian, and Dalmatian (extinct), among others. Of all the so-called families of languages, the Romance group is perhaps the simplest to identify and the easiest to account for historically. Not only do Romance languages share a good proportion of basic vocabulary—still recognizably the same in spite of some phonological changes—and a number of similar grammatical forms, but they can be traced back, with but few breaks in continuity, to the language of the Roman Empire. So close is the similarity of each of the Romance languages to Latin as currently known from a rich literature and continuous religious and scholarly tradition that virtually no one doubts the relationship. For the layman, the testimony of history is even more convincing than the linguistic evidence; Roman occupation of Italy, the Iberian Peninsula, Gaul, and the Balkans accounts for the "Roman" character of the major Romance languages. Later colonial and commercial contacts with parts of the Americas, of Africa, and of Asia readily explain the French, Spanish, and Portuguese still spoken in those regions.

The Roman origins of Romance

The name Romance indeed suggests the ultimate connection of these languages with Rome: the English word is derived from a French form of Latin Romanicus, used in the Middle Ages to designate a vernacular type of Latin speech (as distinct from the more learned form used by clerics) as well as literature written in the vernacular. The fact that the Romance languages share features not found in contemporary Latin textbooks suggests, however, that the version of Latin they continue is not identical with that of Classical Latin as known from literature. Nonetheless, although it is sometimes claimed that the other Italic languages (the Indo-European language group to which Latin belonged, spoken in Italy) did contribute features to Romance, it is fairly certain that it is specifically Latin itself, perhaps in a popular form, that is the precursor of the Romance languages.

Distribution of the Romance languages

By the early 1980s, at least 550,000,000 people claimed a Romance language as their mother tongue. To this number may be added the not-inconsiderable number of Romance creole speakers (a creole is a simplified or pidgin form of a language that has become the native language of

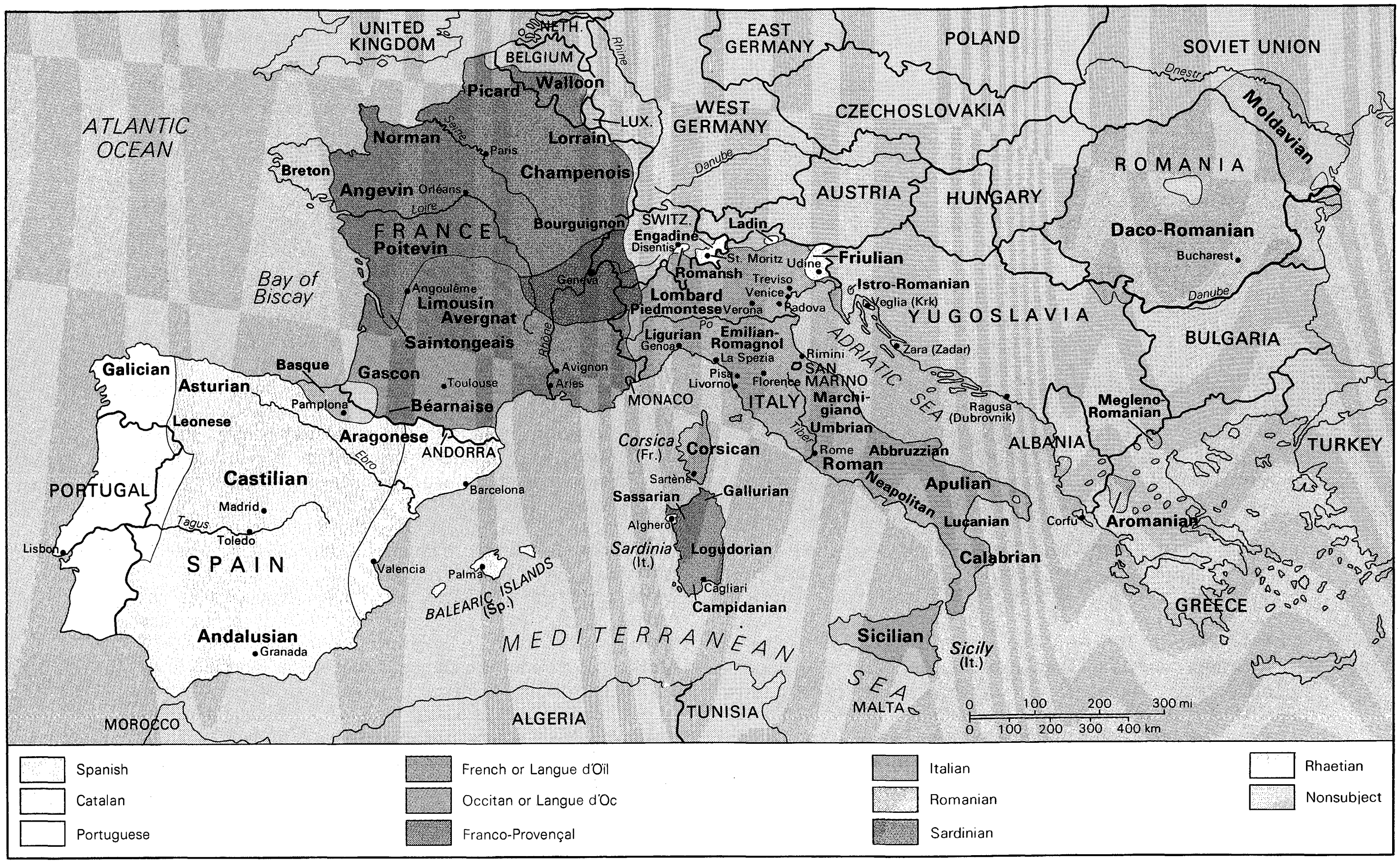


Figure 9: Distribution of Romance languages in Europe.

a community) scattered around the world. French creoles are spoken in the West Indies (with about 5,700,000 speakers), in North America (*e.g.*, Louisiana), and islands of the Indian Ocean (*e.g.*, Mauritius, Réunion, the Seychelles); Portuguese creoles in Africa, India, and Malaysia (probably fewer than 300,000 speakers); and Spanish creoles in the West Indies (more than 240,000 speak Papiamento) and the Philippines. Many speakers use creole for informal purposes and the standard language for formal occasions. Romance languages are also used formally in some countries where one or more non-Romance languages are used by most speakers for everyday purposes. French, *e.g.*, is used alongside Arabic in Tunisia, Morocco, and Algeria; it is the official language of 13 countries in West and Equatorial Africa and of the Malagasy Republic (Madagascar). Portuguese is the official language of Angola, Mozambique, and Guinea-Bissau; Italian is widely used in Somalia.

French is still widely used today as a second language in many parts of the world. Although its influence has waned before the growing popularity of English as an international language, it is still used by more than a third of the delegations at the United Nations; the wealth of French literary tradition, its precisely formulated grammar bequeathed by 17th- and 18th-century grammarians, and the pride that Frenchmen feel in their language may ensure French a lasting importance among languages of the world. By virtue of the vast territories in which Spanish and Portuguese hold sway, they will continue to be of prime importance. The beauty of the Italian language, associated with Italy's great cultural heritage, assures its popularity among students, even though territorially it has comparatively little extension. Some lesser Romance languages, such as Catalan and Romanian, retain their vitality, but others, such as Sardinian and the Rhaetian and Occitan dialects, are surely doomed to the extinction that has already overtaken a number of Romance tongues.

LANGUAGES OF THE FAMILY

Distinguishing languages and dialects

What constitutes a language, as distinct from a dialect, is a vexing question, and opinion varies on just how many Romance languages are spoken today: estimates range between five and 11. The political definition of a language—one that is accepted as standard by a nation or people—is the least ambiguous one; according to this definition, French, Spanish, Portuguese, Italian, and Romanian are certainly languages and possibly also Romansh (a national language of Switzerland since 1938 but probably related to other Rhaetian dialects spoken in Italy) and Catalan (the official language of Andorra but also widely used in parts of Spain and France). On linguistic grounds Sardinian (not the language of an independent nation since the 14th century) and Occitan (the medieval Provençal) are usually regarded as languages rather than dialects, though in modern times Occitan has grown so near to French as to be intelligible to French speakers with comparative ease. The Rhaetian dialects of Italy (Ladin in the Dolomites and Friulian around Udine) are usually regarded as non-Italian. Sicilian is different enough from northern and central Italian dialects to be given separate status often, but in Italy all neighbouring dialects are mutually intelligible, with differences becoming more marked with geographical distance. Franco-Provençal (the name given to a group of dialects spoken around the Alpine region of France and Italy) is often also assumed to be a different language from both French and Occitan, though some think it is merely a transitional dialect. Only a few persons know it in France today, though it still survives in the Italian Valle d'Aosta (where French, rather than Italian, remains the language of culture).

Judeo-Spanish

Judeo-Spanish is normally regarded not as an independent language but as an archaic form of Spanish preserving many features of the Castilian of the 15th century, when the Jews were expelled from Spain. There are possibly about 200,000 speakers, mostly originating in the Balkans and Asia Minor but, since World War II, scattered around the world; about 11,000 speakers now reside in Israel, and many live in New York City and Buenos Aires.

Some linguists believe that creoles are often different languages from their metropolitan counterparts; Haitian, for instance, is said to be mutually unintelligible with

French. Intelligibility varies so much with the speaker and the hearer, however, that it is difficult to formulate firm criteria on this basis.

The Dalmatian dialect

Many Romance dialects have virtually ceased to be spoken in the last century. Of these, Dalmatian is the most striking, its last known speaker, one Antonio Udina, having been blown up by a land mine in 1898. He was the main source of knowledge for his parents' dialect (that of the island of Veglia, or Krk), though he was hardly an ideal informant; Vegliot Dalmatian was not his native language, and he had learned it only from listening to his parents' private conversations. Moreover, he had not spoken the language for 20 years at the time he acted as an informant, and he was deaf and toothless as well. Most of the other evidence for Dalmatian derives from documents from Zara (modern Zadar) and Ragusa (modern Dubrovnik) dating from the 13th to 16th centuries. It is possible that, apart from isolated pockets, the language was then replaced by Croatian and, to a lesser extent, by Venetian (a dialect of Italian). It is certain, even from scanty evidence, that Dalmatian was a language in its own right, noticeably different from other Romance languages and presenting difficulties of classification.

On the Istrian Peninsula of the Yugoslav mainland close to the island of Veglia, another Romance language precariously survives (5,000 speakers); known as Istriot, it may be related to Vegliot, though some scholars dispute this and connect it with Rhaetian Friulian dialects or with Venetian dialects of Italian; others maintain that it is an independent language. There are no texts except those collected by linguists. A little farther north in the same peninsula, another Romance dialect, Istrio-Romanian, is threatened with extinction (fewer than 1,000 speakers). Usually classified as a Romanian dialect, it may have been carried to the Istrian Peninsula by Romanians taking refuge from the Turks in the 16th and 17th centuries and has undergone strong Croatian influence. There is evidence for its existence from a short list of words in a 1698 historical work, but it is otherwise unwritten. Another isolated Romanian dialect that may be nearing extinction is Megleno-Romanian, from a mountainous region of Macedonia, just west of the Vardar River, on the border between Yugoslavia and Greece. In 1914 there were 13,000 speakers, but many have emigrated to Asia Minor, Yugoslavia, and Romania, where small pockets survive. The only texts are those transcribed from oral traditions.

Mozarabic

Other Romance tongues earlier ceased to be spoken; there is evidence, for instance, of a form of Spanish spoken in Arab-occupied Spain until shortly after its liberation by the Spanish, accomplished at the end of the 15th century. Usually known as Mozarabic, from the Arabic word for an "Arabized person," or as ʿ*ajamī* "barbarian language," it was originally the spoken language of the urban bourgeoisie, who remained Christian while the peasantry generally converted to Islām, but it appears that many Arabs also came to use it, even though Arabic remained the only written language. Because most of the evidence, apart from a 15th-century glossary from Granada, is written in Arabic script (which uses no vowel signs), it is difficult to reconstruct the phonology of the language, but it appears to be a very conservative Spanish dialect. Much of modern information about Mozarabic comes from medical and botanical works that give Mozarabic terms alongside the Arabic. To this was added recently the discovery of Romance refrains (*kharjahs*) inserted in Arabic love ballads (*muwashshaḥs*) of the 11th and 12th centuries; study of these began only in 1946. For much of the Muslim period (beginning in 711), Christians were treated tolerantly and became culturally Arabized. Even after persecution by fanatical Muslim newcomers in the 12th century, the Mozarabs were often in conflict with Westernized "liberators" from the north. Their language died out soon after the Arabs were driven out of Spain at the end of the 15th century, though it is sometimes claimed that Mozarabic has left its mark on the dialects of southern Spain and Portugal.

Other Romance languages may have developed in peripheral regions of the Roman Empire only to die out under pressure from neighbouring non-Italic languages. Often these extinct Romance dialects are known from words borrowed into surviving languages; Berber, for instance, bears witness to the long and brilliant Roman period in North Africa that was to end in the 7th century AD with Arab invasions, and British Celtic (especially Welsh) retains many traces of what appears to have been a conservative Romance dialect, otherwise eliminated by Anglo-Saxon in about the 5th century. Albanian has so many Romance words that some style it "semi-Romance," and farther north, in what was formerly the Roman province of Pannonia (modern northwestern Yugoslavia and western Hungary), Romance speech was probably not dead at the time of the Magyar invasion at the end of the 10th century. Thus, there is reason to believe that Romance dialects may have been spoken at one time over much of southeastern Europe. It is also evident that Romance languages have been retreating south before German for some time, and it is probable that Romance tongues were used in the whole of Switzerland and parts of Bavaria and Austria until about the 9th century. Some scholars maintain that the modern Rhaetian dialects of Switzerland and northern Italy are remnants of an earlier Germano-Romance speech form.

From R.A. Hall Jr., *Introductory Linguistics;* originally published by Chilton Books, now distributed by Rand McNally & Co.

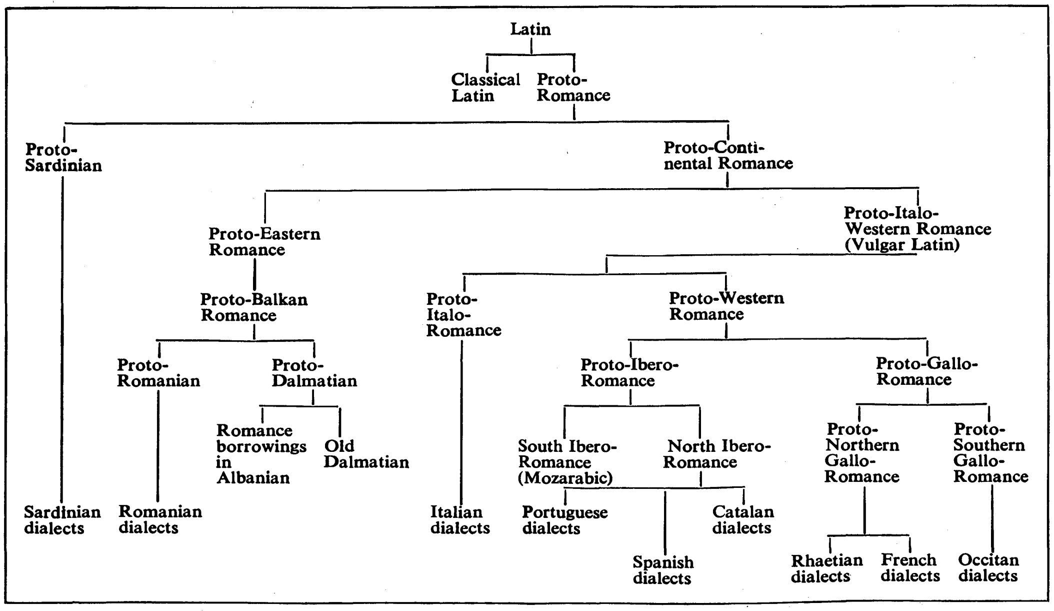


Figure 10: Derivation of Romance languages from Latin.

Classification methods and problems. Though it is quite clear which languages can be classified as Romance, on the basis primarily of lexical (vocabulary) and morphological (structural) similarities, the subgrouping of the languages within the family is less straightforward. Most classifications are, overtly or covertly, historico-geographical—so that Spanish, Portuguese, and Catalan are Ibero-Romance; French, Occitan, and Franco-Provençal are Gallo-Romance; and so on. Shared features in each subgroup that are not seen in other such groups are assumed to be ultimately traceable to languages spoken before Romanization. The first subdivision of the Romance area is usually into West and East Romance, with a dividing line drawn across Italy between La Spezia and Rimini. On the basis of a few heterogeneous phonetic features, one theory maintains that separation into dialects began early, with the Eastern dialect areas (including central and south Italy) developing popular features and the school-influenced Western speech areas maintaining more literary standards. Beyond this, the substrata (indigenous languages eventually displaced by Latin) and superstrata (languages later superimposed on Latin by conquerors) are held to have occasioned further subdivisions. Within such a schema there remain problem cases. (1) Is Catalan, for instance, Ibero-Romance or Gallo-Romance, given that its medieval literary language was close to Provençal? (2) Do the Rhaetian dialects group together, even though the dialects found in Italy are closer to Italian and the Swiss ones closer to French? Sardinian is generally regarded as linguistically separate, its isolation from the rest of the Roman Empire by incorporation into the Vandal kingdom in about 455 providing historical support for the thesis. The exact position of Dalmatian in any classification is open to dispute.

Systems for classifying the Romance languages

A family-tree classification, such as that of Figure 10, is commonly used for the Romance languages. If, however, historical treatment of one phonetic feature is taken as a classificatory criterion for construction of a tree, results differ. Classified according to the historical development of stressed vowels, French would be grouped with North Italian and Dalmatian but not with Occitan, while Central Italian would be isolated. Classifications that are not based on family trees usually involve ranking languages according to degree of differentiation rather than grouping them; thus, if the Romance languages are compared with Latin, it is seen that by most measures Sardinian and Italian are least differentiated and French most (though in vocabulary Romanian has changed most). By most nonhistorical measures, standard Italian is a "central" language (*i.e.,* it is quite close and often readily intelligible to all other Romance languages), whereas French and Romanian are peripheral (they lack similarity to other Romance languages) and require more effort for other Romance speakers to understand them. Spanish and Portuguese are even today so close in most respects that they can be regarded from a linguistic point of view as dialects of the same language, even though structural criteria would assign them to different broad classes.

Extent of mutual intelligibility

In general, it is possible to maintain that all of the standard Romance languages are to some extent mutually intelligible (especially in their Latin-based written forms) and that they have become more so during the course of history, because of much borrowing from one another and remodelling on Latin, the religious language of most speakers. In 19th-century Romania, for instance, national pride prompted a turning toward other Romance cultures, especially the French, in order more clearly to differentiate Romania from neighbouring Slavic countries, with the consequence that Romanian has become "more Romance" in vocabulary if in no other way. Local dialects in all the countries are less affected by such converging movements, but even here encroachment on the local dialect by the standard form of the language leads to the ironing out of dialect peculiarities, often ending with the loss of the local dialect, replaced by a regional variety of the standard dialect; this process is particularly evident in southern France, much less so in Italy.

Minor languages. *Occitan.* Occitan is the modern name given by linguists to the group of dialects spoken by some 13,000,000 people in the south of France (or about one-fourth of the whole French population). All Occitan speakers now use French as their official and cultural language, but their local dialects remain lively and, across most of the area, remarkably homogeneous. The name Occitan derives from the name of the area Occitanie (formed on the model of Aquitania). The medieval language is often called *langue d'oc* (from the word for "yes," compared with *langue d'oïl,* Northern French, and with the *si* languages, Spanish and Italian). In the area itself, the names Lemosí (Limousin) and Proensal (Provençal) were formerly used, but today these are often considered too localized to designate the whole range of dialects. Members of a vigorous literary movement in the Provence region, however, still prefer to call their language Provençal.

Occitan was rich in poetic literature in the Middle Ages until the north crushed political power in the south (1208–29). The standard language was, however, well established and did not really succumb before French until the 16th century, while only after the French Revolution did the French language penetrate into popular use in place of Occitan. In the mid-19th century, a literary Renaissance led by the Félibres (from an old word meaning "wisemen"), based on the dialect of the Arles-Avignon region, lent new lustre to Occitan, and a modern standard dialect was established. The most famous figure of this movement was a Nobel Prize-winning poet, Frédéric Mistral. Almost contemporaneously, a similar movement, based in Toulouse, arose and concentrated on problems of linguistic and orthographic standardization to provide a wider base for literary endeavour.

The Occitan dialects have changed comparatively little since the Middle Ages, though now French is influencing them more and more. Perhaps this influence has helped them to remain more or less mutually intelligible. The main dialect areas are Limousin, in the northwest corner of the Occitan area; Auvergnat, in the north central region of this area; northeastern Alpine-Provençal; and Languedocian and Provençal, on the west and east of the Mediterranean seacoast, respectively.

Gascon and Franco-Provençal

Gascon, in the southwest of France, is usually classified as an Occitan dialect, though to most other southerners it is today less readily comprehensible than Catalan. Some scholars claim that it has always been distinct from Occitan, because of the influence of a non-Celtic Aquitanian pre-Roman population. The Roman name of the region, Vasconia (from which the name Gascony derives), suggests the relationship of its original population with the non-Indo-European Basques. Although poets from this region used the Occitan literary language during the Middle Ages, there is evidence that their spoken language was noticeably different (the 14th century *Leys d'amor* calls it *lengatge estranh* "strange language"). Some of the region remained politically independent for a long period (the Kingdom of Béarn, which used its own standardized dialect as an official language, was not incorporated into France until 1620), and popular use of French is evident only from about 1700. Documents in the dialect are few, however; they date from about the 12th century.

Northeast of the Occitan region, along the French, Swiss, and Italian frontiers, is located a group of dialects that historically have shared most vowel developments with languages to the south and many consonant changes with those to the north. For the last 100 years claims have been made for the linguistic autonomy of these dialects, usually called Franco-Provençal; today it is estimated that somewhat fewer than 2,000,000 speakers use them (urban speakers are hard to find, and even in the countryside young speakers are few). Dialects are extremely diversified and heavily influenced by French, which has been used extensively in the area since the 13th century. Even during the Middle Ages there was no standard form of Franco-Provençal, though some 12th–13th-century documents exist. The dialect of Geneva (now extinct except in some rural communes) was the official language of the Swiss republic for some time, but otherwise none of the dialects has had official status. Some claim that a section of a manuscript, the so-called Alexander fragment, dating from the 11th–12th century and apparently part of a lost poem, is Franco-Provençal in character, but others maintain that it, like other literary texts from the region, is mainly Provençal with some French features. Since the 16th century, there has been local dialect literature, notably in Savoy, Fribourg, and Geneva.

Historic use of Catalan

Catalan. Currently spoken by about 6,200,000 people in Spain and 210,000 in France (in Roussillon), as well as by 10,000 in Andorra and 15,000 in Alghero (Sardinia), Catalan has lost little of its former lustre, even though it is no longer an important national language (as it was between 1137 and 1749, as the official language of Aragon). Although in the Middle Ages there is no evidence of dialectalization, perhaps because of the standardizing influence of its official use in the Kingdom of Aragon, since the 16th century the dialects of Valencia and the Balearic Isles, especially, have tended to differentiate from the Central (Barcelona) dialect. Nevertheless, some degree of uniformity is preserved in the literary language, which continues to flourish in spite of the little encouragement received from Madrid since the Spanish Civil War. Although there were no publications of any sort in Catalan between 1939 and 1941, and only 12 titles were published in 1946, in 1968 the number of titles published had reached 520.

The earliest surviving written materials in Catalan date from the 12th century (a charter and six sermons), with poetry flourishing from the 13th century, before which time Catalan poets wrote in Provençal. The first true Catalan poet was Ramon Llull (1235–1315), and the language remained vigorous (its greatest poet was Ausiàs March, 1397–1459, a Valencian) until the union of the Aragonese and Castilian crowns in 1474 marked the beginning of its decline. After that, although mainly grammatical works appeared, the language was to wait for its renaissance until the late 19th century. In 1906 the first Catalan Language Congress attracted 3,000 participants, and in 1907 the Institut d'Estudis Catalans was founded. Yet not until 1944 was there a course in Catalan philology at the University of Barcelona; a chair of Catalan language and literature was not founded there until 1961.

It is much disputed whether Catalan is more closely related to Occitan or to the Hispanic languages. Medieval Catalan was so close to Lemosí, the literary dialect of Occitan in southern France, that it is thought by some to have been imported from beyond the Pyrenees in the resettlement of refugees from the Moors. In more modern times, Catalan has, however, grown closer to Aragonese and Castilian, so that its family-tree classification becomes less relevant. It was occasionally called Llemosí by 19th-century Catalan revivalists, however, who wished to emphasize its independence from other Iberian tongues by stressing its relation to Occitan. Certainly, by most standards, Catalan merits the distinction of being deemed a language in its own right, and it shows little sign of decline.

Historical extent of Sardinian

Sardinian. Sardinian is currently spoken by more than 1,000,000 people, but it has many dialect differences, and there is virtually no literature, nor even a newspaper in the language (although satirical journals do appear from time to time). In earlier times the language was probably spoken in Corsica, where a Genoese dialect of Italian is now used (although French has been Corsica's official language for two centuries). Since the early 18th century Sardinia's destiny has been linked with that of the Italian mainland, and Italian is now the official language. From the 14th century till the 17th century, Catalan (at that time the official language of Aragon, which ruled Sardinia) was used extensively, especially for official purposes; a Catalan dialect is still spoken in Alghero. Castilian began to be used in official documents in 1600 but did not supplant Catalan in the south of the island until later in the 17th century. Sardinia was more or less independent from 1016, when Arab occupation was ended, until the arrival of the Aragonese in 1322, though much influenced by the Genoese and Pisans. The first documents in Sardinian are legal contracts dating from about 1080; in the north of the island Sardinian was used for such documents until the 17th century. The main dialect groupings are Logudorian (Logudorese), the central, most conservative dialect, which appears to have been used throughout the island in earlier times and which (in a northern form) provides the basis for a *sardo illustre* (a conventionalized literary language used mainly for folk-based verse); Campidanian (Campidanese), centred around Cagliari in the south, heavily influenced by Catalan and Italian; Sassarian (Sassarese) in the northwest; and Gallurian (Gallurese) in the northeast. It is sometimes said that these last two are not Sardinian dialects but rather Corsican. Gallurian in particular is related to the dialect of Sartène in Corsica, and it may have been imported into the Gallura region in the 17th and 18th centuries by refugees from vendettas.

Sardinian is unintelligible to most Italians and, with its harsh consonants and hammered accent, gives an acoustic impression more similar to Spanish than Italian. It is clearly and energetically articulated but has always been regarded as barbarous by the soft-speaking Italians; Dante, for instance, said that Sardinians were like monkeys imitating men. It retains its vitality as a "home language," but dialect diversification is such that it has little chance of development to greater prominence. Perhaps the development of the island as a tourist centre, with better communications with the mainland, will lead to the eventual decline of the language.

Rhaetian. The Rhaetian, or Rhaeto-Romanic, dialects derive their conventional name from the ancient Raetics of the Adige area, who, according to classical authors, spoke an Etruscan dialect. In fact, there is nothing to connect Raetic with Rhaetian except geographical location, and some scholars would deny that the different Rhaetian dialects have much in common, though others claim that they are remnants of a once-widespread Germano-Romance tongue. Three isolated regions still use Rhaetian.

Romansh and other Swiss dialects of Rhaetian

In Switzerland, Romansh (Rumantsch), the standard dialect of Graubünden canton, has been a "national" language, used for cantonal but not federal purposes, since 1938. The proportion of Rhaetian speakers in Graubünden fell from 39.8 percent in 1880 to 23 percent in 1970, with a corresponding increase in the Italian-speaking population, but interest in Romansh remains keen, and there are five Romansh newspapers. The main Romansh dialects are usually known as Sursilvan (or Surselvan) and Sutsilvan (or Subsilvan), spoken on the western and eastern banks of the Rhine, respectively. Another important Swiss Rhaetian dialect, Engadine, is spoken in the Protestant Inn Valley, east of which there is now a German-speaking area that has encroached on former Romance territory since the 16th century. The dialects from the extreme east and west of the Swiss Rhaetian area are mutually intelligible only with difficulty, though each dialect is intelligible to its neighbour. Sursilvan (spoken around the town of Disentis) has one text dating from the beginning of the 12th century but then nothing else until the work of Gian Travers (1483–1563), a Protestant writer. The Upper Engadine dialect (spoken around Samedan and Sankt Moritz) is attested from the 16th century, notably with the Swiss Lutheran Jacob Bifrun's translation of the New Testament. Both dialects have had a flourishing local literature since the 19th century. In many ways the Swiss Rhaetian dialects resemble French, and speakers seem to feel more at home with French than with Italian.

Ladin and Friulian

In the Alto-Adige and Dolomites area of Italy, 15,000 persons speak a language they call Ladin. Some Italian scholars have claimed that it is really an Italian (Veneto-Lombard) dialect. German is the other main language spoken in this now semi-autonomous region, much of which was Austrian until 1919. Though it is sometimes said that Ladin is threatened with extinction, it appears to retain its vitality among the mountain peasantry, as distinct from German-speaking hotel owners. Ladin newspapers are on sale in village shops, and speakers welcome visitors (there mainly to ski or climb) who show interest in their language, which is comprehensible without too much difficulty to a student of Romance languages. As it appears that these remote valleys were very sparsely populated until recently, it may even be that the number of speakers there has in fact grown. Since World War II, Ladin has been taught in primary schools in the Gardena and Badia valleys, in different conventionalized dialect forms. Although a Ladin document of the 14th century (from Val Venosta, to the west of the modern Ladin region) is known from references, the earliest written material in Ladin currently possessed dates from the 18th century, a word list of the Badia dialect. In more recent times there have been a few literary and religious texts.

In Italy, north of Venice, stretching to the Yugoslav border on the east and to the Austrian border on the north, its western boundary almost reaching the River Piave, is the Friulian (Friulan, Frioulian) dialect area, centred around Udine, with more than 500,000 speakers. This dialect is much closer to Italian than Ladin or Romansh, and it is often claimed to be a Venetian dialect. Venetian proper has gained ground at the expense of Friulian both to the east and west since the 1800s. Friulian retains its vitality today in the well-populated, industrialized region, however, and supports a vigorous local literature; its most notable poet was Pieri Zorut (1792–1867). The first text in Friulian (apart from a doubtful 12th-century inscription) is a short one dating from about 1300, followed by numerous documents in prose, as well as some poems, up to the end of the 16th century, when a rich poetic tradition began.

Creoles. The French, Spanish, and Portuguese creoles, together with their metropolitan equivalents, share many things in common. Indeed, some scholars regard them as in some sense related, either in sharing an African grammatical base, with a superimposed Romance lexicon, or in historical derivation from a Portuguese pidgin lingua franca used by colonizers and slavers, with later addition of vocabulary from metropolitan languages, such as French and Spanish, with which they came into contact. Other scholars maintain that the creoles are continuators of French, Spanish, and Portuguese in the same way as these are themselves continuators of Latin but that, under the conditions that attended the slave trade, linguistic change was exceptionally rapid, so that the origins of the creoles are often hardly recognizable.

Origin of the term creole

"Creole" is a word first found in Spanish (*criollo;* 1590), meaning a Spaniard born in the colonies or his black household servant. It most probably originated in Portuguese, in which it is related to such words as *criança* "child" and *criada* "maid-servant," generally indicating a household dependent, although the word *crioulo* is not known until 1632. Today, "creole" has come to indicate a pidgin or trade language that has become the mother tongue of a population, often black; the conditions under which this has happened have included forcible transplantation and intermingling of people with mutually unintelligible native languages and imposition of the master's language, during the slave-trade era.

Of Romance creoles used today, French creoles are most widespread. In Haiti, for instance, there are more than 5,000,000 creole speakers, of whom only about 10 percent know French; the island of Santo Domingo (Hispaniola), of which the eastern half forms the Haitian Republic, was settled in 1665. The Lesser Antilles (Martinique, Guadeloupe, Dominica, St. Lucia, St. Kitts, etc.) were colonized in 1635, and many still use French creoles, even when change of ownership led to the imposition of English as the official language. French creoles are also used in French Guiana and, though dying out, in Louisiana. In all, almost 6,000,000 speakers use French creoles in the Americas. On islands of the Indian Ocean, too, French creoles are spoken; in Mauritius (600,000 speakers), owned by France from 1715 to 1810, the creole retains its hold as a lingua franca even though English is the official language and though a large part of the population use Indian dialects as home languages. In the Seychelles (65,000 speakers), owned by France from 1768 until 1814, when they became British, and in Réunion (originally L'Île Bourbon; 485,000 speakers), where French is still the official language, French creoles are still in use. Some French-creole speakers claim that creoles from other far-off regions are easily intelligible to them. Others contest this, however, pointing out that the creole used by educated speakers is often heavily larded with standard French on all but very informal occasions. Certainly, the linguist can easily discern similarities, especially in grammatical structure, that make the various French creoles seem more like each other than like standard French.

Portuguese creoles were purportedly once widely used in Asia, though probably more frequently as trade languages than as mother tongues. They survive today in Macau and Hong Kong and to some extent in Malaysia and Goa, India. In Africa a Portuguese creole is used by more than 260,000 people in Guinea-Bissau, the Cape Verde Islands, and some Gulf of Guinea islands (Annobón in Equatorial Guinea, where it is losing ground to Spanish, and São Tomé and Príncipe, where, of the 86,000 inhabitants, about two-fifths speak creoles). In South America a Brazilian creole is still used in the interior, although at one time this language was more widespread in the country (spreading even to Surinam, where Portuguese Jews and their slaves fled from Brazil in the 17th century).

Papiamento

Papiamento, spoken by more than 240,000 people on the islands of Aruba, Curaçao, and Bonaire in the Netherlands Antilles, is today classed as a Spanish creole, though some claim that it was once a Portuguese creole that later acquired many new words from Spanish. A Spanish creole also survives precariously in the Philippines among descendants of mixed Spanish–Filipino stock.

On the whole, creoles are rarely used for literature, except satirical and comic pieces. Most speakers regard them as "bad" versions of the standard language and in formal situations try to improve their usage on the model of the standard, though they admit to feeling more relaxed speaking natural creole. Sometimes "purer" creole speech can be heard among speakers not as much exposed to the standard form of the language, as in West Indian islands where English is the official language, while a French creole is the home language.

Major languages. *French.* Probably the most internationally important of the Romance languages, French is used as the official language in 21 countries and as a co-official language in several more (including Algeria, Belgium, Canada, Luxembourg, Switzerland, Jersey). In France and Corsica about 44,000,000 use it as their first language; in Canada, 6,250,000; in Belgium, 3,200,000; in Switzerland (cantons of Neuchâtel, Vaud, Genève, Valais, Fribourg), more than 1,000,000; in Monaco, 13,000; in the Italian Valle d'Aosta, 100,000; and, in the United States (especially Maine and Louisiana), about 2,400,000. Moreover, about 5,000,000 Africans and 4,000,000 Indo-Chinese use it as their principal international language; many creole French speakers, too, use standard French in formal situations.

Standard French is based on the dialect of Paris (in the so-called Île de France with its Francien dialect), which assumed importance in about the second half of the 12th century; it was basically a north central dialect with some northern features. Before that, other dialects, especially Norman (which developed in Britain as Anglo-Norman, widely used until about the 14th century), and northern dialects, such as Picard, had more prestige, especially in the literary sphere. The Edict of Villers-Cotterêts (1539), however, established Francien as the only official language, as against both Latin and other dialects. From then on, standard French began to smother local dialects, which were officially discouraged until recent times, though the standard language did not spread to popular usage in all regions until well into the 19th century. Dialectal features, still admired and cherished by 16th-century writers, were ridiculed in the 17th and 18th centuries, when the grammar and vocabulary of the modern language were standardized and polished to an unprecedented degree.

Linguistic change in French

Linguistic change was more rapid and more drastic in northern France than it was in other European Romance regions, and influence from Latin was comparatively slight (though borrowing of Latin vocabulary has been great since the 14th century). The influence of the Germanic Frankish invaders is often held to account for exotic features in Old French, such as strong stress accent and abundant use of diphthongs and nasal vowels; but the change in about the 15th century to a more sober (even monotonous) intonation and loss of a stress accent on each word can hardly be attributed to influences from neighbouring languages. The popularity of French as a first foreign language, in spite of numerous pronunciation difficulties for nearly all foreign speakers, is perhaps as much the result of the precise codification of its grammar, effectuated especially during the 18th century, as of the brilliance of its literature at all periods. Thus, although Italian would be an easier language for them to master, most foreign speakers of Romance languages become acquainted with French at school and often retain an affection and admiration for the language.

The first document apparently written in French purports to date from 842; known as the Strasbourg Oaths, it is a Romance version of an oath sworn by two of Charlemagne's grandsons. Some claim that the text is thinly disguised Latin constructed after the event to look authentic, for political propaganda purposes; others suppose that its Latinizing tendencies reveal the struggle of the scribe with the problems of spelling French as it was spoken at the time. If the language of the Strasbourg Oaths is Northern French, it is difficult to decide what dialect it represents—some say that of Picard, others Franco-Provençal, and so on. The second existing text in Old French is a short sequence of the writings of St. Eulalia, precisely dated (AD 880–882) and localized (Valenciennes); it is definitely Picard in character. Two 10th-century texts (the *Passion du Christ* and the *Vie de St. Léger*) seem to mingle Northern and Southern dialect features, while another (the "Jonas fragment") is obviously from the far north. After that the Norman dialect seems to dominate literature, though here it is probable that the language is better described as a standard language with elimination of gross dialect features. Two manuscript traditions—one from the far north and one from the west—seem to have developed, and it is possible that the intersection of the two produced the Francien dialect that was eventually to reign supreme.

Modern dialects of French

Modern dialects are classified mainly on a geographical basis, and most survive only in the peasant speech. Walloon, a dialect spoken mainly in Belgium, is something of an exception in that it has had a flourishing dialect literature since about 1600. Other dialects are grouped as follows:

Central: Francien, Orléanais, Bourbonnais, Champenois
Northern: Picard, Northern Norman
Eastern: Lorrain, Bourguignon (Burgundian), Franc-Comtois
Western: Norman, Gallo (around the Celtic Breton area), Angevin, Maine
Southwestern: Poitevin, Saintongeais, Angoumois.

Outside France, apart from the creoles, the French of Canada, originally probably of Northwestern dialect type, has developed the most individual features. Although in the 18th century Canadian French was regarded as exceptionally "pure" by metropolitan commentators, it began to diverge from Parisian French as English influences took over from the French after 1760. It is less clearly articulated, with less lip movement and with a more monotonous intonation than standard French; some change in consonantal sounds occurs (*t, d* shift to *ts, dz,* respectively, and *k* or *g* followed by *i* or *e* become palatalized [pronounced with the tongue touching the hard palate, or roof of the mouth]); nasal vowels tend to lose the nasal element; vocabulary and syntax are heavily anglicized. Though intellectuals turn toward France for cultural inspiration (some university-educated French Canadians may not even know English), the pronunciation and usage of standard French is sometimes derided by French Canadians; this may be because their English compatriots are taught Parisian at school. The French-speaking population of Canada is growing relatively fast, and at present 80 percent of the population of Quebec Province use French as their normal language. Even today, however, French is not as socially prestigious as English; the activities of the separatist movement are evidence of the feeling of grievance that many French Canadians still have.

Spanish. Spanish, the Romance language spoken as a first language by the most people in the world, is the official language of 18 American countries as well as that of Spain, and, though many South and Central Americans use native Indian languages as their first language, Spanish is spreading and achieving continuing educational progress. Estimated numbers of speakers are as follows (in order of numerical importance): Mexico 55,000,000; Spain 27,000,000; Colombia, 26,000,000; Argentina, 24,000,000; the United States, 9,000,000; Dominican Republic, 5,000,000; El Salvador, 4,500,000; Guatemala, 3,500,000; Puerto Rico, 3,200,000; Honduras, 3,000,000; Uruguay, 2,500,000; Nicaragua and Paraguay, 2,400,000 each; Bolivia, Costa Rica, and Panama, 2,000,000 each; and the Philippines, 900,000. There are also a few hundred thousand Judeo-Spanish speakers and about 160,000 Spanish speakers in Africa.

The dialect spoken by nearly all these speakers is basically Castilian, and indeed Castellano is still the name used for the language in several American countries. In the north of Spain two other Spanish dialect groups (Asturo-Leonese and Aragonese) survive but seem doomed to extinction. The dialect of the Northwest (Galician, or Gallego) is properly a Portuguese dialect, and, on the east, Catalan can be held to be a different language, as noted above. The now-unchallenged ascendancy of Castilian among Spanish dialects is the result of the particular circumstances of the Reconquista (the conquest of Moorish Spain by the Spanish, completed in 1492), with which the language went south. Having established itself in Spain, the Castilian dialect, possibly in its southern, or Andalusian, form, was then exported to the New World during the 16th century.

The standard Spanish dialect, Castilian

Standard Castilian is no longer the language of Old Castile, which was already regarded as rustic and archaic in the 15th century, but a modified form developed in Toledo in the 16th and 17th centuries and, more recently, in Madrid. American countries have developed their own standards, differing mainly in phonology (in which they often agree with the southern Spanish dialects) and in vocabulary (in which loanwords from English are more frequent), but differentiation is comparatively slight, and some Americans still regard true Castilian as their model. On the whole, American forms of Spanish are more musical and suave than the harsh Castilian of Madrid, but it is remarkable how little deformation, or creolization, of the language has occurred, even in the mouths of uneducated Indian speakers.

The first texts in Spanish consist of scattered words glossing two Latin texts of the 10th century, one from Rioja and the other from Castile; the language in the two documents shows few dialect differences. Another document, dating from about 980, seems to be Leonese in character. The Mozarabic verse forms known as *kharjahs* are the next-oldest surviving texts, but by the middle of the 12th century the famous epic poem *El cantar de mío Cid* ("The Song of the Cid") appeared in a language that is basically Castilian. Literary works in Leonese appear till the 14th century and in a conventionlized Aragonese till the 15th century, but Castilian was destined from the first to gain the upper hand, even making an impact on Portuguese, especially in the 15th and early 16th centuries.

Judeo-Spanish (Jewish-Spanish, Sefardi, Ladino) is the continuation of an archaic form of Castilian, reflecting the state of the language before 16th-century standardization. The expulsion of the Jews from the Iberian Peninsula in 1492 affected mainly the humbler classes, with the rich preferring "conversion," but the latter often later chose voluntary exile to settle in England and Holland, where their Sefardic tongue precariously survives as a religious language in a few communities. Earlier refugees fled to the Middle East and, once settled, continued to produce learned works in a literary archaic form of their language, Ladino, written in an adapted Hebrew script. The spoken dialects have differentiated considerably from Ladino, mainly by borrowing from Hebrew and local languages, and, after further dispersion during and after the World War II, these dialects are now threatened with extinction, though Ladino survives with a mainly religious function.

Portuguese. Portuguese owes its importance largely to its position as the language of Brazil, where more than 110,000,000 people speak it. In Portugal itself there are about 10,000,000 speakers. The Galician (Gallego, Galego) dialect of northwestern Spain is historically a Portuguese dialect, though now much influenced by the standard Castilian Spanish; about 3,000,000 speakers use Galician as their home language. It is estimated that there are also about 400,000 Portuguese speakers in Africa (some of whom also use creole) and about 500,000 in the United States, with a few thousand in former Portuguese possessions, such as Goa, Mozambique, and Guinea-Bissau.

Portuguese dialect groups

There are four main Portuguese dialect groups, all mutually intelligible: (1) Northern, or Galician; (2) Central (Beira); (3) Southern (Estremenho, including Lisbon, Alemtejo, and Algarve); and (4) Insular, including the dialects of Madeira and Brazil. Standard Portuguese was developed in the 16th century, basically from the dialects spoken between Lisbon and Coimbra, to the north. Brazilian (Brasileiro) differs in several respects, in syntax as well as phonology and vocabulary, but many writers still use an academic metropolitan standard. A creolized form, once widespread in Brazil, seems now to be dying out. A Jewish Portuguese is attested in 18th-century Amsterdam and Livorno (Leghorn, Italy), but virtually no trace of this remains today.

Portuguese speakers have little difficulty in understanding and speaking Spanish, in spite of considerable acoustic and grammatical differences between the two languages. In Portugal, however, they often show considerable resentment at being addressed in Spanish, and there are signs of social resistance to this neighbouring tongue, perhaps because Portugal has so frequently had to play a subordinate role to Spain in the course of its history. In the region of northwestern Spain that adjoins Portugal, the Galician dialects lack uniformity and are closer to Spanish. Even in Castile, where standard Spanish (Castilian) originated, Galician was the conventional language of the courtly lyric until about 1400, but it lost ground in the 15th century, and Castilian replaced Galician as the official language of Galicia in 1500. Dialect poetry in Galician has flourished from the 18th century, with an upsurge in the 19th century.

Before the reconquest of Moorish Spain, largely completed in the 13th century, Galician and Portuguese were indistinguishable. The first evidence for the language consists of scattered words in 9th–12th-century Latin texts; continuous documents date from about 1192, the date assigned to an extant property agreement between the children of a well-to-do family from the Minho Valley. Literature began to flourish especially during the 13th and 14th centuries, when the soft Gallego-Portuguese tongue was preferred by courtly lyric poets all over the Iberian peninsula except in the Catalan area. In the 16th century, Portugal's Golden Age, Galician and Portuguese grew further apart, with the consolidation of the standard Portuguese language.

Distribution of Italian speakers

Italian. Italian is currently spoken by more than 65,000,000 people, of whom the vast majority live in peninsular Italy (including the Republic of San Marino), with about 5,000,000 in Sicily and 1,500,000 Italian speakers in Sardinia. France, including Corsica, has about 1,200,000 Italian speakers and Switzerland about 450,000; there are, in addition, about 300,000 in Yugoslavia. For a large, if decreasing, proportion of these speakers, standard Italian is not the language of the home, where dialectal forms are used. Overseas (*e.g.,* in the United States, where it is estimated that there are 4,500,000 Italian speakers; in Argentina, with 1,300,000; and in Brazil, with about 500,000), speakers sometimes do not know the standard language and use only dialect forms. A speaker of an Italian dialect, even one as superficially different as Sicilian, can with effort understand standard Italian, however, and can even teach it to himself by such means as listening to radio programs. For most Italians their first contact with the standard language comes in primary school, in which until recently it was the only dialect used; standard Italian is virtually the only dialect of culture in modern Italy, and with immigration from the south to the industrial north it is becoming more and more the language of intercommunication. Standard Italian is widely used in Malta and Somalia. In Libya and Ethiopia it is now dying out of use.

Standard Italian began to be developed in the 13th and 14th centuries as a literary dialect. At first basically a Florentine dialect, stripped of local peculiarities, it has since acquired some characteristics of the dialect of Rome in particular and has always been heavily influenced by Latin. It overlies a wide variety of dialects, sometimes considered to represent a fundamental differentiation between northern and southern Italy that dates from Roman times. Today, however, these variant dialects form a continuum of intelligibility, although geographically distant dialects may be radically different. The northern dialects include what are often called the Gallo-Italian dialects (Piedmontese, Lombard, Ligurian, Emilian-Romagnol), in which some linguists discern the influence of a Celtic (Gaulish) substratum (*i.e.,* the traces of a language previously spoken in the region). The other northern group of dialects, spoken in northeastern Italy, is called Venetan (including Venetian, Veronese, Trevisan, and Paduan dialects, etc.). Istriot, a language spoken in Yugoslavia, is sometimes considered yet another northern Italian dialect, rather than an independent language. The Tuscan dialects (including those of Corsica) are often held to form a linguistic group of their own, while in the south and east three broad dialect areas are grouped loosely together: (1) the dialects of the Marche (Marchigiano), Umbria, and Rome; (2) Abruzzian, Apulian, Neapolitan, Campanian, and Lucanian; and (3) Calabrian, Otrantan, and Sicilian, believed by some to be influenced by the Greek once spoken there (which still survives in isolated pockets on the toe and heel of the peninsula).

Italian dialect regions

Outside Italy, Italian dialects are heavily influenced by contact with other languages (English in New York; Spanish in Buenos Aires). A pidgin Italian can still be heard in Addis Ababa but has little extension. Relics of a Jewish Italian survive within Italy; a colony of 6,000 Jews, who used a Venetan dialect as a home language in Corfu, was exterminated during World War II.

Early texts from Italy are written in dialects of the language that only later became standard Italian. Possibly the very first text is a riddle from Verona, dating from perhaps the 8th century, but its interpretation is obscure and its language Latinized. More surely Italian are some 10th-century documents from Montecassino, after which there are three Central Italian texts of the 11th century. The first literary work of any length is the Tuscan *Ritmo Laurenziano* ("Laurentian Rhythm") from the end of the 12th century, followed soon by other compositions from the Marches and Montecassino. In the 13th century, lyric poetry was first written in a conventionalized Sicilian dialect that influenced later developments in central Italy.

In modern Italy, although dialects are still the primary spoken idiom, standard Italian is virtually the only written language and, with the spread of literary regional varieties of the standard language, may eventually replace the dialects. Neorealism, especially in the cinema, has introduced a limited use of dialect into cultural media, but the relevance of such a development is hotly debated.

Romanian. There are about 23,500,000 speakers of Romanian (or Rumanian), of whom about 20,000,000 live in the Socialist Republic of Romania, 3,000,000 in the U.S.S.R., and about 300,000 in Yugoslavia, Bulgaria, Greece, and Albania. There are about 60,000 Romanian speakers in the United States.

Standard Romanian

The standard language of Romania is based on a Walachian variety of so-called Daco-Romanian, the majority group of dialects; it was developed in the 17th century mainly by religious writers of the Orthodox Church and includes features from a number of dialects, though Bucharest usage now provides the model. Daco-Romanian is fairly homogeneous but shows greater dialectal diversity in the Transylvanian Alps, from which region the language may have spread to the plains. Moldavian, the variety of Romanian spoken in the U.S.S.R., is still written in a form of Cyrillic script, and some claim that it is a language in its own right, though most Western linguists see it definitely as a variant of Daco-Romanian. Other dialects of Romanian are barely mutually intelligible with the standard, and some can be counted as separate languages; these include Megleno-Romanian (Vlaši) and Istrio-Romanian, both already mentioned as nearly extinct. More vigorous is the Aromanian, or Macedo-Romanian, group of dialects scattered throughout Greece (60,000 speakers in the early 1980s), Yugoslavia (150,000), Albania (16,000), and Bulgaria (4,000). Numbers have probably decreased considerably, but certainly before the war Aromanians were often prominent businessmen in their localities. The first known inscription in Aromanian, dated 1731, was found only in 1952 at Ardenita, in Albania; texts date from the end of the 18th century, and literary texts have been published in the 19th and 20th centuries (mostly in Bucharest).

The first known Daco-Romanian text is a letter dated 1521, though some manuscript translations of religious texts show Transylvanian dialect features and may be earlier. The vast majority of early texts are written in Cyrillic script, the Roman (Latin) alphabet having been adopted in 1859 at the time of the union of Walachia and Moldavia. Literature in Romanian began to flourish in the 19th century, when the emerging nation turned toward other Romance countries, especially France, for cultural inspiration. Today, in spite of efforts at industrialization, peasant life continues almost unchanged in many regions, and the linguistic standards of the capital have little impact; an autonomous Magyar-speaking (Hungarian) region in the middle of Romania adds further complication to the linguistic situation.

HISTORICAL SURVEY

Latin and the Italic languages

Latin and the protolanguage. Latin is traditionally grouped with Faliscan among the Italic languages, of which the other main member is the Osco-Umbrian group. Oscan was the name given by the Romans to a group of dialects spoken by Samnite tribes to the south of Rome. It is well attested in inscriptions and texts for about five centuries before Christ and was used in official documents until *c.* 90–89 BC. The absence of great dialectal variations in the texts suggest that they are written in a standardized form. In early times, Umbrian was spoken northeast of Rome, to the east of the Etruscan region, possibly as far west as the Adriatic Sea at one period. It is attested mainly in one series of texts, the Tabulae Iguvinae, dated from 400 to 90 BC, and it is similar to Oscan. Probably Latin and Osco-Umbrian were not mutually intelligible; some claim they are not closely related genetically but that their common features arose from convergence as a result of contact.

The Roman dialect was originally one of a number of Latinian dialects, of which the most important was Faliscan, the language of Falerii, 32 miles (51 kilometres) north of Rome. The Faliscans were probably a Sabine tribe that early fell under Etruscan domination. The dialect is known mainly from short inscriptions dating from the 3rd and 2nd centuries BC and probably survived until well after the conquest of Falerii by the Romans in 241 BC. It shares one phonetic feature with Osco-Umbrian (medial *f* from Indo-European **bh* [the asterisk marks a hypothetical reconstructed form] when Latin has *b—e.g.,* Faliscan *carefo* = Latin *carebo*), but other are like Latin (*e.g.,* Faliscan *cuando* = Latin *quando* = Umbrian *pan(n)u*). Some Latin diphthongs, however, appear as simple vowels in Faliscan (*e.g.,* Latin *ae* = Faliscan *e*), and Latin final consonants are often absent (*e.g.,* Faliscan *cra* = Latin *cras*).

The earliest Latinian text is an inscription on a cloak pin (fibula) of the 6th century BC, from Palestrina (Praeneste); the inscription is definitely dialectal and seems to have Oscan features (*e.g.,* a reduplicated syllable in the perfect form—*fhefaked* = Latin *fecit* "he did, made"). Other Latinian inscriptions show marked differences from Roman Latin, for which there is, however, little evidence before the end of the 3rd century BC. What is certain is that the language changed so rapidly between the 5th century (the date of a mutilated inscription said to mark the tomb of Romulus and of the Twelve Tables, the contents of which are known from later evidence) and the 3rd century BC that older texts were no longer intelligible.

During this period the Romans subjugated their Latin neighbours (by 335 BC), and their language began to establish itself as a standard form, absorbing features from other dialects. The first author of any note was the comedian Plautus (254–184 BC), whose language is thought to reflect a spoken idiom, some features of which appear to have survived into Romance.

Literary development in Latin

By 265 Rome had conquered Magna Graecia, in the south of the Italian peninsula, and had begun to absorb some of its Greek literary and cultural ideals. Poetic language was especially influenced by Greek until Latin poetry reached its zenith with Virgil. In the 1st century BC a literary prose was to be developed, with emphasis placed on rejection of vulgarity and rusticity and pride of place accorded to elegance and clarity. Grammatical rules were codified and tightened and vocabulary pruned, and the cult of the harmonious, balanced period held sway in rhetorical circles. With Cicero, Golden Age prose style attained its highest point; for the linguist, the distinction Cicero makes between the style of his letters and that of his speeches is especially interesting in that it provides evidence that even educated speech differed from written language. When Cicero uses the *sermo plebeius* ("plebeian speech"), his language is more elliptical, with shorter, less complex sentences and more colourful vocabulary (including plentiful diminutives). It seems obvious that truly popular language differed even more from the elaborate, sophisticated, classical literary idiom; there is evidence that archaic features, banned from literary style, survived in vulgar speech right through to the Romance stage of the language. It is sometimes claimed that the language of Roman historian and politician Sallust (86–35 BC) approximated popular usage, but it is more probable that his archaizing style derives more from conscious imitation of old Roman poetry. The Roman "judge of elegance" Petronius Arbiter (died AD 65/66), too, is often thought to imitate vulgar speech, but many of the odd features found in his "Cena Trimalchionis" ("Trimalchio's Dinner") may represent the broken Latin spoken by Greeks and such.

Some characteristics of Classical Latin. *Pronunciation.* Evidence for pronunciation of the Latin of the classical era is often difficult to interpret. Orthography is conventionalized, and grammarians' commemts lack clarity, so that to a considerable extent it is necessary to extrapolate from later developments in Romance in order to describe it. On the whole, linguists think that Latin probably sounded something like Italian, though areas of uncertainty exist.

Among these uncertainties, the most important concerns Latin intonation and accentuation. The way vowels developed in prehistoric Latin suggests that there was a heavy stress accent on the first syllable of each word, but in later times the accent fell on the penultimate syllable or, when this had "light" quantity, on the antepenultimate (much as in modern Italian). The nature of this accent is hotly disputed: contemporary grammarians seem to suggest it was a musical, tonal accent and not a stress accent. If this were so, the acoustic effect of Latin would be quite different from Romance and similar, perhaps, to modern West African languages or even Chinese. Some scholars claim, however, that Latin grammarians were merely slavishly imitating their Greek counterparts and that the fact that in Latin accent is linked with syllable vowel length makes it unlikely that such an accent was tonal. Probably it was a light stress accent that was normally accompanied by a rise in pitch; in later Latin evidence suggests that the stress became heavier.

The system of syllable quantity, connected with that of vowel length, must have given Classical Latin distinctive acoustic character. Broadly speaking, a "light" syllable ended in a short vowel and a "heavy" syllable in a long vowel (or diphthong) or a consonant. The distinction must have been reflected to some extent in late Latin or early Romance, for, even after the system of vowel length was lost, light, or "open," syllables often developed in a different way from heavy, or "closed," syllables.

Vowel length

Because the system of vowel length was lost after the classical period, it is not known with any certainty how vowels were pronounced at that period; but, because of later developments in Romance, the assumption is that the vowel-length distinctions were also associated with qualitative differences, in that short vowels were more open, or lax, than long vowels. Standard orthography did not distinguish between long and short vowels, although in early times various devices were tried to remedy this. At the end of the Roman Republic an "apex" (one form was like this: ›) often was used to mark the long vowel, but this was replaced in imperial times by an acute accent (′). In Classical Latin the length system was an essential feature of verse, even popular verse, and mistakes in vowel length were regarded as barbarous. In later times, however, many poets were obviously unable to conform to the demands of classical prosody and were criticized for allowing accent to override length distinctions.

Besides the vowels *ā, ē, ī, ō, ū* (long vowels) and *ă, ĕ, ĭ, ŏ, ŭ* (short vowels), educated speech at the classical period used a sound taken from Greek upsilon and pronounced rather like French *u* (the symbol [y] in the International Phonetic Alphabet—IPA) in words borrowed from Greek; in popular speech this was probably pronounced like Latin *ŭ*, though in later times *ī* sometimes substituted for it. A neutral vowel was probably used in some unaccented syllables and was written *u* or *i* (*optumus, optimus* "best"), but the latter rendering became standard. A long *ē*, from earlier *ei*, had probably completely merged with *ī* by the classical period. Classical pronunciation also used some diphthongs pronounced by educated Romans much as they are spelled, especially *ae* (earlier *ai*), pronounced perhaps as an open long *e* in rustic speech, *au* (rustic open long *ō*), and *oe* (earlier *oi*, late Latin *ē*).

The consonant system

The Classical Latin consonant system probably included a series of labial sounds (produced with the lips), *p, b, m, f,* and probably *w*; a dental or alveolar series (produced with the tongue against the front teeth or the alveolar ridge behind the upper front teeth), *t, d, n, s, l,* and possibly *r;* a velar series (produced with the tongue approaching or contacting the velum or soft palate), *k, g,* and perhaps *ng*; and a labiovelar series pronounced with the lips rounded, k^w and g^w. The *k* sound was written *c*, and the k^w and g^w were written *qu* and *gu*, respectively.

Of these, k^w and g^w were probably single labialized velar consonants, not clusters, as they do not make for a heavy syllable; g^w occurs only after *n*, so only guesses can be made about its single consonant status. The sound *ng* (as in English "sing"; represented in the International Phonetic Alphabet by [ŋ]), written *ng* or *gn*, may not have had phonemic status (in spite of the pair *annus/agnus* "year"/"lamb," in which [ŋ] may be regarded as a positional variant of *g*). The Latin letter *f* probably represented by classical times a labiodental sound pronounced with the lower lip touching the upper front teeth like its English equivalent) but earlier it may have been a bilabial (pronounced with the two lips touching or approaching one another). The so-called consonantal *i* and *u* were probably not true consonants but frictionless semivowels; Romance evidence suggests that they later became a palatal fricative, [j] (pronounced with the tongue touching or approaching the hard palate and with incomplete closure) and a bilabial fricative, [β] (pronounced with vibration of the lips and incomplete closure), but there is no suggestion of this at the classical period. Some Romance scholars suggest that Latin *s* had a pronunciation like that of modern Castilian (with the tip, rather than the blade, raised behind the teeth, giving a lisping impression); in early Latin it was often weakened in final position, a feature that also characterizes eastern Romance languages. *R* was probably a tongue trill at the classical period, but there is earlier evidence that in some positions it may have been a fricative or a flap.

The nasal consonants were probably weakly articulated in some positions, especially medially before *s* and in final position; here probably there was mere nasalization of the preceding vowel.

In addition to the consonants shown, educated Roman speakers probably used a series of voiceless aspirated stops, written *ph, th, ch,* originally borrowed from Greek words but also occurring in native words (*pulcher* "beautiful," *lachrima* "tears," *triumphus* "triumph," etc.) from the end of the 2nd century BC.

Another nonvocalic sound, *h,* was pronounced only by educated speakers even in the classical period, amd references to its loss in vulgar speech are frequent.

Doubled consonant sounds

Consonants written double in the classical period were probably so pronounced (a distinction was made, for instance, between *anus* "old woman" and *annus* "year"). When consonantal *i* appeared intervocalically, it was always doubled in speech. Earlier than the 2nd century BC consonant gemination (doubling of sounds) was not shown in orthography but was probably current in speech. Among the Romance languages, the eastern ones on the whole retained Latin double consonants, whereas in the west they were often simplified.

Morphology and syntax. Latin reduced the number of Indo-European noun cases from eight to six by incorporating the sociative-instrumental (indicating means or agency) and, apart from isolated forms, the locative (indicating place or place where) into the ablative case (originally indicating the relations of separation and source). The dual number was lost, and a fifth noun declension was developed from a heterogeneous collection of nouns (principally verbal abstracts in *-ie*). Of the other declensions, the Indo-European *ā*- and *ō* classes remained, with the introduction of new genitive singular forms in *-ae* and *-i,* while consonantal and *-i* stem nouns were amalgamated into a "third" declension, which also took in adjectives formerly of the *ū*-class (a "fourth" declension). Probably before the Romance period the number of cases was further reduced (there were two in Old French—nominative, used for the subject of a verb, and oblique, used for all other functions—and Romanian today has two, nominative–accusative, used for the subject and the direct object of a verb, and genitive–dative, used to indicate possession and the indirect object of a verb), and words of the fourth and fifth declension were absorbed into the other three or lost.

New tense forms

Among verb forms, the Indo-European aorist (indicating simple occurrence of an action without reference to duration or completion) and perfect (indicating an action or state completed at the time of utterance or at a time spoken of) combined, and the conjunctive (expressing ideas contrary to fact) and optative (expressing a wish or hope) merged to form the subjunctive mood. New tense forms that developed were the future in *-bō* and the imperfect in *-bam;* a passive in *-r,* also found in Celtic and Tocharian, was also developed. New compound passive tenses were formed with the perfect participle and *esse* "to be" (*e.g., est oneratus* "he, she, it was burdened")—such compound tenses were to develop further in Romance. In general, the morphology of the classical period was codified and fluctuating forms rigidly fixed. In syntax, too, earlier freedom was restricted; thus, the use of the accusative and infinitive in *oratio obliqua* ("indirect discourse") became obligatory, and fine discrimination in the use of the subjunctive was insisted on. When earlier writers might have used prepositional phrases, classical authors preferred bare nominal-case forms as terser and more exact. Complex sentences with subtle use of distinctive conjunctions were a feature of the classical language, and effective play was made with the possibilities offered by flexible word order.

In the postclassical era, Ciceronian style came to be regarded as laboured and boring, and an epigrammatic, compressed style was preferred by such writers as Seneca and Tacitus. Contemporaneously and a little later, florid, exuberant writing—often called African—came into fashion, exemplified especially by Apuleius (2nd century AD). Imitation of classical and postclassical models continued even into the 6th century, and there seems to have been continuity of literary tradition for some time after the fall of the Western Roman Empire.

The growth of the empire spread Roman culture over much of Europe and North Africa. In all areas, even the outposts, it was not only the rough language of the legions that penetrated but also, it seems, the fine subleties of Virgilian verse and Ciceronian prose. Recent research suggests that in Britain, for instance, Romanization was wider spread and more profound than hitherto suspected and that well-to-do Britons in the colonized region were thoroughly imbued with Roman values. How far these trickled down to the lower classes is difficult to tell; because Latin died out in Britain, it is often thought that it had been used only by the higher classes of the population, but some suggest that it was a result of wholesale slaughter of the Roman British. It is, however, more likely that the pattern of Anglo-Saxon settlements was not in conflict with the Romano-Celtic and that the latter were gradually absorbed into the new society.

Development of Romance from Latin

In the lands in which Romance is still spoken, it is of course certain that, sooner or later, Latin in some form was the normal language of most strata. Whether, however, the Romance languages continue rough peasant dialects of

Latin (or even slave creoles) or the usage of more cultured urban communities is open to question. There are those who maintain that the Latin used in each area differentiated as soon as local populations adopted the conqueror's language for any purpose. According to this belief, dialects of Latin result from "interference" from indigenous languages (substrata), even though clear evidence for dialectal diversification cannot be found in extant texts. It is obvious that Latin usage must have differed over a wide area, but it can be questioned whether the differences were merely phonetic and lexical variations—regional accents and usage—not affecting mutual intelligibility or whether they were profound enough to form the basis of further differentiation when administrative unity was lost. The latter hypothesis would suggest a long period of bilingualism (up to about 500 years), as experience shows that linguistic interference between languages in contact rarely outlives the bilingual stage. Virtually nothing is known about the status of the indigenous languages during the imperial period, and only vague contemporary references can be found to linguistic differences within the empire. It seems odd that no one among the numerous Latin grammarians should have referred to well-known linguistic facts, but the absence of evidence is not sufficient to justify the assertion that there was no real diversification during the imperial era. Historical parallels are lacking—the British Empire did export English to widely different lands, but it lasted a comparatively short time, and its linguistic contribution was backed by modern communications media, besides being to some extent negated by nationalist feeling.

Latin and Romance coexisting

What is certain is that, even if popular usage within the empire showed great diversification, it was overlaid by a standard written language that preserved a good degree of uniformity until well after the administrative collapse of the empire. As far as the speakers were concerned, they apparently thought they were using Latin, though they were often conscious that their language was, through sheer ignorance, not quite as it should be. Not until about the 8th or 9th century—later in some parts—did it strike them that Classical Latin was perceptibly a different language, rather than merely a more polished, cultured version of their own.

Later Latin (3rd century AD onward) is often called Vulgar Latin—a confusing term in that it can designate the popular Latin of all periods and is sometimes also used for so-called Proto-Romance (*roman commun*), a theoretical construct based on consistent similarities among all or most Romance languages. All three Vulgar Latins in fact share common features but, given their different theoretical status, can hardly be called identical or even comparable. Written Vulgar Latin attained wide diffusion as the language of the Christian Church, officially adopted by the empire from the 4th century on. Its "vulgarisms" often called forth apologies from Christian authors, whose false humility seems akin to pride in that they did not succumb to the frivolities of pagan literary style.

Aside from the numerous inscriptions from all over the empire, there is no shortage of texts in Vulgar Latin. One of the first is the so-called *Appendix Probi* (3rd–4th centuries AD; "Appendix of Probus"), which lists correct and incorrect forms of 227 words, probably as an orthographical aid to scribes, but as a result illustrates some phonological changes that may have already occurred in the spoken language (*e.g.,* loss of unstressed penultimate syllables and loss of final *m*). The Vulgate, St. Jerome's translation of the Bible (AD 385–404), and the works of St. Augustine (AD 354–430) are among Christian works in Vulgar Latin. Particularly amusing and linguistically instructive also is the so-called *Peregrinatio Etheriae* ("Pilgrimage of Etheria"), written probably in the 4th century by a nun, describing her visit to the Holy Land. Medical and grammatical works also abound from the 4th to the 7th century (among the writers were the provincials Cosentius, from Gaul; Virgilius Maro, from southern Gaul; and St. Isidore of Seville, from Spain).

New vocabulary

Some of the characteristics of Vulgar Latin recall popular features of classical and preclassical times and foreshadow Romance developments. In vocabulary, especially, many of the sober classical words are rejected in favour of more colourful popular terms, especially derivatives and diminutives: thus, *portare* "to carry" (French *porter,* Italian *portare,* etc.) is preferred to *ferre; cantare* "to sing again and again" (French *chanter,* Spanish and Portuguese *cantar,* etc.) to *canere; vetulus* "little old man" (Romanian *vechi,* Italian *vecchio,* French *vieux,* etc.) to *vetus.* In grammar, classical synthetic constructions are often replaced by analytic; thus, the use of prepositions often makes case endings superfluous. *Ad regem* for *regi* "to the king," for instance, or anomalous morphological forms are simplified and rationalized (*e.g., plus sanus* for *sanior* "healthier"). Shorter, less complex sentences are preferred, and word order tends to become less flexible.

The most copious evidence for Vulgar Latin is in the realm of phonology, though interpretation of the evidence is often open to dispute, consisting as it does of the confused descriptions of grammarians and the misspellings of bewildered scribes. Much of the evidence points to a strengthening of stress accent during the late period, leading to the shortening and swallowing of unaccented syllables: thus, *viridem* "green" becomes *virdem* (*verde* in several Romance languages); *vinea* "vine" becomes *vinia* (French *vigne,* Spanish *viña* ["vineyard"], etc.). It is often thought that Classical Latin had a tonal, not a stress, accent; though this is uncertain, any stress on accented syllables was probably light and less acoustically perceptible than an accompanying rise in pitch. There is some scant evidence that a stress accent was used in popular and dialectal preclassical speech (*e.g., vinia* in a 3rd-century-BC epitaph, Oscan *minstreis* for Latin *minister* "attendant"), but the first undisputed testimony is to be found in 3rd-century-AD texts.

Among other phonological features of Vulgar Latin, probably the most striking is the loss of the system of long and short vowels. On the whole, long vowels became tense and short vowels lax, resulting in a wholesale change in the rhythm of the language. In the texts there is evidence of the confusion of *ĭ* and *ē* and of *ŭ* and *ō* that has occurred in the western Romance languages. A similar collapse of *ō* and *ŭ* seems to have occurred in Oscan, but it is unlikely that this is connected with later developments. It is to be remembered that even popular Latin verse used measures of vowel length, and there is no evidence to suggest that vowel-length distinctions were lost in vulgar preclassical speech.

An archaic feature that does recur in Vulgar Latin is the loss of word-final *-m,* of which virtually no trace remains in Romance. It is possible, however, that the written letter of Classical Latin was no more than an orthographical convention for a nasal twang: in scanning Latin verse, the *-m* is always run in (elided) before a vocalic initial. Reduction of the diphthongs *ae* (to *e*) and *au* (to *o*) seems also to be a popular and dialectal feature reflected in Vulgar Latin texts; in the latter case, however, the Romance languages do not support the hypothesis that the diphthong was reduced early, for it remains in Old Provençal and in Romanian and, probably, in early Old French.

Borrowing from Latin into non-Romance languages

The prestige of Rome was such that Latin borrowings are to be found in virtually all European languages, as well as in the Berber languages of North Africa, which preserve a number of words, mainly agricultural terms, lost elsewhere. Basque has borrowed a good number of words, mainly from administrative, commercial, and military spheres, though it is difficult in some cases to determine whether the terms were later borrowings from Spanish, rather than from Latin. This is not a problem in the case of the 800 Latin words found in three British Celtic languages (Welsh, Cornish, and Breton)—words drawn from a wide sphere of activities. In the Germanic languages, borrowed Latin words principally involve trade and often reflect archaic forms. The very large number of Latin words in Albanian form part of the basic vocabulary of the language (including kinship terms) and cover such spheres as religion, although there is doubt about whether some of them were later borrowings from Romanian. In other cases Latin words in Albanian have survived in no other part of the former Roman Empire. Greek and Slavic languages have comparatively few Latin words, many of them administrative or commercial in character.

Latin has had a continuous influence on the Romance languages and their neighbours in its capacity as a language of religion and culture. With Christianity, Latin penetrated to new lands, and it was perhaps the cultivation of Latin in a "pure" form in Ireland, whence it was exported to England, that paved the way for an 8th-century reform of the language by Charlemagne. Conscious that current Latin usage was falling short of classical standards, Charlemagne invited Alcuin of York, a scholar and grammarian, to his court at Aix-la-Chapelle (Aachen), where he remained from 782 to 796, inspiring and guiding an intellectual renaissance. It was perhaps as a result of the revival of so-called purer Latin that vernacular texts began to appear, for it now became obvious that the vernacular and Latin were not the same language. Thus, in 813, just before Charlemagne's death, the Council of Tours decreed that sermons should be delivered in *rusticam Romanam linguam* ("in the rustic Roman language") to make them intelligible to the congregation.

Latin as the language of religion and education

Latin has remained the official language of the Roman Catholic Church and as such has been in constant use by most Romance speakers; it is only very recently that church services have begun to be conducted in vernacular. As the language of science and scholarship, Latin held sway until the 16th century, when, under the influence of the Reformation, nascent nationalism, and the invention of the printing press, it began to be replaced by modern languages. Nevertheless, in the west, along with Greek, the Latin language has remained a mark of the educated man throughout the centuries, although since World War II the popularity of classical languages in schools has declined, and a generation of scholars who know no Latin, except for the numerous terms borrowed by all European languages, will soon be seen.

The emergence and development of the Romance languages. *Earliest period.* The question of when Latin ended and Romance began, which has occupied scholars in the past, is largely a problem created by terminology. In some senses, today's Romance languages are regional varieties of one uniform set of speech patterns that resembles the Vulgar Latin of attested texts fairly closely—indeed, the analyses of generative phonologists make the modern "underlying forms" (as distinct from their phonetic representation in speech) look almost identical with the reconstructed ancestor of the Romance languages, Proto-Romance. On the other hand, speakers are conscious that today they are speaking a "different language" from their neighbours, even though they may understand a good deal of their neighbours' discourse. Perhaps the speaker's consciousness is the best measure of divergence; when, one may ask, did Romance speakers realize that they were not using Latin in their everyday speech? Some scholars suggest that the realization must have dated from about the 5th century, when barbarians were streaming into the Roman Empire and, supposedly, hindering communication. Others prefer to rely on positive textual evidence, indicative of efforts to make up a written form of Romance distinct from Latin. Such evidence begins to appear only in the 9th century, first in northern France and then in Spain and Italy. The reforms of Charlemagne, reestablishing more classical standards in written Latin, may have been at once cause and result of the development of conventional written forms for vernacular Romance. Perhaps it was also the emergence of a new type of social organization, feudalism, that had linguistic effects as a result of the splitting of the open society of Roman tradition into small closed territorial units.

Romance glosses to Latin texts

From the 7th century onward, consciousness of linguistic change was strong enough to prompt scribes to gloss little-known words in earlier Latin texts with more familiar terms. Though the glosses often reflect Romance forms, however, they are usually given in a Latinate form, and one gains the impression of a few superficial adjustments to archaic but fundamentally comprehensible texts. The best known set of glosses—to the Vulgate Bible of St. Jerome—formerly belonged to the abbey of Reichenau, on an island in Lake Constance, Germany, and probably dates from the 8th century. The vocabulary of the Reichenau glosses appears to be French in flavour (*e.g., arenam* "sand" glossed by *sabulo,* French *sable; vespertiliones* "bats," by *calvas sorices,* French *chauvesouris*), and some words of Frankish origin appear (*e.g., scabrones* "beetles" is glossed by *wapces* "wasps," *respectant* "they look about" by *rewardant*). The glosses provide some evidence of morphological simplification (*e.g., saniore* "healthier" is glossed by *plus sano* "more healthy" and *cecinit* "he sang" by *cantavit*), but for the most part only lexical items are regarded as meriting comment. Another well-known glossary, known as the Kassel (or Cassel) glosses, probably dates from the very early 9th century. It gives Latin equivalents of German (Bavarian) words and phrases and provides evidence of lexical and phonetic differentiation within Latin that permits scholars to localize the work as probably French or Rhaetian (*e.g., mantun* "chin," as compared with modern French *menton*). Although orthographically eccentric, however, the text is obviously meant to represent Latin, not a Romance tongue; when phrases rather than isolated words are glossed, the Latin is often very close to classical models.

Beginnings of Romance literature

Later in the 9th century (with the Strasbourg Oaths, possibly, and more clearly in the Eulalia poem), deliberate attempts were made to write vernacular Romance, though the resources of the Latin alphabet were not wholly adequate to the task. That northern French texts were the first to appear is not surprising, for in that region Latin had changed more radically than elsewhere. By the 10th century the need to couch legal documents in more readily comprehensible vernacular, rather than Latin, was felt in other regions. Vernacular literature did not really get under way, however, until around the 12th century, when the arts flourished throughout western Europe. Rhaeto-Romance and Romanian, however, had to wait for the Reformation period to take on literary form.

Late-medieval period to the Renaissance. There was a good deal of cross-fertilization between Romance literary languages during the period of development of medieval poetry; the example of the Provençal lyric especially left its mark on all vernacular literatures, and borrowing of lexical items from one language to another was abundant. The 13th century saw some shift of linguistic influence from southern to northern France and from Sicily to Tuscany, toward the politically and economically more powerful regions. Portuguese and Catalan developed flourishing literatures somewhat later, taking over some of the traditions of the badly battered southern French region and dominating the literary scene of the Iberian Peninsula. French was fast losing its hold in England, which, a century earlier, had boasted a rich Anglo-Norman literature, and within France the central Parisian dialect began to dominate. In Italy, the Florentine dialect was showing signs of rising to prominence and providing the base for a literary standard.

Influence of Classical Latin on Romance

The rediscovery of classical literature and art, first in Italy and then in other Romance regions, had some considerable effect on the languages in the shape of extensive borrowing from Latin and Greek and, often, conscious attempts to model grammatical constructions in the vernacular on Classical Latin. The Italian standard language, in particular, owes much to the influence of Latin, which it resembles more closely than do the spoken dialects. French, except in the 16th century, was influenced grammatically less by Latin, but from the 14th century onward the habit of preferring words with a quasi-Latin shape to inherited forms became well established, so that much of the French vocabulary has a "learned" appearance. The trickle of Latinisms into Spanish became a flood in the 15th century, and, though Spanish has been more reluctant than French to reject old words, they today form a considerable proportion of the lexicon.

Standardization of the Romance languages in the 17th and 18th centuries. It was in Italy first that the "question of the language" became a matter of hot dispute. Dante himself made an important contribution to the debate on what should constitute a *volgare illustre* (an "illustrious popular speech") capable of rivalling Latin for literary and scholarly purposes. Controversy did not reach its peak, however, until the 16th century. In the Spain of 1492 the completion of the reconquest of Spain from the Arabs and the discovery of America were matched linguistically by the appearance of Antonio de Nebrija's *Gramática Castellana* ("Grammar of the Castilian Language"), which argues the need for an ennobled language fit for imperial exportation. In France during the 16th century, with the Renaissance backed by the Reformation and the advent of printing, French really took over the remaining functions of Latin—scholarly, scientific, and religious—and efforts were made to put together a worthy national language from dialect and Latin sources. The choice of standard was not made definitively, however, until the late 17th century, when, with political power and social influence centred exclusively at the royal court, the only acceptable usage became that of the court. It would seem that social acceptance and advancement were inextricably bound up with correct behaviour, especially linguistic behaviour, so that the well-to-do bourgeoisie set out to ape the speech habits of their "betters"—hence the popularity of works describing *le bon usage* "good usage." The influence of French, resplendent with the achievements of French dramatic poet Racine and of Louis XIV, was destined to remain dominant within the Romance languages; the Golden Age of Spain and Portugal had already passed, and Italy was going through a period of comparative stagnation.

"Correct" language

The French grammarians of the 18th century had lasting effect on all the Romance standards, concerned as they were with maintaining "purity," eliminating "vulgarity," and strictly codifying usage, often more in accord with logical than linguistic considerations. The belief that correct language is not a birthright but a tool to be carefully fashioned and skillfully handled, that conscious effort was required to allow it to mirror thought with the minimum of distortion, is one that has persisted in Romance and that still has important effects on educational practice. To many English speakers it seems ludicrous that the criterion of competence in a language should be strict adherence to grammar-book rules rather than native-like performance, but in Romance countries a foreigner is often frowned upon if he permits himself the "negligence" of native usage, rather than the more stilted correct expression. Educated Romance speakers often speak very formally, with flowing, complex sentences and precise vocabulary, in contrast with the casual, slangy expression of the less educated (who openly envy the speech habits of their "betters"). The passionate interest shown in subtleties of language usage (including regular articles in the better class newspapers) is something that characterizes all Romance speakers, though perhaps only the French take it to excess.

Modern developments. The Romantics of the early 19th century were eager to break the stranglehold of intellectual, aristocratic language, but their attempts to introduce more colourful expressions did not bring them nearer to popular usage, for their efforts were mainly directed toward enriching the vocabulary while leaving grammar intact. Sentimental idealization of peasant existence aroused interest in dialectal usage, and egalitarian sentiments provoked some groups of speakers to proclaim the worthiness of their own mother tongues to rival more politically important languages. Occitan, Catalan, Rhaeto-Romance, and, indeed, Romanian were to develop literatures under the impact of such ideas, which took political form in the demands of regional separatists.

Popular usage in literature

The introduction into literature of conventionalized popular usage is mainly a 20th-century trend, but in the Romance languages it remains more limited than, for example, in English. Other, more literary attempts to break out of the straitjacket of standard usage are connected with such artistic movements as Symbolism and Surrealism, in which syntax, as well as vocabulary, suffers onslaught, and sentence construction tends to cut loose from logical ties. Yet, even within these movements, fine, elegant style continues to be appreciated by many, and it is the traditional forms that have wider appeal. The use of a weighty bureaucratic style in nonliterary writings is also evident, often characterized by an excessive use of Latinism and by the use of verbal nouns, but educational systems continue to place emphasis on plain, classic style. It is difficult to imagine that the long-continuing tradition of interest in "correctness" in language will die out in the Romance countries; recent educational reforms of France and Italy seem rather to emphasize the value of such a tradition, and organizations and individuals, no matter how revolutionary in political outlook, rarely work to counteract the cultural values so long regarded as paramount in their homelands.

CHARACTERISTICS

As a group, the Romance languages share many characteristics besides that which defines the family (*i.e.,* the presence of a significant proportion of lexical cognates). In comparison with Germanic languages, for instance, they seem musical and mellifluous—probably because of the relatively greater importance of vowels than consonants. On the whole, the vowels are clear and bell-like and articulation energetic and precise, though Portuguese and Romanian convey a more muted acoustic impression. Foreigners often think that Romance speech is particularly rapid and voluble, no doubt because individual words receive only light stress (or, in French, no stress), and elision, the running of words into each other within stress groups, is common. Romanian is something of an exception in that speech tempo is comparatively slow. Intonation patterns, surface manifestations of nonlexical meaning, such as interrogation, exclamation, scorn, surprise, and so forth, seem to some to denote excitability and emotional expressiveness in the speakers. Northern French is comparatively sober, with typically about a one-octave range in intonation, but Italian seems to be sung, with sinuous pitch movement over two octaves, and Castilian jumps jerkily and up and down over about an octave and a third.

Retention of the synthetic character of Latin

Grammatically, the modern languages have retained to a greater or lesser extent some of the synthetic character of Latin, principally in the verb, but in Romanian also in the noun. French, since about the 14th century, has undergone most radical changes in grammatical typology, so that much greater reliance is placed on word order and intonation to convey sentence meaning than on morphological form. Other languages allow a little more flexibility of word order but far less than in Classical Latin.

Dominant purist grammarians have always opposed influence from foreign languages and reproved their fellows for sullying their language with lavish borrowing (at present primarily from English), but they have never been able to stem the flood of neologisms. French vocabulary, particularly, has always been receptive to change and has been as quick to lose old words as to adopt new. Codification of grammar, on the other hand, has had a permanent effect on the stability of the standard languages, even feeding back into spoken usage via the education system. Acceptance of the most minor changes follows long debate and deliberation and requires governmental edicts that decree what can be marked as correct in all-important examinations. Curiously enough, this rigidity and consequent self-confidence have resulted in greater teachability, so that standards of correctness of, for instance, French among Africans or Spanish among American Indians are remarkably high. The moves toward codification were, indeed, originally linked to a desire to give the languages inernational importance, and language teaching is, in the Romance ethos, indissolubly linked to the diffusion of cultural and moral values.

Linguistic typology of the Romance languages. As stated previously, the most "central" Romance language is standard Italian, which has retained and even re-adopted many Latin characteristics. In some ways its morphology lacks the elegance and efficiency of Castilian, which has most ruthlessly eliminated anomalies during the modern period; there are signs in Italian of historical inertia, a harking back to a glorious past, that has hindered popular development. Romanian remains closest in grammatical type to Latin, though its noun-declension system, based on the definite article placed after the noun, and its frequent use of the subjunctive mood may owe much to its Balkan neighbours (or to an earlier linguistic substratum). Its vocabulary has incorporated so many Slavic and Turkish words, however, that it often appears less Romance than the rest. French, by any standard, has diverged most—

Causes of the divergences in French

radical phonetic changes that transformed the outward appearance of the language must have preceded the earliest surviving (9th-century) texts. Such changes are usually ascribed to Celtic and Frankish influence. Another wave of change, with loss of word accent and of many morphological markers, probably dates from around the 15th century, but it is difficult to find external motivation for these phenomena. Occitan and Catalan are conservative in character; the long persistence of Roman schools in South Gaul is often seen as the cause of stability there. Spanish and Portuguese are close enough to lead some scholars to assign their shared characteristics to Iberian substratum and Moorish superstratum influence. Castilian's forceful character and receptivity to grammatical innovation contrast sharply with Portuguese softness and its inertia in retaining morphological oddities, however. One might conceivably see the differences as connected with climatic and geographical conditions, though just how would be difficult to discern. Rhaeto-Romance and Dalmatian peculiarities can most easily be connected with the impact of other languages (mainly German, Italian, and Serbo-Croatian), while Sardinian is often regarded as an extremely conservative, peasant language, some dialects of which have been penetrated by features from Italian and Spanish.

Phonology. Some important phonological developments, such as the loss of the system of contrasting vowel lengths and the strengthening of the stress accent, must have occurred during the Vulgar Latin period, while some degree of unity still existed among the various Romance dialects. Certain other changes shared by the Western Romance languages, especially the collapse of *ē* and *ĭ*, might have postdated the linguistic separation of Sardinia and parts of southern Italy from the other areas, while the distinct development of *ō* and *ŭ* in Romanian and Vegliot suggests a split between Eastern and Western Romance at a later date.

Unaccented vowels

Vowels. Everywhere, unaccented vowels have had a different history from accented, and in some languages they have so weakened as to disappear altogether in certain positions. At the end of a word, for instance, even *-a,* the most sonorous of the vowels, has weakened to a neutral vowel in Romanian, Portuguese, and some Catalan and Rhaetian dialects—in some French dialects it is still pronounced as a neutral vowel sound (such as the second vowel in English "alph*a*bet"), but it has been lost completely in the standard language. Final *-o,* from Latin *-o* or *ŭ,* was lost very early in French, Occitan, Catalan, and Rhaetian and remains only before an article following the word in Romanian; in Portuguese it is closed to a *u* sound (such as the *u* in English "l*u*nar"). Final *-e* is even more evanescent, regularly remaining as a full vowel only in parts of central and southern Italy and Sardinia.

Under the main stress accent of the word, Latin vowels have often become diphthongs in Romance, perhaps as a result of lengthening under heavy stress or as a consequence of the raising influence of following high vowels (a process known as breaking, similar in action to German umlaut). The vowels most affected are the "open" *e* sound (as in "m*e*t"), from Latin *ĕ,* and to a lesser extent the "open" *o* sound (similar to the *aw* sound in "law" in many American English dialects and to the *o* in British English "ingot"), from Latin *ŏ,* while high close vowels *i* and *u* are virtually untouched. Transformation of short *e* to a diphthong (usually a *ye* sound, as in "yet") is so common that some believe it occurred during the Vulgar Latin period. The conditions of this process (and similar ones) vary, however; in some languages (notably French and Italian) it happens only in open syllables (*i.e.,* those ending in a vowel in Vulgar Latin), whereas Romanian, Vegliot, Spanish, and perhaps Rhaetian show similar developments in all accented syllables. Portuguese possibly did not join in the diphthong-forming process at all, though, as in Occitan, Catalan, Sardinian, and some Italian dialects the short *e-* and *o-* sounds may at one time have developed into diphthongs under the influence of a following high vowel (*i* or *u*), later to be reduced once more to a single vowel. Table 10 illustrates treatment of stressed Latin *ĕ* and *ŏ* in different languages.

Table 10: Occurrence of Diphthongs Replacing Stressed Short Vowels in Romance Languages

	pĕde "foot"	*hĕrba* "grass, herb"	*mŏrit* "he dies"	*mŏrtem* "death"
Sardinian	*pe*	*erva*	*móridi*	*morte*
Portuguese	*pe*	*herva*	*morre*	*morte*
Catalan	*peu*	*herba*	*mor*	*mort*
Occitan	*pe*	*erba*	*mor*	*mort*
French	*pied*	*herba*	*meurt* (Old French *muert*)	*mort*
Italian	*piede*	*erba*	*muore*	*morte*
Romanian	—	*iarbă*	*moare*	*moarte*
Spanish	*pié*	*hierba*	*muere*	*muerte*
Rhaetian (Sursilvan)	*pei*	*jarva*	*miere*	*mort*
Rhaetian (Friulian)	*pid*	—	—	*muart*
Vegliot	*pi*	*járba*	—	*muart*

Reflexes of Latin *ō* (and *ŭ*) and *ē* (ĭ) became diphthongs *ou* and *ei* in Northern French at an early period (after the 5th but before the 9th century); the 12th-century phonetic results *eu* and *oi* provided the present-day spellings, though the sounds thus represented have changed considerably since (compare *fleur* "flower," from *flour,* from *flore*). The greater extension of spontaneous diphthong formation in French than in other Romance languages (including perhaps also reflexes of *a*—compare *mer* "sea," from **maer* [?], from Latin *mare*) is often attributed to the effects of the heavy stress presumably used by the Frankish superstratum.

Effects of nasal consonants on vowels

In nearly all Romance languages a following nasal consonant has caused peculiar development in a preceding vowel. In most cases the effect is limited to a raising or closing influence, but in two major languages, French and Portuguese, phonological nasalization has taken place (*i.e.,* a series of vowels distinguished by the presence of nasal resonance has developed). Here, as well as in some other dialects (especially Chilean, Caribbean, and Andalusian Spanish, in the Romanian spoken in Albania, and in northern Occitan), nasal vowels are distinct from their oral counterparts and not mere variants (*i.e.,* they are phonemic). Thus, they serve to differentiate one meaningful form from another: *e.g.,* French *pin* "pine," pronounced *pɛ̃* (ɛ stands for a short *e* sound, and ~ marks nasalization) versus *paix* "peace," pronounced *pɛ*; Portuguese *lã* "wool" versus *la* "there"; Andalusian *cantã* "they sing" versus *canta* "he sings." Occasionally, nasalization of a vowel is caused by a preceding consonant (*e.g.,* Portuguese *mãe* "mother," from *matre*), but this is comparatively rare.

Nasalization in both French and Portuguese was probably noticeable by the 10th century, though it may not have become phonemic until much later. Some claim that even today nasal vowel resonance is merely a surface manifestation of a latent underlying nasal consonant. It would appear that in both languages nasal vowels were more frequent in the Middle Ages than today; in about the 16th century in France, denasalization took place when the nasal consonant was intervocalic, and the *n* sound was retained—in, for example, French *bon* "good [masculine]" (pronounced *bõ*) and *bonne* "good [feminine]" (pronounced *bon* or [bon]). In Portuguese the consonant did not always reappear after denasalization (compare *boa* "good [feminine]," from *bõa,* from *bona*), though between *i* and *a* or *o* the palatal nasal consonant (close to *ny* in "canyon") is inserted (*vinho* "wine," from *vĩo,* from *vinu*).

Nasalization has sometimes, though without much conviction, been attributed to Celtic substratum influence. A better case can be made for the effect of such influence in the French *u* sound, [y], pronounced like German *ü* or Greek upsilon, though ignorance of Gaulish and certain chronological and geographical discrepancies make it difficult to argue in detail. The French *u* sound is also found in most Occitan dialects (in which it may be a recent introduction from French), in Rhaetian, and in parts of Portugal and Italy; elsewhere it is sometimes a characteristic of affected speech.

French uvular *r*

Consonants. Another French pronunciation that is often imitated by socially pretentious speakers is that of the Parisian uvular *r* (produced by vibration of the uvula, an appendage at the back of the mouth), which was not accepted in standard French until after the Revolution, though probably used by the Parisian bourgeoisie from the 17th century. It probably developed from the Latin double *-rr-*, differentiated from single *-r-*, which in Middle French tended to be pronounced with local friction, almost as a *th* or *z* sound (compare *chaise* "chair" and *chaire* "chair—throne, pulpit"). In most dialects of Provence today the distinction between the two *r* sounds is still made (though Occitanian dialects in general are adopting the French pronunciation). Brazilian Portuguese uses a similar contrasting pair of *r* sounds, with the usual trilled *r* represented in orthography by "*r*" and a velar, or "rough," *r* represented by "rr": Brazilian *caro* "dear" and *carro* "cart." Elsewhere only Puerto Rican Spanish and a few North Italian and Romanian dialects use the velar *r* regularly, though it is heard sporadically nearly everywhere.

One phonological development that is thought by many to be indicative of a very early split between the Eastern and Western Romance areas concerns the treatment of consonants between vowels. To the north and west of a line drawn between La Spezia and Rimini, in Italy, most dialects voiced Latin voiceless consonants between vowels and simplified geminates (doubled consonants); southern and eastern dialects to a greater extent retain the Latin voiced–voiceless–geminate system. The dividing line appears also to run through Sardinia, so that northern dialects are "Western" and southern ones "Eastern." Table 11 shows the treatment of intervocalic *p* and *t*.

Table 11: Development of Latin Intervocalic p and t in Romance Languages

	ripa "bank"	*rota* "wheel"
Vegliot	*raipa*	—
Romanian	*rîpă*	*roata*
Italian	*ripa*	*ruota*
Logudorian	*riba*	*roda*
Occitan	*riba*	*roda*
Catalan	*riba*	*roda*
Spanish	*riba*	*rueda*
Portuguese	*riba*	*roda*
French	*rive*	*roue* (Modern French) *ruede* (Old French)
Rhaetian	*riva*	*roda, ruede*

Some believe that the voicing of voiceless sounds is connected with a similar, though not identical, process known as lenition in Celtic. Lengthening and subsequent development into diphthongs of accented vowels may be linked to the reduction of Latin doubled consonants to single consonants, as some recent theories suggest.

One noticeable difference between Latin and all the Romance languages is that the consonantal systems of the latter include a number of palatal and palato-alveolar consonants, which did not exist in Latin. (Palatal consonants are formed with the tongue touching the hard palate; palato-alveolar sounds are made with the tongue touching the region of the alveolar ridge or the palate.) One consequence of the strengthening of the stress accent in the later Latin period was that unstressed *ĭ* and *ĕ* following consonants became shortened to a nonsyllabic palatal *y* sound (called *jod*). The effects of this new sound on preceding consonants are varied, but in many cases these have been pronounced with the tongue raised more toward or against the roof of the mouth, or palate (a process classified under the general heading assimilation), sometimes ending up eventually as a dental fricative (such as *z* or *th*) or affricate (such as *ch*) and perhaps modifying the preceding vowel. That this process began early is suggested by the not-infrequent confusion of *-tĭ-* and *-cĭ-* in orthography, sometimes represented even as *tz* in inscriptions. This palatal shift in pronunciation led to developments such as French *rouge,* Portuguese *ruivo,* Catalan *roig,* and Italian *rosso* from Latin *rubeum* "red" and French *feuille,* Portuguese *folha,* Italian *foglia,* and Sardinian *fodza* from Latin *folia* "leaf."

Shifts involving palatal consonants

Another source of palatal consonants in Romance has been back (velar) consonants when immediately followed by a front sound: the velar consonant has often moved forward in the mouth, sometimes eventually to dental or alveolar position but often settling on a palatal or palato-alveolar position. This process, too, probably began early, first affecting velar consonants *k* and *g* preceding front vowels *e* and *i.* That it had not occurred at the classical period is shown by its absence in early loanwords into other languages (Berber, Basque, Celtic, Germanic, Albanian, and Greek). As central Sardinian dialects retain velar pronunciation in the environment of front vowels, it may be assumed that palatalization postdated the separation of the island from the rest of the empire. Vegliot evidence is difficult to interpret, as *ē* does not seem to have provoked palatalization, whereas *ĕ, ĭ,* and *u* did so. It was this sound change that resulted in the pronunciation of "soft" *c* before *e* and *i* (in most Romance languages this is an *s* or *ts* sound; in Italian and Rhaetian it is a *ch* sound). Before *a, o,* and *u* the *c* retained its "hard" pronunciation (that is, a *k* sound). In Classical Latin, before the sound change occurred, all *c* sounds were "hard." Hence, Latin *centum* ("kentum") gave rise to Italian *cento* ("chento"), Portuguese *cento* ("sento"), and Spanish *ciento* ("siento" or, in Castilian, "thiento").

In north central France, Latin *a* must have advanced to a front position, with the result that it, too, palatalized preceding *k* and *g* sounds. The results give the palato-alveolar sounds of *sh* and *zh* (written in the International Phonetic Alphabet as [ʃ] and [ʒ], respectively), via [tʃ], the *ch* sound in "church," and [dʒ], the *j* sound in "jam"; *e.g.,* French *chanter* "to sing" developed from Latin *cantare, joie* "joy" from *gaudia.* West Rhaetian dialects show a similar development (compare Sursilvan *tgaun,* Engadine *chaun,* French *chien,* from Latin *canem* "dog"), as do Franco-Provençal and Northern Occitan dialects, but Picard and some Norman dialects do not (Picard *canter,* with an unpalatalized *c,* from Latin *cantare; kier* "dear," from *carum*). The change is assumed to have taken place at a later period than the palatalization of *k* when followed by *e* or *i,* which did not affect Frankish words. These, on the other hand, succumbed to the type of palatalization in which *k* changed to *ch* [tʃ] and then to *sh* [ʃ] (**skina* > *échine* "backbone").

Romanian consonant developments

In Romanian, velar consonants were moved forward under the influence of a following *i* and *e,* and dental consonants were moved back to a palatal position under the same influence; *e.g., ţară* from *terram* "earth"; *şi* "and" from *sic* "thus." Labial consonants are also affected in some dialects: *k'ept* from *piept* from *pectum* "chest"; *jin* from *vin* from *vinum* "wine." Romanian also has, in final position, a series of "soft" consonants, reminiscent of the Slavic sounds. These are transparently derived from earlier "hard" consonants followed by *i,* performing certain important morphological functions: *lupi* [lup'] "wolves" / *lup* [lup] "wolf"; *cînţi* [kints'] "thou singest" / *cînt* [kint] "I sing."

Palatalization of consonants in Romance was effected not only by following front vowels but also by juxtaposed front consonants, especially when a velar (such as Latin *c* or *g*) was next to a dental (such as *t, s, n*) or a lateral (*l* sound) in medial position, sometimes as a consequence of the loss of an unaccented vowel during the Vulgar Latin

Table 12: Results of Palatalization of Consonant Clusters

	noctem "night"	*coxam* "hip"	*piscem* "fish"	*pugnum* "fist"	*oc'lum* "eye"
Vegliot	*nwat*	—	*pask*	—	*vaklu*
Romanian	*noapte*	*coapsă*	*peşte*	*pumn*	*ochi*
Sardinian	*notte*	*koša*	*piske*	*pundzu*	*okru*
Italian	*notte*	*coscia*	*pesce*	*pugno*	*occhio*
Occitan	*nôit, nuech*	*cuoissa*	*peis*	*ponh*	*uelh*
Catalan	*nit*	*cuixa*	*peix*	*puny*	*ull*
Spanish	*noche*	*cojo*	*pez*	*puño*	*ojo*
Portuguese	*noite*	*coxa*	*peix*	*punho*	*olho*
Rhaetian					
Sursilvan	*notg*	*queissa*	*pesch*	*pugn*	*egl*
Engadine	*not*	—	—	*puoñ*	—
Friulian	*ñot*	—	*pes*	—	—
French	*nuit*	*cuisse*	(*poisson*)	*poing*	*oeil*

period. Results of this process vary from language to language. Table 12 gives examples of these changes.

It will be noted that in Romanian a labial consonant has been substituted for the velar in the Latin clusters *-ct-*, *-x-* [ks], and *-gn-*. Perhaps there was first assimilation of the velar to the dental—as in Italian *-tt-* from Latin *-ct-* and Sardinian *-nn-* from Latin *-gn-* (*linna* from *ligna* "line")—followed by differentiation of the first element of the geminate. It is notable that Latin *l* regularly becomes *jod* after another consonant in Italian (*piacere* from *placere* "to please"; *fiore* from *flore* "flower"; *chiave* from *clave* "key"; *ghianda* from *glanda* "acorn") and after velars in Romanian (*plăcea, floare,* but *cheie* [kjej], *ghindă* [gjində]). In Spanish and Portuguese a following *l* in Latin often palatalizes labial consonants (*p, f*) as well as velars, in initial as well as medial position; *e.g.,* Latin *planum* becomes Spanish *llano* "plain," Portuguese *chão;* Latin *afflare* becomes Spanish *hallar* "to find," Portuguese *achar.*

Grammar. Item for item, the Romance languages all appear grammatically close to Latin and to each other: superficial resemblances in individual expressions may, however, mask differences of content and construction that are difficult to describe. The most obvious difference between Latin and Romance is in the comparative autonomy of morphemic units, especially words. In Romance, Latin inflectional endings have been much reduced, and more reliance is placed on syntactic construction to convey sentence meaning; that is, Romance languages are more "analytic" than the predominantly "synthetic" Latin. A corollary of this is that word order is less flexible in Romance, as it has become the principal means of showing relationship between words in the sentence.

Reduction of inflectional endings

Forms of nouns and adjectives. The inflectional endings have been lost most in nouns and adjectives. The Classical Latin five-case declensional system has everywhere been replaced (with a couple of doubtful exceptions) by a two-gender system, in which normally masculine gender is marked by survivors of the second (*-us*) declension endings of Latin (Italian *cavallo,* Portuguese *cavalu,* Romanian *calul,* Sardinian *kaḍḍu,* Rhaetian *cavagl,* from Latin *caballus* "horse"), and feminine is marked by first (*-a*) declension endings (Italian *capra,* Spanish *cabra,* Rhaetian *caura,* Romanian *capră,* from Latin *capra* "goat"). Cognates of third-declension Latin noun forms are incorporated into the same system, but their gender is marked by changes in the article or accompanying adjective (agreement or accord) rather than by overt markers in the word itself (for example, masculine Italian *il monte,* Catalan *es munt,* from Latin *mons, montem* "mountain"; feminine Italian *la notte,* Catalan *sa nit,* from Latin *nox, noctem* "night"). In modern French, although gender is marked in the written language, however inconsistently, by the presence or absence of final *-e,* any overt morphological markers the spoken language may have are more complex in character, and more reliance is placed on syntactic agreement; thus, *chatte* "she-cat" is distinguished from *chat* "cat" by the presence or absence of the final consonant sound *-t* in pronunciation, but *(le) tour* "tour, trick" and *(la) tour* "tower" have identical phonetic shapes though they belong to different gender classes.

All the Romance languages continue to mark plurality in nouns and adjectives morphologically, though in modern spoken French this is not done consistently. In Western Romance the sign of the plural is usually *-s,* derived from the Latin accusative plural flection: Spanish *caballos, cabras, montes;* Occitan *cavals, cabras, mons;* Catalan *cavalls, cabres, muntes;* Sardinian *kaḍḍos, krabas, montes;* Old French *chevals, chèvres, monts.* In Italian and Romanian, however, plurality is shown by a final *-i* (which in Romanian "softens" the preceding consonant) or, in the case of some feminine nouns, by a final *-e:* Romanian *cai, capre, munți, nopți;* Italian *cavalli, capre, monti, notti.* These endings may derive from Latin nominative plural first- and second-declension endings *-ae* and *-ī,* or they may represent a somewhat irregular development of the *-s,* favoured elsewhere.

Loss of case system

The Latin nominal case system has disappeared in all modern languages except Romanian, in which the inflected article distinguishes the nominative and accusative from the genitive and dative (see Table 13). Thus, when other Romance languages would use a preposition to indicate a certain relationship between words, Romanian resembles Latin in using an inflected form (*e.g.,* Latin *matris* "the mother's: Romanian *mamei,* French *de la mère,* Italian *della madre*).

Table 13: Declensional System of Romanian

	singular	plural
Masculine "son"		
Nominative–accusative	*un fiu, fiul*	*fii, fiii*
Genitive–dative	*unui fiu, fiului*	*unor fii, fiilor*
Feminine "mother"		
Nominative–accusative	*o mamă, mama*	*mame, mamele*
Genitive–dative	*unei mame, mamei*	*unor mame, mamelor*

In Old French and Old Provençal some remnants of a case system remained, in that the masculine nominative (subject of the verb) was distinguished from the other cases (collectively called oblique). Today such grammatical information is conveyed by word order in most Romance languages, as in English, with the subject normally preceding the verb: French *Pierre appelle Paul* "Peter calls Paul"; Portuguese *Pedro chama Paulo;* Italian *Piero chiama Paulo.* Some Romance languages pick out the object of the verb, if it is a person, by an additional particle: Spanish *Pedro llama a Pablo;* Romanian *Petru cheamă pe Pavel.* Several Italian dialects, as well as Sardinian and occasionally Engadine and Portuguese dialects, have similar constructions: Calabrian *Chiamu a Petru* "I call Peter"; Elba *Ô visto a ttuo babbo* "I saw your grandpa"; Engadine *Amè a vos inimihs* "Love your enemies." It is notable that the Italian-based lingua franca used by Mediterranean sailors since the 16th century also picks out the personal object (*e.g., Mi mirato per ti* "I saw you").

The definite and indefinite articles were unknown in Latin but developed everywhere in Romance, usually from the Latin demonstrative *ille* "that" (though in a few parts from reflexive *ipse* "himself") and the numeral *unus* "one." The articles seem to have played some part, during the older stages of the languages, in distinguishing subject from object; the article is more often used where a Latin nominative would have occurred than in other cases, perhaps to give prominence to the topic of the sentence. Today the use of the article has so extended that such distinction is no longer possible; in French, for instance, a common noun is always accompanied by a determiner such as an article, demonstrative, or possessive, so forms remaining from the earlier stage, such as *avoir faim* "to be hungry," are often regarded as idiomatic and inexplicable in terms of modern structure.

The system of verbs. In the passage from Latin to Romance, verbal inflection has survived much more than noun declension. Although the four regular Latin conjugations have been virtually reduced to two, with only the *-a-* class remaining truly productive, other features of the verb seem almost unchanged. In most languages, for instance, the person markers are directly traceable to Latin origins (*i.e.,* to Latin *-ō, -s, -t, -mus, -tis, -nt*). Modern spoken French is the only major language in which the personal endings no longer serve the same function as Latin. Today, person is marked in French principally by pronouns derived mainly from the Latin emphatic nominative forms of the personal pronoun: *J'aime* [ʒɛm] "I love," *tu aimes* [tyɛm] "you love" from (*ego*) *amo,* (*tu*) *amas.* The creoles have taken this process even further, in that their verb forms are usually invariable but are prefixed by elements indicating person, tense, aspect, etc., as in many West African languages: Louisiana French [motegẽ] "I was having" from *mon* [mo] *étais* [te] *gagner* [gẽ]; and similarly [ilagẽ] "he will have."

Verb conjugations

In the metropolitan languages, verbal modalities are shown, as in Latin, by inflection. Some Latin verb endings, such as that of the *-r* passive or of the future, have disappeared; others, such as the pluperfect indicative and subjunctive, have survived in a few languages with modified function. But most languages today have reflexes of the present, perfect, and imperfect indicatives and of one or more subjunctive tenses. The imperfect indicative, a Latin innovation, survives almost intact, though the evolution of its form, not to mention its function, presents problems. The *-ī-* stem form in Latin *-iēba-* is thought to have coalesced early with the *-ē-* stem *-ēba-* form, but a few languages (notably Italian, Friulian, and some Spanish and Portuguese dialects) today have reflexes of an *-ība-* form that might have survived from popular Latin. The Latin *-āba-* form survives almost everywhere, though in most French dialects its older reflexes, *-eve* and *-oue,* have been replaced in modern times by forms derived from Latin *-ēba-*. These latter are thought to be widespread but are puzzling phonologically as they have very often irregularly lost their *-b-* (Spanish, Portuguese, etc., *-ía,* French *-ais*).

The Latin perfect of the type *amāvit* "he has loved" is known by all the literary languages but is rare in speech in French, Italian, and Romanian, in which it has been replaced by a new compound past made up of the verb for "to have" and a past participle. The latter structure is known to some extent in all Romance languages, often being used to express a more recent past than the preterite *amāvit* form, which also indicates action in the past (without reference to duration or repetition): Romanian *au cîntat,* Italian *ho cantato,* French *j'ai chanté,* Spanish *he cantado,* Old Portuguese *hei cantado,* Engadine *ha chantà, hè chantò,* Sardinian *kantau appo,* from Latin *habeo cantatum* "I have sung." In Modern Portuguese the preferred auxiliary is *ter* "to have, to hold" rather than *haver,* producing forms such as *tenho cantado,* while modern Catalan more commonly uses the verb for "to go" plus the infinitive, giving *vaig cantar* rather than the pan-Romance type *hé cantat*.

Representation of the future tense

The disappearance of the Latin future has been remedied in most Romance languages by the development of new forms of periphrastic origin. Many of these forms use some reflex of *habēre* "to have" joined to an infinitive. From Latin *cantāre habēo* "I will sing" are derived Italian *canterò,* Spanish, Catalan *cantaré,* Portuguese *cantarei,* French *je chanterai,* Rhaetian *c(h)antero, c(h)antera,* Occitan *cantarai; habēo cantāre* gives southern Italian *aggio cantà* (similar forms are seen in earlier Spanish, Portuguese, and northern Italian). Latin *habēo ad cantāre* produces Sardinian *ap a kantare,* and *habēo de cantāre* gives Portuguese *hei-de cantar* (more popular than *cantarei*).

A periphrastic future of the type known in English "I'm going to sing" enjoys popularity in Romance, mainly to indicate a less distant future event than the more formal future tense (*e.g.,* French *je vais chanter,* Spanish *voy a cantar*). Other periphrases used in Romance are "I will (wish to) sing," as in Romanian *voi cînta;* "I must sing," as in Sardinian *deppo kantare;* "I'm coming to sing," Sursilvan *jeu vegnel a cantar;* and "I have that I should sing," as in popular Romanian *am să cînt.* Notably, Dalmatian does not seem to know periphrastic Romance futures but uses a form *kantuora* (perhaps from Latin *canāverō*) as both future and conditional.

The Romance conditional, or "future in the past," a form not found in Latin, is in many languages related to the new future. In the Western languages it is composed of the future stem (or infinitive) plus a past-tense marker related to reflexes of *habēre.* In some cases an imperfect form is used, in others a perfect form; examples are French *je chanterais* "I would sing," Spanish, Portuguese, Occitan, and Catalan *cantaría,* and Italian *canterei, -ebbe,* etc. In Romanian the conditional marker can either precede or follow the infinitive and may be derived from the imperfect of *vrea* "to wish": for example, *aşi cînta, ar cînta,* etc., or (more literary) *cîntare-aş, cîntare-ar,* etc.

Word order

Word order is the means most used by modern Romance languages to show the grammatical relationship between words; statistically the most frequent order in statements is subject–verb–noun object. In many of the Romance languages, interrogation can be shown by inversion of the subject and verb, placing the verb, as the element on which the interrogation falls, at the beginning of the sentence (Spanish *¿Vino el hombre?,* Italian *É venuto l'uomo?* "Has the man come?"). In such examples, however, it is the intonation (represented in writing by the question mark) rather than the word order alone that marks the question. Inversion, without interrogative intonation, is not infrequent in emphatic assertions. Unambiguous question markers—such as the Latin particles *-ne, nonne,* and *num*—are lacking in most Romance standards; popular speech, though relying everywhere principally upon intonation, often has developed new particles to reinforce interrogation. Romanian has *oare, şi* (*Oare a venit?* "Has he come?"; *Si te ai culcat?* "Have you been to bed?"); Italian uses dialectal *ce, che,* or *o* (Vulgar Tuscan *Che è venuto?* "Has he come?"; *O come si chiame?* "What is he called?"); Sardinian has *a (A mosse kkŭstu kăne?* "Does this dog bite?"); and French and Limousine have *ti* (generalized from such forms as *a-t-il?;* French *Je suis-ti bête?* Limousine *Sieu-ti nesci?* "Am I stupid?"). In modern standard French great use is made of *est-ce que* as an interrogative particle: *Est-ce qu'il est venu?* "Has he come?" *Comment est-ce qu'il s'appelle?* "What is his name?"

Negation in Latin was expressed by a range of special items (*non, nemo, nihil, nullus, nunquam,* etc.). Although some of the others survive in Romance, continuators of *non* have taken over the main burden of negative expression and are regularly prefixed to the verb. Nuances within negation are usually expressed by the adjunction of other items. In France, both north and south, and in northern Italy and some of the Swiss Rhaetian areas, the *non* particle has been so weakened phonetically that it no longer can express unambiguously the important distinction between negative and positive; hence, formerly positive adjuncts have acquired its negative meaning.

French *personne / une personne* signifies "no one / a person"; *pas / un pas* means "not / a step"; and *plus* can mean "more / no more." In popular speech the *non* particle is frequently omitted altogether in areas that use these additional forms (*e.g.,* French *Je (ne) le vois pas;* Occitan *Lou vese pas* for *Noun lou vese* "I don't see it").

Reduction of the subjunctive

Morphologically, the verb system survived comparatively intact from Latin to Romance; if the schoolbooks, heavily influenced by Latin grammar, are right, the ways in which the verb forms are used are not so very different from Latin either. The most obvious change has been the reduction of uses as well as of forms of the subjunctive, with, at the extreme, modern French treating them as automatically determined variants to be used obligatorily after certain phrases and conjunctions and virtually eliminating tense differences within the subjunctive mood. When the subjunctive retains a function in Romance—that is, in contexts in which it can contrast with the indicative—it has developed emotive overtones, especially suggesting doubt, unreality, or some sort of hypothetical futurity. It is used especially in subordinate clauses dependent on verbal expressions of command and exhortation, emotion, or doubt: Romanian *voi să vină* "I want him to come"; Engadine *Mieu bap voul ch'eau lavura* "My father wants me to work"; French *Je doute qu'il vienne* "I doubt that he's coming"; Portuguese *Duvido que seja feliz* "I doubt that he is happy"; Italian *Temo che sia tarde* "I'm afraid it's late"; Spanish *Temo que él lo diga* "I'm afraid he'll say it." The subjunctive also regularly follows subordinating conjunctions that project action forward into the future, such as "until," "before," "in order that": French *avant que vous soyez venu* "before you came"; Spanish *hasta que sea feliz* "until he is happy"; Italian *perchè potessi fare in tempo* "so that I might do it in time"; Portuguese *antes que eu o veja* "before I see it"; Catalan *abans que vingui* "before he comes."

On the whole, however, the Romance languages use the subjunctive less than Latin, with recession particularly, when no doubt is implied, in indirect speech and in temporal and concessive clauses (in French, use of the subjunctive after concessive conjunctions such as *bien que* and *quoique* "although" was imposed by 18th-century grammarians). The infinitive is often used in subordinate constructions when Latin would have used a subjunctive; *e.g.,* French *dites-lui de s'en aller,* for *dites-lui qu'il s'en aille* "tell him to go away." Romanian, on the other hand, has even extended the use of the subjunctive in such constructions, perhaps reflecting a substratum influence that is felt, too, in other Balkan languages. Greek influence is sometimes credited with similar constructions (usually using the indicative rather than the subjunctive) found in northeast Sicily, northern Calabria, and the Salentine Peninsula.

Conditional clauses

One area of syntax in which the Romance languages vary widely in the extent to which they retain and in the manner in which they replace the Latin subjunctive is that of past-tense hypothetical conditional clauses. The Latin formula *si habuissem dedissem* "if I had had it, I would have given it," though challenged by a type using the indicative tense since Ciceronian times, has sporadically survived into Romance, especially in the older stages of the languages and in scattered parts of southern Italy (*Se potessi, facessi* "If I could, I would do it"), Rhaetian (Sursilvan *Jeu vegness, sche jeu vess peda* "I'd come, if I had time") and Romanian (*'dacă aşi avea destui bani, aşi cumpăra-o* "If I had enough money, I'd buy it").

In most languages, however, a new conditional form replaces the subjunctive in "if" clauses. Thus, in Spanish, Portuguese, and most Italian dialects, sentences of this type are seen: Spanish *si yo tuviese bastante dinero, lo compraría;* Italian *se avesse abbastanza danaro, lo comprerei;* Portuguese *se tivesse bastante dinheiro compraríao* ("if I had enough money, I'd buy it"). Spoken Catalan usually prefers a similar construction (*si estudiessis ho sabries* "if you studied, you would know it"). Another construction that replaces the subjunctive by the imperfect indicative in the "if" clause is, however, considered more correct in Catalan and is normal in French as well as in Corsica and Sardinia: Catalan *si estudiaves ho sabries* ("if you studied, you would know"); French *si j'avais assez d'argent, je l'achèterais* ("if I had enough money, I'd buy it"); Logudorian *si denía abba deo dia buffare* ("if I had water, I'd drink"). Other constructions using the imperfect indicative or the conditional in both clauses are found mainly in substandard styles—both types are common in French, the former in Tuscany, southeastern Italy, and Spain and the latter in much of southern Italy.

Forming new words

Word formation. Romance methods of forming new words from native sources are in part inherited from Latin (the morphological device of adding a suffix and that of prefixing an element that modifies the original meaning) and in part later developments (mainly that of combining two or more free forms to make compound words and of changing or extending the syntactic distribution of an already existing word).

Derivation by means of suffixes is the most popular and widespread device; verbs in particular must be morphologically marked as members of a conjugation, of which those corresponding to Latin *-āre* form by far the most frequent and indeed in modern times virtually the only productive class (thus Latin *plantāre* "to plant," Italian *plantare,* Engadine *plaunter,* French *planter,* Catalan *plantar,* from *planta* "plant"). Infixes, inserted between the verbal root and the conjugation marker, are common. Sometimes they continue Latin infixes, such as the frequentative (compare *jactāre* for *jacere* "to throw," Italian *gettare,* French *jeter,* Catalan *getar,* etc.); sometimes they add semantically to the root meaning (compare pejorative Italian *lavoracchiare* "to slack off" from *lavorare* "to work," French *criailler* "to bawl" from *crier* "to cry"). The Greek verbal infix *-iz* (as in English "ize") is particularly popular in Romance today (*e.g., latinisare, automatiser*).

Among noun suffixes, diminutives are frequent and, except perhaps in French, still productive. Romanian uses *-aş* (*degetaş* "little finger", but the other languages prefer derivatives of Latin *-ittus* (especially in Spanish: *arbolito* "little tree," *señorita* "Miss, young lady," etc.; but also French *sachet* "little sack," Italian *foglietta* "little leaf," etc.) or of Latin *-īnus* (preferred in Italian: *tavolina* "little table, desk," *signorina* "young lady"; and Portuguese: *copinho* "little drinking glass," *senhorinha* "young lady"). The Latin *-ōne* suffix has, conversely, acquired augmentative meaning in several languages (Romanian *căloiu,* Italian *cavallone,* Spanish *caballón* "large horse").

Other frequent suffixes sometimes have a "learned" modern form alongside the older "popular" one; *e.g.,* Latin *-atione:* Italian *-agione / -azione,* French *-aison / -ation,* Spanish *-azón / ación,* Portuguese *-azão / ação;* also Romanian *-ăciune,* and Occitan *azó.*

Suffixes that remain extremely productive include the Latin verbal adjectival *-bilis* (not found in Romanian): Italian *bastevole* "enough," French *admirable,* Spanish *amable* "pleasing"; and verbal nominal *-mentum:* French *abonnement* "subscription," Spanish *cobijamiento* "lodging," Italian *abboccamento* "interview, parley," Romanian *acoperămînt* "cover."

Prefixing of modifying elements remains frequent in all languages (Italian *autostrada* "highway," Spanish *contraveneno* "antidote," French *photocopie* "photocopy"), although some older prefixes may hardly be recognized as such today. The "repetitive" verbal prefix *re-* remains particularly active (Romanian *răsări* "rebound," Italian *ricattare* "to recover," French *racheter* "to buy back," etc.).

Compound words, though less frequent than in the Germanic languages, are not uncommon (*e.g.,* French *cheflieu* "principal town," Italian *primavera* "spring," Spanish *lavamanos* "wash basin").

Adverb formation

Originally a compounding process, the most common method of forming adverbs from adjectives (suffixing of Latin *mente* "mind") has become in most languages a morphological process, although Spanish and Portuguese retain traces of the earlier stage in phrases such as *severa e* (*y*) *cruelmente* "severely and cruelly."

Among the syntactic means that most Romance languages use to extend vocabulary is the potent device, unavailable to Latin, of juxtaposing to any part of speech an article or other determiner and using it as a noun (*e.g.,* Italian *il perchè* "the reason," Spanish *lo útil* "utility, something useful," French *un je ne sais quoi* "an I-don't-know-what"). In French and Spanish, verbal infinitives are frequently so treated (*le devoir* "duty," *el poder* "power," etc.); Romanian also uses infinitives as verbal nouns, but they are differentiated formally by retaining the full form (*e.g., cîntare* "singing"), compared with the shortened verbal form (*cînta*). In earlier stages of most Romance languages the verbal root (most often as it appears in the 3rd-person singular present indicative) could be used as a noun, a process known as back-formation (compare Romanian *laudă* "praise," Italian *domanda* "question," French *approche* "approach," *désir* "desire," Spanish *baila* "dance," Portuguese *muda* "change").

Just as former adjectival forms are frequently used as substantives, so are nouns used with adjectival function; there seem to be few restrictions on this use, though practice varies as to whether agreement should be made (French *les frères ennemis* "enemy brothers," with agreement; *une femme médecin* "a woman doctor," without agreement). Past-participial forms normally act as adjectives, as in English.

Romance makes use of gender classification to extend and modify its vocabulary, especially by relating the gender markers to sex differences (*e.g.,* Romanian *nepot, nepoată;* Occitan *nebut, nebudo,* Spanish *nieto, nieta,* Portuguese *neto, neta,* Catalan *net, neta* "nephew, niece," with Italian invariable *nipote,* and French lexically differentiated *neveu, nièce*). Modern French makes particularly fruitful use of gender differences (originally via ellipse); thus, *le* (*vin de*) *champagne* (the drink) / *la Champagne; La Normandie* (the province) / *le Normandie* (the ship).

Importance of Latin in vocabulary formation

Vocabulary. The basic vocabularies (the most frequently used lexical items) of all the Romance languages are in the main directly inherited from Latin. This applies equally to "function" words, such as *de* "of, from" (Romanian *de,* Italian *di,* Rhaetian *da,* French *de,* Spanish *de,* Portuguese *de*), as to common lexical items, such as *facere* "to do" or *aqua* "water" (Romanian *a face, apă,* Italian *fare, acqua,* Logudorian *fágere, abba,* Engadine *fer, ova,* French *faire, eau,* Catalan *fer, aygua,* Spanish *hacer, agua,* Portuguese *fazer, água*). In some cases different Romance languages inherit words perhaps from different strata of Roman society. Thus, for "lamb," forms derived from Latin *agnus* remain in southern Italy and Galician (*año*), but forms derived from diminutive *agnellus* prevail in Romanian (*miel*), Italian (*agnello*), French (*agneau*), Rhaetian (Engadine *agné,* Friulian *añel*), Occitan (*anhel*), and Catalan (*anyell*), with Sardinian and some Calabrian dialects using another form derived from Latin *agnone* (Logudorian *andzone*). Spanish and Portuguese, however, prefer a derivative of a different word, *chorda* (*cordero, cordeiro*), referring perhaps to the birth process; this word is also found in Occitan and Catalan.

Some words shared by the majority of the Romance languages are not of Latin origin but were probably borrowed from other languages before Latin unity was disrupted. These include especially words of Celtic origin, such as Latin *carrum* "cart," Romanian *car,* Italian *carro,* Logudorian *karru,* Rhaetian *k'ar,* French *char,* Occitan and Catalan *car,* Spanish and Portuguese *carro.*

In Christian Latin a great many Greek ecclesiastical terms were borrowed, which survived in most Romance languages. For example, the Greek word *episkopos* (literally, "overseer") was borrowed into Latin as *episcopus* "bishop," which gave rise to Vegliot *pasku,* Logudorian *pískamu,* Italian *vescovo,* Engadine *ovaisch,* Friulian *veskul,* French *evêque,* Occitan *avesque,* Catalan *bisbe,* Spanish *obispo,* and Portuguese *bispo.*

Germanic words did not penetrate into Latin very frequently before the separation of the various Romance languages from Latin, so that few of them have more than limited extension. Only one Germanic word is known for certain to be found in both Eastern and Western Romance—*sapōne* "soap," recorded in Pliny and occurring as Romanian *săpun,* Vegliot *sapaun,* Italian *sapone,* Logudorian *sabone,* Engadine *savum,* French *savon,* Occitan and Catalan *sabó,* Spanish *jabón,* and Portuguese *sabão.*

"Learned" words

Many Latin words are widespread throughout the Romance languages even though they do not date back directly to the imperial period; these are the "learned" words that have freely entered the languages at virtually every period, borrowed from Latin used as a scholarly language. Because of this later borrowing, such words as *capital, natura, adulterium,* and *discipulus* appear in Romance virtually unchanged from Latin, as they do in other European languages; Romance Latinisms, however, are quite normally used in contexts in which similar words would sound stilted and pedantic in English (*e.g.,* French *supprimer* "suppress" but often used to mean "to do away with").

However similar the Romance vocabularies are to each other, considerable differences nevertheless exist. Some of these may be traced back to imperial times, when provinces may have developed their own vocabulary preferences. For instance, for "oak" Eastern Romance seems to have preferred Latin *quercus* (Logudorian *kerbu,* southern Italian *quercia,* etc.), whereas the West preferred the alternative *robur* (Italian *rovere,* Occitan and Catalan *roure,* Spanish and Portuguese *roble,* Old French *rouvre*—modern French *chêne* is of Celtic origin, while Romanian *stejar* is perhaps of Balkan origin). In some cases the conservative peripheral areas have retained a word that was displaced in more central regions; thus, for "beautiful," *formosus* is preferred in Romanian (*frumos*), Spanish (*hermoso*), and Portuguese (*formoso*), whereas *bellus* is more popular in Vegliot (*bial*), Italian (*bello*), Rhaetian (*bal, biel*), French (*beau*), Occitan (*bel*), and Catalan (*bell*).

Borrowing from substratum and superstratum languages

When Romance borrowed vocabulary from the substratum, differentiation must have taken place early (certainly before the indigenous languages died out). Thus, Spanish *vega,* Portuguese *veiga* "wooded ground by a river" (probably from a non-Indo-European Iberian language, compare Basque *ibaiko* "riverbank"), French *charrue* "plow," *borne* "boundary stone" from Celtic, and Romanian *barză* "stork" (perhaps from Dacian, compare Albanian *bar*) probably were used during Roman times in some form. The debt of Romance vocabulary to substrata languages is probably great but difficult to estimate with any certainty. When there is no known source form or cognate for a word, scholars often suggest an Iberian, Dacian, Ligurian, or Gaulish origin, but, as little is known of these languages, some such theories are mere speculation.

After the influx of barbarian invaders, Romance vocabularies differentiated further as each borrowed from its own superstratum (language superimposed upon Romance). French, for instance, is estimated to have taken some 700 words from Frankish (a Germanic language), not all of which have survived but some of which have passed via French into other Romance languages. Many of these were concerned with agriculture (*jardin* "garden," *houe* "hoe," *blé* "wheat," *gerbe* "sheaf," etc.) or with war (*guerre* "war," *héaume* "helmet") or social organization (*sénéchal* "seneschal," *chambellan* "chamberlain," *maréchal* "marshal," *baron* "baron"). The occupation of much of northern Italy by speakers of Langobardic (also a Germanic language) left less of a mark on Italian vocabulary, though dialects retain more words (estimated at about 300) than the standard language. Standard Italian borrowed little in the way of administrative or military terms but accepted a number of words from rural life (*melma* "mud," *zecca* "sheep tick," *stamberga* "hut," etc.). The Visigoths, who occupied Iberia, were more Romanized than the other Germanic invaders and indeed had abandoned their Germanic tongue by the 7th century AD. Thus, borrowings from Visigothic into Spanish and Portuguese are less frequent, though still not inconsiderable; some (such as *estaca* "stake," *brotar* "to bud") are common to all the Iberian Peninsular languages.

Slavic infiltration into the Balkans led Romanian to adopt a very large number of Slavic words, some in the basic part of the vocabulary. At exactly what stage in history they were borrowed is uncertain, for the earliest Romanian texts, of the 16th century AD, are saturated with Slavic terms from different dialectal sources, though South Slavic predominates. Possibly the borrowings occurred in the 9th century, when the Hunnish Bulgarians, who had adopted Slavic speech, established a powerful state and embraced Christianity, and Slavic pressures were already very strong. Among common Romanian words of Slavic origin one may mention *a trăi* "to live," *hrană* "food," *ceas* "hour," *bogat* "rich," *prieten* "friend," *a munci* "to work." The Magyars (modern Hungarians) also lent a smaller number of words to their Romanian neighbours (*e.g., oraş* "town").

Words borrowed from Arabic

Islāmic invaders into Europe from the 8th century had considerable effect on the vocabulary of the Western Romance languages, even though occupation was confined to southern regions. With its superior cultural and agricultural skills, the Arab world had much to teach Europe of the Dark Ages. Words entered via two routes, Sicily and Spain, and usually their form gives clues about their provenance—if the Arabic definite article (*al*) has coalesced with the root, the word is from Moorish Spain (thus Spanish *algodón* "cotton," Portuguese *algodão,* Old French *auqueton* via Spain, but Italian *cotone,* French *coton* via Sicily). The Arabs introduced into Europe many exotic plants and fruits and with them their names, such as oranges (Spanish *naranja*), lemons (Spanish *limón*), and artichokes (Spanish *alcachofa,* Italian *carciofo*). In some cases the Iberian Peninsula has adopted the Arabic word for such plants, while other languages prefer words of other origin—"rice" is *arroz* in Spanish and Portuguese, *arròs* in Catalan, but Italian and French prefer a Greek word (*riso, riz*), as do Vegliot (*rize*), Rhaetian (Friulian *ris*), and Romanian (*orez*). Apart from the numerous Arabic words known throughout Romance (especially "arithmetic," "algebra," and the like), many are peculiar to the Hispanic languages, such as administrative terms such as Spanish *alcalde* "mayor" or *alguacil* "senior police officer," commercial terms such as *almacén* "warehouse, department store," as well as everyday words such as *ahorrar* "to save," *alboroto* "noise."

Many of the words individual languages borrowed from other sources or fashioned themselves from native sources did not remain private property for long. Interchange among the Western languages has been common since the earliest times and especially from the 16th century. Perhaps French has been the greatest supplier of words throughout the ages, often displacing native words. But French, too, has borrowed heavily from the other languages, especially when they have been purveyors of new objects (such as *patate, banane, tabac,* introduced into Europe by Spanish and Portuguese explorations) or of special cultural values (Italian musical and architectural terms, as well as words to do with banking). Borrowing into minor languages from prestigious neighbours has, naturally, been prolific. Passage of words in the other direction is rare and usually employed for comic or other emotive effect (though Occitan in its heyday supplied a good many words of all sorts—even, it is said, *amour* "love" to French).

Anglicisms

Borrowings from non-Romance languages are less frequent and often frowned on by purists, but far from negligible. Any contact in specialized spheres has produced a crop of loanwords, especially since the 17th century, when French in particular began to borrow a fair number from its Germanic neighbours. In recent times, the influx of anglicisms has become a flood, resisted to the death by some purists. Many are, however, ephemeral or specialized, and none affects the basic vocabulary in which Latin-inherited words continue to predominate.

Orthography. Today the Romance languages are all written in the Latin alphabet, with certain modifications, though until the mid-19th century Romanian was normally written in Cyrillic (still used in Moldavia), and, in the Middle Ages, Arabic script was used for some Spanish dialects.

As soon as scribes first made attempts to write in vernacular Romance, they found the resources of the Latin alphabet inadequate to represent the non-Latin sounds of their spoken language. One device used to overcome these difficulties was to add the letter *h* to another, to indicate a deviant pronunciation: thus, *ch* might represent the *ch* sound in Spanish (*e.g., muchacho* "boy") or the *sh* (earlier *ch*) sound in French (*e.g., chef* "chief"). *C* would normally be used for the *k* sound (before *a, o, u*) or an *s* or *th* sound (before *i, e*). In Italian, conversely, *ch* serves to distinguish the *k* sound, followed by *e*, from the *ch* sound (compare *che* [key] "that, who" with *c'è* [chey] "there is"). *H* was also sometimes added to *n* and *l* to indicate a palatal pronunciation (similar to the *ny* in English "canyon" and *li* in "scallion"), as today in Portuguese *vinho* "wine" and *filho* "son." Another device frequently used to stretch the capacity of the Latin alphabet was to distinguish the letters *i* and *j, u* and *v*, which were originally each single letters *i* (with variant form *j*) and *v* (with variant form *u*, and in Latin pronounced *u* or *w*). In Romance, *v* and *j* came to represent consonants, while *u* and *i* retained their vowel values.

The palatal consonants *n* and *l* are also often depicted by doubled letters or other combinations of letters: the palatal *n* as *nn* (or its scribal variant *ñ*), *gn, nj,* or *in;* palatal *l* as *ll, gl, lj, il,* or *yl,* as well as the combinations *nh* and *lh,* already mentioned. The Latin letter *x,* an abbreviation for *ks,* was also put to other uses in Romance; in Portuguese, Catalan, Sicilian, and Old Spanish it represents an *sh* sound, in modern Spanish a strong *h* sound, more commonly spelled with a *j,* and in northern Italian dialects the *z* sound. Other letters pressed into use for new consonantal sounds were *z* (used in Italy for *ts* and *dz* sounds, Germanic *k* and *w,* and the Visigothic *ç* for *ts* and sometimes *s,* as today in French and Portuguese).

Diacritical markings

Vowels were less of a problem for early Romance scribes—diphthongs were simply shown as vowel combinations such as *ie, uo.* Later, the diaeresis (¨) was sometimes introduced to distinguish diphthongs from adjoining vowels that were to be pronounced separately. Non-Latin vowels are rarely clearly distinguished: French *u* (pronounced like German *ü*) for instance, was written *u* and not consistently distinguished from Latin *u* (pronounced as in "lunar" and, in modern French, written *ou*). Nasal vowels in French are marked by a following *n* or *m;* in Portuguese a tilde (~) is often used for final nasal vowels and diphthongs (*ã, ãi*). Use of diacritics was not consistent until modern times; thus, so-called long and short *e,* still not always distinguished in Italian, are shown as *é* and *è* or *ê* (e.g., *élève* "student") in French (since the 18th century) and as *e* and *é* in Portuguese (since about 1930). Romanian established the use of *î* and *ă* only in the 20th century.

In most of the languages with a long history of writing, the original attempts by scribes at phonological transcription were followed by an "etymological" period in which Latinized spelling gained ground. Castilian was least subject to this fashion, and, because its phonology has changed comparatively little since the Middle Ages (when its spellings became more or less fixed), it has few orthographical problems today. Standard Italian retains a fairly etymological orthography that covers up various minor regional differences of pronunciation; small reforms have been made through the centuries (in the 17th century, for instance, the use of *h*—except in *ho, ha, hanno*—was discontinued; in the 20th century, *î* and *j* for *ii,* as in *studii,* have virtually disappeared), but chaos still reigns in the use of accents. Romanian suffered from etymologizing orthography in the 19th century, but successive edicts of the Romanian Academy, of which the most important was dated 1932, have established a more or less phonetic spelling (the notable exception being the depiction of final "soft" consonants by a following *i*). Modern Catalan, like other "minor" languages, has had the aid of expert linguists in the establishment of its orthography. A standard was proposed by Antonio María Alcover Sureda, a Catalan priest, philologist, and writer from Majorca, in 1913, which is accepted, with small variations, by most writers.

Only two of the Romance languages, French and Portuguese, have had major orthographical problems, mainly resulting from the radical transformations that have affected their phonology since the Latin period. Portuguese has attempted to overcome its difficulties by a series of governmental reforms during the 20th century, but, in spite of official agreements between Portugal and Brazil in 1931 and 1945, there is still little consistency in usage, with Brazilian writers, especially, remaining more conservative (*i.e.,* etymological). In France, in spite of vociferous demands for reform since the 16th century, only minor changes have been accepted (usually originally from unofficial sources, such as printers), so that French orthography today reflects 12th-century phonology, overlaid by the etymologizing of Middle French legal scribes. Battles still rage between the reformers, who deplore the absurdly large proportion of school time devoted to teaching spelling, and the defenders of tradition, who point out that the phonological character of French, with no consistent phonetic markers for the word, make it unsuitable for phonetic transcription and that written French has its own structure, not identical with that of the spoken language.

(Re.P.)

Germanic languages

The Germanic languages, a branch of the Indo-European language family, include a number of extinct languages as well as the earlier and present forms of German, Netherlandic, Afrikaans, English, Frisian, the Scandinavian languages, Yiddish, and their many dialects.

In numbers of native speakers, English, with 285,000,000, clearly ranks second among the languages of the world (after Chinese); German, with 98,000,000, probably ranks seventh (after Hindi-Urdu, Spanish, Russian, and Japanese). To these figures may be added those for persons with another native language who have learned one of the Germanic languages for commercial, scientific, literary, or other purposes. English is unquestionably the world's most widely used second language.

Table 14 presents information on each of the modern standard Germanic languages.

The earliest historical evidence for Germanic is provided by isolated words and names recorded by Latin authors beginning in the 1st century BC. From *c.* AD 200 there are Scandinavian inscriptions carved in the 24-letter runic alphabet. The earliest extensive Germanic text is the (incomplete) Gothic Bible, translated *c.* AD 350 by the Visigothic bishop Ulfilas (Wulfila), and written in a 27-letter alphabet of the translator's own design. Although the Gothic alphabet was hardly used outside of this Bible translation, later versions of the runic alphabet were used sparingly in England, Germany, and particularly Scandinavia—in the latter area down to early modern times. All extensive later Germanic texts, however, use adaptations of the Latin alphabet.

The names and approximate dates of the earliest recorded Germanic languages are recorded in Table 15.

The Germanic languages are related in the sense that they can be shown to be different historical developments

Table 14: Modern Standard Germanic Languages

	where spoken	native speakers (1981 estimate)	use as a 2nd language
English	Great Britain, Ireland, United States, Canada, Australia, New Zealand, Republic of South Africa	285,000,000	extreme
German	Germany, Austria, Switzerland (part)	98,000,000	extensive
Netherlandic (Dutch-Flemish)	The Netherlands, Belgium (part)	20,000,000	moderate
Swedish	Sweden, Finland (part)	9,000,000	slight
Danish	Denmark	5,400,000	slight
Norwegian	Norway	4,800,000	slight
Yiddish	various countries	4,000,000	slight
Afrikaans	Republic of South Africa (part)	3,300,000	slight
Frisian	The Netherlands, Germany	445,000	—
Icelandic	Iceland	218,000	—
Faeroese	Faeroe Islands	43,000	—

of a single earlier parent language. Although for some language families there are written records of the parent language (*e.g.*, for the Romance languages, which are variant developments of Latin), in the case of Germanic no written records of the parent language exist. Much of its structure, however, can be deduced by the comparative method of reconstruction (a reconstructed language is called a protolanguage; reconstructed forms are marked with an asterisk). For example, a comparison of Runic *-gastiR,* Gothic *gasts,* Old Norse *gestr,* Old English *giest,* Old Frisian *iest,* and Old Saxon and Old High German *gast* "guest" leads to the reconstruction of Proto-Germanic **ǥastiz.* Similarly, a comparison of Runic *horna,* Gothic *haurn,* Old Norse, Old English, Old Frisian, Old Saxon, and Old High German *horn* "horn" leads scholars to reconstruct the Proto-Germanic form **hornan.*

Recon-struction of parent language

Such reconstructions are, in part, merely formulas of relationship. Thus the Proto-Germanic **o* of **hornan* in this position gave *au* in Gothic and *o* in the other languages. In other positions (*e.g.*, when followed by a nasal sound plus a consonant) **o* gave *u* in all the languages: Proto-Germanic **dumbaz,* Gothic *dumbs,* Old Norse *dumbr,* Old English, Old Frisian, and Old Saxon *dumb,* Old High German *tumb* "dumb." What may be deduced is that this vowel sounded more like *u* in some environments, but like *o* in others; it may be written as **u~o,* indicating that it varied between these two pronunciations.

The above example shows that such reconstructions are more than mere formulas of relationship; they also give some indication of how Proto-Germanic actually sounded. Occasionally scholars are fortunate enough to find external confirmation of these deductions. For example, on the basis of Old English *cyning,* Old Saxon and Old High German *kuning* "king," the Proto-Germanic **kuningaz* can be reconstructed; this would seem to be confirmed by Finnish *kuningas* "king," which must have been borrowed from Germanic at a very early date.

By pushing the comparative method still farther back, it can be shown that Germanic is related to a number of other languages, notably Celtic, Italic, Greek, Baltic, Slavic, Iranian, and Indo-Aryan (Indic). All of these language groups are subsequent developments of a still earlier parent language for which there are, again, no written records but which can be reconstructed as Proto-Indo-European (see above *Indo-European languages*).

Table 15: Earliest Recorded Germanic Languages

	dates* (AD)
Early Runic	200–600
Gothic	350
Old English (Anglo-Saxon)	700–1050
Old High German	750–1050
Old Saxon (Old Low German)	850–1050
Old Norwegian	1150–1450
Old Icelandic	1150–1500†
Middle Netherlandic	1170–1500†
Old Danish	1250–1500†
Old Swedish	1250–1500†
Old Frisian	1300–1500†

*Indicates approximate range of dates.
†Cutoff date for beginnings of modern Germanic languages.

CHARACTERISTICS

The special characteristics of the Germanic languages that distinguish them from other Indo-European languages result from numerous changes, both phonological and grammatical.

Grimm's law

Phonology. *Consonants.* Proto-Indo-European had 12 stop consonants: *p, t, k,* k^w*; b, d, g,* g^w*; bh, dh, gh,* g^wh*;* and one sibilant, *s.* (Stops are produced with momentary complete stoppage of the breathstream at some point in the vocal tract.) By a change known as the Germanic consonant shift (or Grimm's law, after the German scholar Jacob Grimm, who was one of the first to describe it), the 12 stops changed in Germanic to voiceless fricatives, voiceless stops, and voiced fricatives, as illustrated in Table 16. A few examples: (1) Proto-Indo-European *p, t, k,* and k^w, as in Latin *piscis, tenuis, centum,* and *quod,* became Proto-Germanic *f, p, x,* and x^w, as in English "fish," "thin," "hund(red)," and "what." Proto-Germanic *x* and x^w early became *h* and h^w in some positions, giving the alternations of *h~x* and $h^w \sim x^w$. (2) Proto-Indo-European *d* and *g,* as in Latin *decem* and *genus,* became Proto-Germanic *t* and *k,* as in English "ten" and "kin." (3) Proto-Indo-European *bh, dh,* and *gh,* as in Sanskrit *bhū-, dhā-,* and *(g)hā-,* became Proto-Germanic *ƀ, ð,* and *ǥ,* which later changed to the stops *b, d,* and *g* in some positions (*e.g.*, English "be," "do," and "go"), giving *b~ƀ, d~ð,* and *g~ǥ.* Proto-Indo-European *s,* as in Latin *sedeō,* was unchanged; Proto-Germanic kept *s,* as in English "sit."

Table 16: Sound Changes in the Germanic Consonant Shift

Proto-Indo-European voiceless stops	p	t	k	k^w
Proto-Germanic voiceless fricatives	f	þ	x	x^w
Proto-Indo-European voiced stops	b	d	g	g^w
Proto-Germanic voiceless stops	p	t	k	k^w
Proto-Indo-European voiced aspirated stops	bh	dh	gh	g^{w}h
Proto-Germanic voiced fricatives	ƀ	ð	ǥ	ǥ^w

In addition to these general changes, there were two special ones. (1) Proto-Indo-European *p, t,* and *k* remained voiceless stops when preceded by *s* or another stop; thus, Proto-Indo-European *sp, st, sk, pt,* and *kt* gave Proto-Germanic *sp, st, sk, ft,* and *xt,* respectively. For example, Proto-Indo-European *sp* and *st,* as in Latin *spuō* and *hostis,* remained *sp* and *st* in Proto-Germanic, as in English "spew" and "guest"; Proto-Indo-European *pt* and *kt,* as in Latin *captus* and *octō,* became Proto-Germanic *ft* and *xt,* respectively, in Old English *hæft* "captured" and *eahta* "eight." (By still another change, Proto-Indo-European *tt* gave Proto-Germanic *ss; e.g.,* Sanskrit *sattá-,* Old English *sess* "seat.") (2) By a change known as Verner's law (named for the Danish scholar Karl Verner, who first described it), early Germanic voiceless *f, þ, x,* x^w, and *s* (from Proto-Indo-European *p, t, k,* k^w, and *s*) were voiced to *ƀ, ð, ǥ,* ǥ^w, and *z,* respectively, when they followed an unaccented syllable, and the first four of these thereby merged with the already existing *ƀ, ð, ǥ,* and ǥ^w (from Proto-Indo-European *bh, dh, gh,* and g^wh). Thus, Proto-Indo-European **bhrā́tēr* became Proto-Germanic **brōþēr* (with *þ* after an accented syllable) and Old English *brōþor* "brother"; but by Verner's law Proto-Indo-European **mātḗr* became Proto-Germanic **mōðēr* (with *ð* after an unaccented syllable) and Old English *mōdor* "mother." (The *th* of modern English "mother" is the result of a subsequent change.)

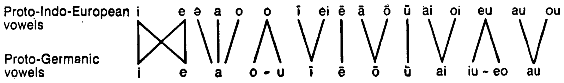
Verner's law

These changes gave the following Proto-Germanic system of consonants: voiceless stops and fricatives, *p, f, t, þ, k, h~x,* k^w, $h^w \sim x^w$*;* voiced stops and fricatives, *b~ƀ , d~ð, g~ǥ,* (g^w~ǥ^w); sibilants, *s, z;* nasals, *m, n;* liquids, *l, r;* and semivowels, *w, j(y).* The sound alternation of g^w~ǥ^w is parenthesized because it early became either *g~ǥ* or *w.* The sounds k^w and $h^w \sim x^w$ occurred as such more or less clearly only in Gothic; elsewhere they became the sequences *kw* and *hw~xw*, or the labial element ^{w} was lost. All remaining consonants except *z* occurred between vowels both singly and doubly (*e.g.*, *-p-* and *-pp-*, *-t-* and *-tt-*).

Ablaut

Vowels. In addition to the above consonants (12 stops and the sibilant *s*), Proto-Indo-European also had vowels and resonants. The vowel of any given root was not necessarily fixed but varied in an alternation called ablaut. Thus, the root that means "sit" was alternately **sed-*, *sod-*, **sd-*, **sēd-*, and **sōd-* (English "sit" is from **sed-*, "sat" from **sod-*, and "seat" from **sēd-*); and the root that means "do" was **dhē*, **dhō-*, and **dhə-* (English "deed" is from **dhē-*, and "do" is from **dhō-*). Other Proto-Indo-European vowels were *a*, *ā*, *ī*, and *ū*. The Proto-Indo-European resonants, which functioned as vowels in some positions and as consonants in others, were *i*, *u*, *m*, *n*, *l*, and *r*. Thus, **bhr̥tó-* (Sanskrit *bhr̥tá-* "borne") had syllabic *r̥* (*i.e.*, *r̥* functioning as a vowel), but **bhéreti* (Sanskrit *bhárati* "he bears") had nonsyllabic *r* (*i.e.*, *r* functioning as a consonant).

This Proto-Indo-European system of vowels contrasting with resonants was reshaped in Germanic by a number of changes. Syllabic *i*, *u*, *m̥*, *n̥*, *l̥*, and *r̥* became in Proto-Germanic the vowels *i* and *u* and the sequences *um*, *un*, *ul*, and *ur*, respectively; nonsyllabic *m*, *n*, *l*, and *r* developed into the nasals and liquids *m*, *n*, *l*, and *r*, respectively; nonsyllabic *i* and *u* before vowels resulted in the semivowels *j* (also symbolized as *y*) and *w*, though after vowels they continued to form diphthongs (*ei*, *ai*, *oi*; *eu*, *au*, *ou*). The Proto-Indo-European vowels and diphthongs then changed into Proto-Germanic sounds as follows:


Proto-Indo-European vowels
i e ə a o ō ī ei ē ā ō ū ai oi eu au ou
Proto-Germanic vowels
i e a o~u ī ē ō ū ai iu~eo au

In this diagram the lines between two sounds indicate that the Proto-Indo-European sound developed into the corresponding Proto-Germanic sound; for example, Proto-Indo-European *i* became either *i* or *e*, and Proto-Indo-European *ə*, *a*, and *o* coalesced in Proto-Germanic as *a*. These changes gave the following vowels for early Proto-Germanic: short vowels, *i*, *e*, *a*, *u~o*; long vowels, *ī*, $\bar{e}^1$, *ū*, *ō*; diphthongs, *ai*, *au*, *iu~eo*. The vowel *ē* is noted here as $\bar{e}^1$ because Proto-Germanic had (or developed) a second $\bar{e}^2$ of uncertain origin. In Gothic the two *ē*'s merged. Elsewhere $\bar{e}^2$ remained a midvowel, but $\bar{e}^1$ was lowered; thus, for example, $\bar{e}^2$ in Old Saxon *hēr* "here" but lowered $\bar{e}^1$ in Old Saxon *dād* "deed." In addition to the above oral vowels, Proto-Germanic also had three nasalized vowels: long *ĩ*, *ã*, and *ũ*, which arose when, in the sequences *inx*, *anx*, and *unx*, the *n* was lost with nasalization and lengthening of the preceding vowel.

Accent. Proto-Indo-European had a variable pitch accent that could fall on any syllable of a word (*e.g.*, on the first syllable in **bhrā́tēr* "brother" but on the last syllable in **mātḗr* "mother." This was replaced in Germanic by a fixed stress accent that always fell on the first syllable: **brṓþēr*, **mṓðēr*. One effect of this strong initial stress seems to have been the progressive weakening and loss of unstressed final syllables; *e.g.*, Proto-Indo-European *sodéionom*, Proto-Germanic **sátjanan*, Old English *settan*, Middle English *sette(n)*, and modern English (to) "set."

Strong initial stress

Strong initial stress is also reflected in the basic unit of old Germanic poetry, which consisted of two half lines, each with one of a small number of stress patterns, linked by the alliteration of stressed initial consonants or vowels (*e.g.*, from *Beowulf*: *Bḗo-wùlf wæs brḗme / bl ǣ́d wíde spràng* "Beowulf was famous; his renown went far").

Grammar. *Declensions.* Proto-Germanic kept the Proto-Indo-European system of three genders (masculine, feminine, neuter) and three numbers (singular, dual, plural), though the dual was becoming obsolete. It reduced the Proto-Indo-European system of eight cases to six: nominative, accusative, dative, genitive, instrumental, and vocative, though the last two were becoming obsolete. In the adjective declensions there were two innovations: (1) To the Proto-Indo-European vowel types (*o-*, *ā-*, *i-*, and *u-*stems), it added some pronominal endings to give the Germanic "strong" adjective declension. (2) It extended the Proto-Indo-European *n*-stem endings to all adjectives to give the Germanic "weak" adjective declension. Contrast, in modern German, strong *gutes Bier* "good beer" with weak *das gute Bier* "the good beer."

Strong and weak verbs

Conjugations. The Proto-Indo-European verb seems to have had five moods (indicative, imperative, subjunctive, injunctive, and optative), two voices (active and mediopassive), three persons (1st, 2nd, and 3rd), three numbers (singular, dual, and plural), and several verbal nouns (infinitives) and adjectives (participles). In Germanic these were reduced to indicative, imperative, and subjunctive moods; a full active voice plus passive found only in Gothic; three persons; full singular and plural forms and dual forms found only in Gothic; and one infinitive (present) and two participles (present and past). The Proto-Indo-European tense-aspect system (present, imperfect, aorist, perfect) was reshaped to a single tense contrast between present and past. The past showed two innovations: (1) In the "strong" verb Germanic transformed Proto-Indo-European ablaut into a specific tense marker (*e.g.*, Proto-Indo-European **bher-*, **bhor-*, **bhēr-*, **bhr̥-* in Old English *beran* "bear," past singular *bær*, past plural *bǣron*, past participle *boren*). (2) In the "weak" verb Germanic developed a new type of past and past participle (*e.g.*, Old English *fyllan* "fill," past *fylde*, participle *gefylled*). Weak verbs fell into three classes depending on the syllable following the root (*e.g.*, Old High German *full-e-n* [from **full-ja-n*] "fill," *mahh-ō-n* "make," *sag-ē-n* "say"). Gothic also had a fourth class: *full-nō-da* "it became full."

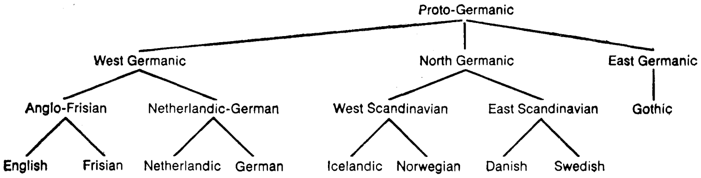
Many Proto-Germanic strong verbs showed a consonant alternation between *f* and *ƀ*, *þ* and *ð*, *x* and *ǥ*, and *s* and *z* that was the result, through Verner's law, of the alternating position of the Proto-Indo-European accent. The forms in Table 17 illustrate changes resulting from Verner's law. In this particular word, English has generalized the *s* (now *z*): "freeze," "froze," "frozen." German has generalized the *z* (now *r*): *frieren*, *fror*, *gefroren*. And Netherlandic still shows the alternation: *vriezen*, *vroor*, *gevroren*. English has kept the alternation in only one verb: singular "was," plural "were." Traces of it still survive, however, in a few now isolated forms: "seethe" (Proto-Germanic *þ*) and its old past participle "sodden" (Proto-Germanic *þ*); "lose" (Proto-Germanic *s*) and its old past participle "(for)lorn" (Proto-Germanic *z*).

Table 17: Illustration of Verner's Law

Proto-Indo-European	Proto-Germanic	Old English	English translation
*préusonom	*freosanan	frēosan	"(to) freeze"
*próuse	*fraus	frēas	"(it) froze"
*prusn̥ t́	*fruzun	fruron	"(they) froze"
*prusénos	*frozenaz	froren	"frozen"

*Unattested, reconstructed form.

Branches of Germanic. Like every language spoken over a considerable geographic area, Proto-Germanic presumably consisted of a number of geographical varieties or dialects, which, in the course of time, developed in different ways to give the different early and modern Germanic languages. Late-19th-century scholars used a family tree diagram to show this splitting into dialects and the relationships among the dialects:



Though there is much truth in such a diagram, it overemphasizes the notion of "splits" into separate "branches" and obscures the fact that the transition from one dialect to another may be gradual rather than abrupt. Modern Netherlandic and German, for example, constitute a single speech area at the level of local dialects; they have "split" only in the sense that they have developed different standard languages.

Dialect similarities

Mid-20th-century scholars, using the findings of archaeology and the methods of geographical linguistics, attempted to correct the distortions of this family-tree model by noting also the linguistic features shared by two or more dialect areas. Archaeological evidence suggests that a relatively uniform Germanic people at *c.* 750 BC were located in southern Scandinavia and along the North Sea and Baltic coasts from The Netherlands to the Vistula. Five hundred years later (*c.* 250 BC) they had spread south, and five general groups are distinguishable: North Germanic in southern Scandinavia, excluding Jutland; North Sea Germanic, along the North Sea and in Jutland; Rhine-Weser Germanic, along the middle Rhine and Weser; Elbe Germanic, along the middle Elbe; and East Germanic, between the middle Oder and the Vistula.

By *c.* AD 250 the division was much the same, though the Elbe group had spread southward to the Danube, and the East Germanic group moved southeast into the Carpathians and beyond. Then, toward the end of the 4th century, began the great Germanic tribal migrations. North Germanic speakers migrated into Jutland, approximately to the modern Danish-German language border; part of the North Sea group crossed the North Sea and conquered much of England; the Elbe group (the later Alamanni, Bavarians, and Langobardi) spread still farther south into part of Switzerland and into Austria and northern Italy; and the East Germanic group left the Oder-Vistula area to begin their many wanderings.

This five-way division of the Germanic peoples is based on archaeological evidence, but it agrees well with deductions that can be made from the earliest linguistic evidence. Five linguistic groups are indeed distinguishable, though they are linked into sets of two, three, or four through shared linguistic innovations.

Linguistic groups

1. North Germanic, North Sea Germanic, Rhine-Weser Germanic, and Elbe Germanic share the change of *z* to *r; e.g.,* Proto-Germanic **maiz-* "more," Gothic *maiza,* contrasting with Old Norse *meire,* Old English and Old Frisian *māra,* Old Saxon *mêro,* and Old High German *mēro.* In addition, they also show *i*-umlaut, as in the raising of *a* to *e* before *j* (pronounced as the *y* in "year"); *e.g.,* Proto-Germanic **satjanan* "set," Gothic *satjan* in contrast to Old Norse *setia,* Old English *settan,* Old Frisian *setta,* Old Saxon *settian,* and Old High German *setzen.* In certain strong verbs they share a new past tense form with *ē*^{2} and without reduplication (the repetition of a part of a word)—*e.g.,* Proto-Germanic **le-lōt* "let," Gothic *lailot* in contrast to Old Norse, Old English, Old Frisian, and Old Saxon *lēt* and Old High German *liez.*

2. North Germanic, North Sea Germanic, and Rhine-Weser Germanic partly share the loss of nasal sounds before voiceless fricative sounds. As noted above, *n* was lost in Proto-Germanic before *x.* North Sea Germanic shows loss of nasals before the remaining fricatives *f, þ,* and *s;* thus, the nasals in Proto-Germanic **fimf* "five," **munþ* "mouth," and **uns* "us" are preserved in Gothic *fimf, munþs,* and *uns,* as well as in Old High German *fimf, mund,* and *uns,* but are lost in Old English *fīf, mūþ,* and *ūs* and Old Frisian *fīf, mūth,* and *ūs.* Rhine-Weser Germanic shows this same loss only sporadically; Old Saxon has *fīf, mūð,* and *ūs,* without the nasals, but also has *andar* (from Proto-Germanic **anþar-* "other"), with the nasal conso-

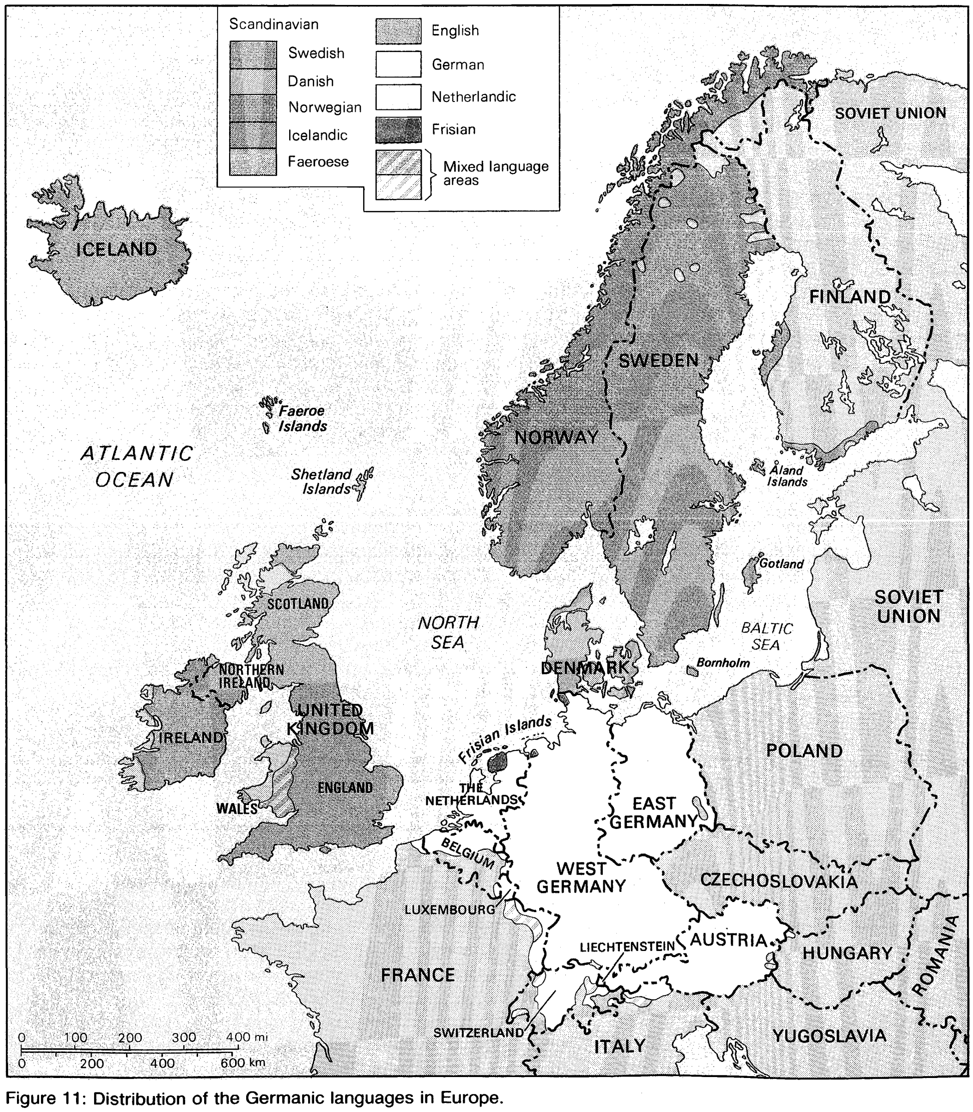


Figure 11: Distribution of the Germanic languages in Europe.

nant, beside *āðar* and *ōðar,* without it. Old Norse, which is North Germanic, shows regular loss of a nasal sound only before *s* (*e.g., oss* "us").

3. North Sea Germanic, Rhine-Weser Germanic, and Elbe Germanic (usually grouped together as West Germanic) share the change of *ð* to *d* in all positions (*e.g.,* Proto-Germanic **blōð-* "blood," and Gothic and Old Norse *blōð-* in contrast to Old English, Old Frisian, and Old Saxon *blōd* and Old High German *bluot*), the loss of *-z* after unstressed vowels (*e.g.,* Proto-Germanic **dagaz* "day," Gothic *dags,* and Old Norse *dagr* in contrast to Old English *dæg,* Old Frisian *dei,* Old Saxon *dag,* and Old High German *tag*), and a 2nd-person-singular past formation in strong verbs different from that of East Germanic and North Germanic (*e.g.,* Proto-Germanic **gaft* "[thou] gavest" occurs in Gothic and Old Norse *gaft,* but Proto-Germanic **gē*^{1}*ƀi* appears in Old English *gēafe,* Old Saxon *gaƀi,* and Old High German *gābi;* Old Frisian has a new analogical form, *iefst*).

In addition, they share the doubling of most consonants in certain positions, especially before *j* (the *y* sound); *e.g.,* Proto-Germanic **satjanan* "set" appears in Gothic as *satjan* and in Old Norse as *setia* but in Old English as *settan,* in Old Frisian as *setta,* in Old Saxon as *settian,* and in Old High German as *setzen.* North Germanic also shows doubling of *g* and *k* before *j* (*y*); *e.g.,* Proto-Germanic **lagjanan* "lay" becomes Gothic *lagjan* but Old Norse *leggia,* Old English *lecgan,* Old Frisian *ledza,* Old Saxon *leggian,* and Old High German *lecken.*

4. North Sea and Rhine-Weser Germanic share a single ending for the 1st, 2nd, and 3rd persons plural of verbs (North Germanic, Elbe Germanic, and East Germanic show two or three different endings), loss of the Proto-Germanic reflexive pronoun **sik,* and loss of *-z* in pronouns (*e.g.,* Proto-Germanic **wīz* or **wĭz* "we," which appears in Gothic as *weis,* in Old Norse as *vēr,* and in Old High German as *wir,* appears in Old English as *wē* and in Old Frisian and Old Saxon as *wī*).

5. North Germanic, Elbe Germanic, and East Germanic share the addition of the Proto-Germanic nominative-accusative neuter singular pronominal ending **-at* in the strong adjective declension (*e.g.,* Proto-Germanic **hailan* "whole," Old English and Old Frisian *hāl,* and Old Saxon *hêl* in contrast to Gothic *heilata,* Old Norse *heilt,* and Old High German *heilaz*).

6. Elbe Germanic and East Germanic share the pronoun reconstructed for Proto-Germanic as **iz* "he": Gothic *is,* Old High German *ir* or *er,* instead of Proto-Germanic **h-* occurring in Old Norse *hann* and in Old English, Old Frisian, and Old Saxon *hē.*

7. North Germanic and East Germanic share the change of *jj* (pronounced as English *yy*) and *ww* to a long stop plus a semivowel (*e.g.,* from Proto-Germanic **twajj-* "of two" and *triww-* "true"—as in Old High German *zweiio* and *triuwi*—to Old Norse *tueggia* and *tryggr,* Gothic *twaddje* and *triggws*).

EAST GERMANIC

The East Germanic languages, all of which have long been extinct, developed from the dialects of the East Germanic group mentioned above; they were spoken by Germanic tribes located between the middle Oder and the Vistula.

History. According to historical tradition, at least some of the Germanic tribes migrated to the mouth of the Vistula from Scandinavia. Little is known of Gepidic, Rugian, and Burgundian; some knowledge of Vandalic, Visigothic, and, especially, Ostrogothic is provided by the names recorded in Greek and Latin writings. The only East Germanic language on which there is extensive information is the Gothic—more specifically, Visigothic—that was spoken along the western shore of the Black Sea around the middle of the 4th century AD. Its special importance lies in the fact that, except for a few scattered runic inscriptions, it is by far the oldest Germanic language preserved.

Gothic language

Knowledge of Gothic is derived primarily from the remains of a Bible translation made for the Visigoths living along the lower Danube by a Visigothic bishop of the Arian church named Ulfilas, who lived during the 4th century. The surviving manuscripts of this translation, which are not originals but later copies thought to have been written in northern Italy during the period of Ostrogothic rule (493–554), include considerable portions of the New Testament and parts of Nehemiah from the Old Testament. Although most of them are palimpsests, manuscripts in which earlier erased writings are found, a handsome exception is the famous Codex Argenteus, written in silver and gold letters on purple parchment and containing (in 188 leaves remaining from an original 330 or 336) portions of the four gospels. Closely related to these biblical manuscripts are eight leaves containing fragments of a commentary (called the *Skeireins* in Gothic) on the Gospel According to St. John. Minor nonbiblical texts include a fragment of a calendar, two deeds containing some Gothic sentences, and a 10th-century Salzburg manuscript that gives the Gothic alphabet, a few Gothic words with Latin transliteration, and some phonetic remarks with illustrative examples.

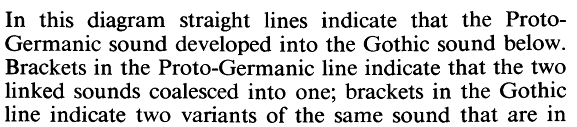
In the 4th and 5th centuries, Gothic (Visigothic and Ostrogothic) must have spread, along with the conquering Goths, at least thinly over much of southern Europe; but there is no evidence for its survival in Italy after the fall of the Ostrogothic kingdom, and in Spain it is doubtful whether the Visigoths retained their language until the Arab conquest. In the 9th century the German monk Walafrid Strabo mentions that Gothic was still being used in some churches near the lower Danube. After that time Gothic seems to have survived only among the Goths of the Crimea, who were last mentioned in the middle of the 16th century by a Flemish diplomat named De Busbecq, who, while on a mission to Constantinople in 1560–62, collected a number of words and phrases showing that their language was still essentially a form of Gothic.

Gothic alphabet

Characteristics. The Gothic alphabet, said to have been created by Ulfilas, contained 27 symbols, two of which functioned only as numbers, while the remaining 25 were used as both numbers and letters. The shape, numerical value, and ordering of the symbols show clearly that the alphabet was based primarily on that of Greek, though a few symbols seem to have been adapted from the Latin alphabet.

Phonology. The Gothic consonant system seems to have been largely identical with that assumed above for Proto-Germanic: *p, t, k,* k^w (this last sound was probably much like the *qu* in "queen"); *f, þ, h,* h^w (this last sound was probably pronounced much like the *wh* in "white"); *b, d, g; s, z; m, n; l, r; w, j.* The nasal *n* was presumably velar before the velar consonants *k, q,* and *g;* in these positions it was usually written (as in Greek) as *g* or *gg.* Examples of this spelling include *dragk* "drank," *igqis* "you two," and *briggan* "bring," although *n* was occasionally used as in Latin (*e.g., þank* "thanks," *inqis* "you two," and *bringiþ* "bring ye").

The Gothic alphabet contained the five simple vowel symbols, *i, e, a, o,* and *u,* from which four compound symbols, *ei, ai, au,* and *iu,* were also made; in addition, *w* was used to transliterate Greek υ and οι (both of which were pronounced as umlauted *u* [ü] in 4th-century Greek). The generally accepted development of the Proto-Germanic vowels in Gothic can be diagrammed as follows:


Proto-Germanic:
i e a o ~ u ī, ī̃ ē², ē¹ ai ā̃ au ō ū, ū̃ eo ~ iu
Gothic:
i ɛ a ɔ u ī ē ɛ̄ ā ɔ̄ ō ū iu
Spelling:
i ai a au u ei e ai a au o u iu

In this diagram straight lines indicate that the Proto-Germanic sound developed into the Gothic sound below. Brackets in the Proto-Germanic line indicate that the two linked sounds coalesced into one; brackets in the Gothic line indicate two variants of the same sound that are in different phonetic environments. Proto-Germanic *i* and *e* apparently first merged as a single vowel and then became Gothic *i* in most positions, but became *ai* before *h,* h^w, and *r.* Similarly, Proto-Germanic **ŏu* became Gothic *u* in most positions, but *au* before *h,* h^w, and *r.*

Special characteristics. Gothic shows a number of archaic features that had been almost or entirely lost by the time the other Germanic languages began to appear in writing; among these are a passive voice and one type of past tense formed with reduplication, a dual number in the 1st and 2nd persons of its verbs and pronouns, and a special vocative case in one noun class. At the same time, Gothic also shows changes from Proto-Germanic, among which are the shortening of most long vowels in final unstressed syllables and the loss of most short vowels (*e.g.*, Proto-Germanic *_erþō_ "earth" became Gothic *airþa,* Proto-Germanic *_stainaz_ "stone" became Gothic *stains*). Finally, voiced fricatives that occurred or came to occur at the end of a word have been unvoiced (*e.g.*, nominative *_hlaiƀaz,_ accusative *_hlaiƀan_ "bread, loaf" changed to *hlaifs* and *hlaif,* respectively [but dative *hlaiba*]).

WEST GERMANIC

The West Germanic languages are those that developed from the North Sea, Rhine-Weser, and Elbe groups mentioned above. Out of the many local West Germanic dialects the following six modern standard languages have arisen: English, Frisian, Netherlandic (Dutch-Flemish), Afrikaans, German, and Yiddish.

English and Frisian changes

English. English and Frisian are descended from North Sea Germanic. The most striking changes that distinguish them from the other Germanic languages are the loss of nasal sounds before the Proto-Germanic voiceless fricatives *f, þ*, and *s* (contrast the following pairs of words, in which English loses the nasal but German preserves it: before *f*—"soft"/*sanft;* before *þ*—"other"/*ander;* before *s*—"us"/*uns,* "goose"/*Gans*); palatalization of Proto-Germanic *k* before front vowels and *j,* giving modern English *ch* (English/German pairs: "chin"/*Kinn,* "birch" [Old English *birce*]/*Birke*); and palatalization of Proto-Germanic *g* before front vowels, giving modern English *y* (English/German pairs include "yield"/*gelten,* "yester-[day]"/*gestern,* "yard" [Old English *geard*]/*Garten;* this palatalized *g* merged with the *j* [*y* sound] from Proto-Germanic *j:* "year"/*Jahr*).

Other changes include palatalization of *gg* before *j* to Old English *cg* (Proto-Germanic *_brugjō,_ pre-Old English *_bruggju,_ Old English *brycg* "bridge"; contrast the unpalatalized *ck* from *gg* of German *Brücke* "bridge"); fronting and raising of *ā* from Proto-Germanic *ē*^{1} (English/German pairs include "deed"/*Tat,* "seed"/*Saat,* "sleep"/*schlafen,* "meal"/*Mahl*); and backing and raising of nasalized *ā,* from Proto-Germanic *ā* and from Proto-Germanic *a* before nasal plus *f, þ*, and *s* (English/German pairs include "brought"/*brachte,* "thought"/*dachte,* "other"/*ander,* and "goose"/*Gans*).

For further information on English, see below *English language.*

Frisian. A thousand years or so ago, Frisian was apparently spoken throughout a North Sea coastal area extending from the modern Netherlands province of Noord-Holland (North Holland) on up to modern German Schleswig and the adjacent offshore islands. During the following centuries, the Frisian of much of this area was gradually replaced by local Netherlandic and Low German dialects, so that Modern Frisian is now spoken in only three remaining areas: (1) West Frisian, in the Netherlandic province of Friesland, including the island of Schiermonnikoog and two-thirds of the island of Terschelling (altogether some 425,000 speakers); (2) East Frisian, in the German Saterland (some 2,000 speakers; this area was apparently settled in the 12th or 13th centuries from the former East Frisian area to the north); and (3) North Frisian, along the west coast of German Schleswig and on the offshore islands of Sylt, Föhr, Amrum, the Halligen, and Helgoland (altogether some 18,000 speakers).

History. The earliest manuscripts written in Frisian date from the end of the 13th century, though the legal documents that they contain were probably first composed, in part, as early as the 11th century. This stage of the language, until about 1575, is known as Old Frisian. The last written document of this period dates from 1573, after which Frisian was hardly used at all as a written language for some three centuries, though it continued to be spoken.

From the start Old Frisian shows all of the features that distinguish English and Frisian from the other Germanic languages. These include loss of the nasal sound before Proto-Germanic *f, þ,* and *s* (*e.g.,* Proto-Germanic *_fimf,_ *_munþ-,_ and *_uns_ became Old Frisian *fīf* "five," *mūth* "mouth," and *ūs* "us"), palatalization of Proto-Germanic *k* before front vowels and *j* (*e.g.,* Proto-Germanic *_kinn-_ and *_lē^{1}kj-_ became Old Frisian *tzin* "chin" and *lētza* "physician" [*cf.* English archaic "leech"]), and palatalization of Proto-Germanic *g* before front vowels (*e.g.,* Proto-Germanic *_geldan-_ became Old Frisian *ielda* "yield"). This merged with the *j* from Proto-Germanic *j,* as in Proto-Germanic *_jē^{1}r-_ or Old Frisian *iēr* "year." In addition, Old Frisian shows palatalization of *gg* from Proto-Germanic *g* before *j* (*e.g.,* Proto-Germanic *_lagjan-,_ with doubling *_laggjan,_ became Old Frisian *ledza* "to lay"); fronting and raising of *ā* from Proto-Germanic *ē*^{1}, as in Proto-Germanic *_dē^{1}ð-, lowered to *dād,_ and raised again to Old Frisian *dēd* "deed"; and backing and raising of nasalized *ã* from Proto-Germanic *ã* and Proto-Germanic *a* before nasal plus *f, þ, s,* as in Proto-Germanic *_brãxt-,_ *_anþar-,_ and *_gans-,_ which became Old Frisian *brocht* "brought," *ōther* "other," and *gōs* "goose."

Around the beginning of the 19th century it appeared that the age-old replacement of Frisian by Netherlandic and German would continue unabated and that the language would soon become extinct. But with 19th-century Romanticism a new interest in local life arose, and societies were formed for the preservation of the Frisian language and culture. Very slowly, the aims of this "Frisian movement" came to be realized, especially in the Netherlands province of Friesland, where in 1937 Frisian was accepted as an optional course in elementary schools; a Frisian Academy was founded in 1938; and in 1943 the first Frisian translation of the Bible was published. Later, in 1955, Frisian was approved as the language of instruction in the first two years of elementary school (though only about one-fourth of all schools use it in this way), and in 1956 the use of Frisian in courts of law was approved.

Status of Frisian

Despite this gradual reemergence of Frisian, Netherlandic still functions as the primary standard language of Friesland. Nearly all school instruction is given in Netherlandic; all daily newspapers are printed in Netherlandic (though they contain occasional articles in Frisian); and all television broadcasts and nearly all radio broadcasts are in Netherlandic. There is a small and enthusiastic Frisian literary movement, but its works are not widely read. Furthermore, though Frisian continues to be widely used as the language of everyday oral communication, it is increasingly a "Netherlandic" Frisian, with numerous borrowings from standard Netherlandic.

The status of Frisian in the East and North Frisian areas of Germany is far more tenuous. There German performs all the functions of a standard language, and Frisian serves only as yet another local dialect, comparable to the many surrounding local dialects of Low German. No standard North Frisian or East Frisian exists.

Characteristics. The following remarks refer to the more or less standard West Frisian that is developing in the province of Friesland.

Frisian has the following system of consonants, given here in the usual spellings: stops, *p, b, t, d, k, g;* fricatives, *f, v, s, z, ch, g;* nasals, *m, n, ng;* liquids, *l, r;* and glides, *w, h, j.* Examples (given here in part to show the close relationship between Frisian and English) include *p, t,* and *k* (unaspirated) in *peal* "pole," *twa* "two," and *kat* "cat"; *b, d,* and the stop symbolized by the letter *g* in *boi* "boy," *dei* "day," and *goed* "good"; *f, s,* and *ch* in *fiif* "five," *seis* "six," and *acht* "eight"; *v, z,* and the fricative symbolized by the letter *g* in *tolve* "twelve," *tûzen* "thousand," and *wegen* "ways"; *m, n,* and *ng* in *miel* "meal," *need* "need," and *ring* "ring"; *l* and *r* in *laem* "lamb" and *reep* "rope"; *w, h,* and *j* in *wy* "we," *hy* "he," and *jo* "you." Word-finally, voiced *b, d, z,* and *g* are generally unvoiced to *p, t, s,* and *ch.*

Frisian has the following system of stressed vowels and diphthongs. The symbols given in Table 18 refer to the actual sounds rather than to Frisian spellings, which are often irregular. Frisian also has an unstressed vowel *ə* (pronounced as the *a* in English "sofa"), which occurs only in unstressed syllables.

Dialects. The Frisian dialects of The Netherlands

Table 18: The Vowel System of Frisian

short vowels	long vowels	rising diphthongs	falling diphthongs
i ü u	ī ǖ ū		ie üö uo
e ö o	ē ȫ ō	öi	eɛ öə oa
ɛ ɔ	ɛ̄ ɔ̄	ɛi ɔu	
a	ā	ai	

Dialects of Friesland

province of Friesland are, with three exceptions, relatively uniform, though it is customary to make a distinction between Wouden Frisian in the east, Klei Frisian in the west (the variety on which standard Frisian is largely based), and Southwest Corner Frisian in the southwest. The three exceptions are the island dialect of East and West Terschelling and the dialects of the city of Hindeloopen and of the island of Schiermonnikoog. These latter two differ so greatly that they are not intelligible to other speakers of West Frisian and are both dying out. Quite different from any of these is the so-called City Frisian (Stedfrysk, or Stedsk) spoken in the cities of Leeuwarden, Franeker, Harlingen, Bolsward, Sneek, Staveren, and Dokkum. Despite the name, this is not Frisian at all but a variety of Netherlandic strongly influenced by Frisian. Similar in nature are the dialects of Heerenveen and Kollum, of the middle section of the island of Terschelling, and of Het Bildt (a coastal area northwest of Leeuwarden, diked in and settled by Hollanders during the 16th century).

East Frisian survives today only in the German Saterland, consisting of the three parishes of Ramsloh, Strücklingen, and Scharrel, each with a slightly different dialect. The area to the north is called East Frisia (German Ostfriesland), and the local dialect East Frisian (German Ostfriesisch), although it is really not Frisian but the local variety of Low German.

Though North Frisian is spoken in only a small geographical area by only some 18,000 persons, it exists in an extraordinary number of local dialects, some of which are mutually unintelligible. Because of this, it would be almost impossible to develop a single standard North Frisian that could be used throughout this area. North Frisian dialects are customarily divided into Insular North Frisian (Sylt, Föhr-Amrum, Helgoland) and Continental North Frisian (the Halligen Islands and the coast of Schleswig), the latter in seven main varieties and further subvarieties. Because this whole area bordered until recently on Danish, it was extensively influenced by the neighbouring Danish dialects. In more recent times it has been heavily influenced by German, both standard German and the neighbouring Low German dialects. Today all speakers of North Frisian are probably bilingual or trilingual; all of them learn Frisian at home and standard German in school, and many also learn dialectal Low German.

Identity of "Dutch" and "Flemish"

Netherlandic (Dutch–Flemish). Netherlandic is the national language of The Netherlands and one of the two national languages (beside French) of Belgium. Popular English usage applies the term Dutch to the Netherlandic of Holland and the term Flemish to the Netherlandic of Belgium, but in fact they are one and the same standard language. In its various forms, standard and dialectal, Netherlandic is the indigenous language of most of The Netherlands (all but the Frisian-speaking province of Friesland), of northern Belgium, and of a small part of France immediately to the west of Belgium. It is also used as the language of administration in the Dutch dependency of the Netherlands Antilles and in a former dependency, the Republic of Suriname. A derivative of Netherlandic, Afrikaans, is one of the two national languages (with English) of the Republic of South Africa.

As a written language, Netherlandic is quite uniform; it differs in Holland and Belgium no more than written English does in the United States and Great Britain. As a spoken language, however, it exists in far more varieties than does the English of North America. At one extreme is Standard Netherlandic (Algemeen Beschaafd Nederlands, "General Cultured Netherlandic"), which is used for public and official purposes and is the language of instruction in schools and universities. It is everywhere quite uniform, though speakers usually show by their accents the general area from which they come. At the other extreme are the local dialects, used among family and friends and with others from the same village.

History. Netherlandic is descended primarily from the language of the Rhine-Weser group, especially from the language of the Franks who entered much of this area during the 4th and 5th centuries AD. At the same time, it shows many forms descended from the speech of the North Sea Germanic inhabitants of the coastal areas. For example, modern *vijf* "five" (Proto-Germanic **fimf*) shows the typical North Sea Germanic loss of a nasal sound before *f.* Modern *mond* "mouth" (Proto-Germanic **munþ-*) and *ons* "us" (Proto-Germanic **uns*), on the other hand, show preservation of *n* before *þ* and *s;* but loss of *n* before *þ* appears in such place names as *IJmuiden* "mouth of the river IJ," and loss of *n* before *s* appears in the widespread dialectal forms *us* and *os* "us."

Documents written in Netherlandic do not begin to appear until toward the end of the 12th century, in the rich literature called Middle Dutch or Middle Netherlandic. From the preceding Old Netherlandic period there are only a few glosses, names, and isolated words appearing in Latin documents. Related to Netherlandic, though not ancestral to it, are the copyings—partly running text, partly isolated words—made from a lost manuscript that apparently contained an interlinear translation from Latin into Old Low Franconian of the book of Psalms.

The development of modern Netherlandic is closely tied to the political and economic history of the area. By the middle of the 16th century the speech of Brabant and its leading cities Antwerp and Brussels was well on its way to becoming standard for the whole Netherlandic speech area. Then came the revolt against Spain, in which the northern province of Holland was split off from the southern Netherlandic provinces.

Northern and Southern Netherlandic differences

This political split between the United Provinces of the Netherlands in the north and the Spanish Netherlands in the south had far-reaching linguistic consequences. In the prosperous and vigorous north a standard language rapidly developed, based on the speech of the big cities; it also showed the influence of the culturally important refugees from Brabant, who fled to the north, above all to Amsterdam, before and especially after the fall of Antwerp (1585). In the south, French came to prevail among the upper classes. The less privileged classes continued to use dialectal Netherlandic ("Flemish"), but no supradialectal standard was developed.

The cultural predominance of French in the south increased during the period of French rule (1795–1814), abated somewhat during the years when Belgium and Holland were united independently (1815–30), and rose again after the founding of the Kingdom of Belgium in 1830. At that time French was the only official language, used exclusively in government, courts, and schools. The long struggle to give Netherlandic equal status with French ended with the Language Act of 1938, which made Netherlandic the only offical language of northern Belgium. There were numerous attempts to set up a standard Flemish different from the Netherlandic of the north, but in the end the standard Netherlandic that had become established in Holland was accepted for northern Belgium as well.

Characteristics. Modern Standard Netherlandic has the following consonants, given here in the usual spellings: stops, *p, b, t, d, k;* fricatives, *f, v, s, z, ch, g;* nasals, *m, n, ng;* liquids, *l, r;* glides, *w, h, j.*

The voiced stops and fricatives *b, d, v, z,* and *g* are unvoiced to *p, t, f, s,* and *ch,* respectively, in word-final position. The spelling shows this in the case of *v* and *z* (plural *dieven* "thieves," *huizen* "houses," but singular *dief* "thief," *huis* "house") but does not show it in the case of *b, d,* and *g* (plural *ribben* "ribs," *bedden* "beds," *dagen* "days," but singular *rib* "rib," *bed* "bed," *dag* "day," pronounced *rip, bet, dach*).

Netherlandic has three classes of vowels and diphthongs: (1) six checked vowels, which are short and always followed by a consonant; (2) ten free vowels and diphthongs, most of them usually long, which need not be followed by a consonant; and (3) a vowel that occurs only in unstressed syllables. They form the system shown in Table 19 (the

Table 19: Vowel System of Netherlandic

traditional spelling						linguistic notation					
checked			free			checked			free		
			ie	uu	oe				ī̆	ǖ̆	ū̆
i	u	o	ee	eu	oo	e	ö	o	ē	ȫ	ō
e		o	ij, ei	ui	ou, au	ɛ		ɔ	ɛi	ɔ̈ü	ɔu
	a			aa			a			ā	
unstressed: e						unstressed: ə					

traditional spelling is to the left, and to the right is a notation, used by some linguists, that indicates the distinctive sounds [phonemes] of the language). Unlike the English spelling system, which in its basic design has remained essentially unchanged since the days of Chaucer (died 1400), the Netherlandic spelling system has undergone a series of official reforms to keep it in line with changes in pronunciation. The principal inconsistencies in the spelling of vowels are the spellings *ij* and *ei,* which both symbolize the same diphthong, pronounced somewhat between the *ai* of English "aisle" and the *ai* of English "maid" (*bijt* "he bites" rhymes with *feit* "fact"), and the spellings *ou* and *au,* which both symbolize the same diphthong, pronounced somewhat between the *ow* of English "now" and the *ow* of English "low" (*bouw* "building" rhymes with *nauw* "narrow"). Free vowels are written with double letters in closed syllables (*vuur* "fire," *boot* "boat"), but with single letters in open syllables (*vuren* "fires," *boten* "boats"). In contrast the checked vowels are always written with single letters.

Netherlandic spelling

Dialects. Although the standard language changes abruptly at the political border separating Holland and Belgium from Germany (Netherlandic being used to the west, German to the east), in local dialect speech there is no such abrupt change. The entire Netherlandic–German territory from the North Sea to the Alps is a single dialect area with only gradual transitions from one village to the next.

In an area bounded roughly by Amsterdam, The Hague, and Rotterdam, rural dialects are very similar to Standard Netherlandic; there are marked differences only in urban dialects, especially those of Amsterdam and Rotterdam. As one travels from this area—the source of the standard language—in any direction, however, the difference between local dialects and the standard language becomes progressively greater; as a result, throughout most of Holland rural inhabitants in effect speak two closely related but distinct languages, Standard Netherlandic and a local dialect, in varying degrees of proficiency. Dialects are traditionally named after the provinces in which they are spoken (*e.g.,* Gronings in Groningen and Limburgs in Limburg).

In Netherlandic Belgium the use of Standard Netherlandic is more limited, and that of local dialects is more extensive. Some of the better educated people speak the standard language fluently and use it regularly, while others prefer French. The less well educated use a local dialect almost exclusively and are often able to handle the standard language only with difficulty.

Afrikaans. In 1652 a party of Netherlanders under the leadership of Jan van Riebeeck landed at the Cape of Good Hope to establish a station for the Dutch East India Company. In the immediately following years they were joined by a wide variety of other Europeans, in particular Germans and, after 1685, French Huguenots. By 1806, a century and a half after the original settlement, the national origins of the white inhabitants are estimated to have been 53 percent Dutch, 28 percent German, 15 percent French, and 4 percent of other nationalities. Shortly before this date, in 1795, the Cape Colony came under British control, and British settlers began to arrive around 1820.

Afrikaans–Netherlandic differences

History. From the start, the dialect of the province of Zuid-Holland (South Holland), which was spoken by Van Riebeeck and his large family, seems to have set the style for what was eventually to become modern Afrikaans. As might be expected in a language used by so many non-native speakers (white and black), some simplification of sounds and forms took place. For example, whereas Standard Netherlandic shows three forms in the present tense of most verbs—*ik loop, hij loopt, wij/zij lopen* "I run, he runs, we/they run"—Afrikaans shows only one form—*ek/hy/ons/hulle loop* "I/he/we/they run" (with no ending). In addition, whereas Standard Netherlandic uses stressed *die man* "that man" in contrast to unstressed *de man* "the man," Afrikaans has only *die man* "the man"; and, whereas Standard Netherlandic distinguishes *wij / we* "we" and *ons* "us," Afrikaans has only *ons* "we/us."

The relatively small numbers of white European speakers of Afrikaans borrowed place-names and names of such cultural novelties as African plants and the like from the immensely larger numbers of black speakers of various Bantu languages, with whom they were in intimate contact. For some two centuries the gradually developing Afrikaans language existed only as a spoken dialect, alongside Standard Netherlandic (by which it was constantly influenced) and, later, English. Then, around the middle of the 19th century, the effort to make Afrikaans a medium of literary expression and a standard written language began. It came gradually to be used in newspapers. It was adopted for use in schools in 1914 and was accepted for use in the Dutch Reformed Church in 1919. In 1925 the South African Parliament declared it to be an official language, replacing Netherlandic. The first complete translation of the Bible into Afrikaans was published in 1933. Thus it came to be recognized as one of the two standard languages (beside English) of the modern Republic of South Africa. Though clearly a separate language, Afrikaans is very similar to Netherlandic. A person who knows Netherlandic can read Afrikaans with little difficulty; and, with some practice, he can easily learn to understand it when spoken.

Characteristics. Afrikaans has the following consonants, given here in the conventional spellings: stops, *p, b, t, d, k, gh/g;* fricatives, *f/v, w, s, z, g;* nasals, *m, n, ng;* liquids, *l, r;* glides, *h, j.* There are numerous differences between Afrikaans and Netherlandic. Netherlandic *-g-* (*-gg-*) is a voiced fricative, but Afrikaans *-g-* (*-gg-*) is a voiced stop. Unlike Netherlandic, Afrikaans also has this voiced stop initially in a few loanwords. Netherlandic has voiced fricatives initially (*v-, z-, g-*); corresponding words have voiceless initial fricatives in Afrikaans. Afrikaans, however, has voiced *z-* in loanwords: *Zoeloe* "Zulu." Netherlandic has initial *s* plus fricative *ch* as in *schoen* "shoe"; corresponding words have *s* plus *k* in Afrikaans: *skoen.* Netherlandic has *-ft, -st,* and *-cht* as in *gift* "poison," *nest* "nest," *nacht* "night"; corresponding words show loss of *-t* in Afrikaans: *gif, nes,* and *nag.*

Afrikaans has the system of vowels shown in Table 20 (usual spelling to the left; notation used by linguists to indicate distinctive sounds to the right). As in Netherlandic, *uu, ee, oo,* and *aa* are written with single letters in open syllables, and single consonant letters are doubled in open syllables to show that the preceding vowel is short.

Table 20: Vowel System of Afrikaans

usual spelling									linguistic notation								
short vowels			long vowels			diphthongs			short vowels			long vowels			diphthongs		
ie	uu	oe	ie	uu	oe				i	ü	u	ī	ǖ	ū			
			ee	eu	oo	(ee	eu	oo)*				ē	ȫ	ō	(iə	üə	uə)*
	i			î						ə			ə̄				
e	u	o	ê/e	û	ô/o	y/ei	ui	ou	ɛ	ɔ̈	ɔ	ɛ̄	ɔ̈̄	ɔ̄	əi	əü	əu
	a			aa			ai			a			ā			ai	

*The spellings *ee*, *eu*, and *oo* are pronounced either as long vowels (ē, ȫ, ō) or as diphthongs (iə, üə, uə).

Dispersion of German

German. German is spoken throughout a large area in central Europe, where it is the national language of Germany and of Austria and one of the four national languages (beside French, Italian, and Romansh) of Switzerland. From this homeland it has been carried by emigration to many other parts of the world; there are German-speaking communities in North and South America, South Africa, and Australia. In the western world it is extensively used as a second language and in this respect is next in importance (along with French) only to English.

As a written language German is quite uniform, differing in Germany, Austria, and Switzerland no more than written English does in the United States and the British Commonwealth. As a spoken language, however, German exists in far more varieties than English. At one exteme is Standard German (Hochsprache), based on the written form of the language and used in radio, television, public lectures, the theatre, schools, and universities. It is relatively uniform, although speakers often show by their accents the areas from which they come. At the other extreme are the local dialects, which differ from village to village. Between these two extremes there is a continuous scale of speech forms that, in cities, are often close to the standard language and are called Colloquial German (Umgangssprache).

History. From the point of view of local dialects the territory within which German and Netherlandic are spoken is a single speech area. It is possible to travel from Austria, northern Italy, and much of Switzerland into Germany, eastern France (Alsace and part of Lorraine), Luxembourg, northern Belgium, and The Netherlands without encountering a village where the local speech is suddenly different. The only sharp breaks occur when one enters the French-speaking parts of France and Belgium or the Frisian-speaking parts of The Netherlands and Germany.

High German consonant shift

The most striking dialect differences within this large area are those that divide Netherlandic-Low German in the lowlands of the north from High German in the highlands of the south. When the Germanic tribes migrated into southern Germany during the early centuries of the Christian era, their speech had the voiceless stops *p, t,* and *k* in much the same distribution as in modern English. Then, probably during the 6th century, there occurred a change customarily called the "High German consonant shift." At the beginning of words and when doubled, *p, t,* and *k* came to be pronounced as affricates; after a vowel they came to be pronounced as long fricatives. The modern results, compared with related English words, are shown in Table 21.

Table 21: Results of the High German Consonant Shift

p-	*p*ound	*Pf*und	*pp*	a*pp*le	A*pf*el	*Vp*†	ho*p*e	ho*ff*en
t-	*t*en	*z*ehn	*tt*	si*tt*ing	si*tz*en	*Vt*†	bi*t*e	bei*ss*en
k-	can	*kh*ann*	*kk*	li*ck*	le*kch*en*	*Vk*†	ma*k*e	ma*ch*en

*Khann and lekchen, with affricates, are southern dialect forms; standard German has stops: *kann, lecken.* †*V* represents any vowel.

These changes occurred in the south of the German speech area and then spread north, some extending farther than others. The situation at the end of the 19th century was as indicated in Figure 12. Line 2, *maken/machen,* is generally chosen as the boundary between Low German and High German, because it is typical for the shift of *p, t,* and *k* after vowels to *ff, ss,* and *ch,* respectively (*hopen/hoffen, bīten/beissen, maken/machen*), and of *t* and *tt* to *z* and *tz,* respectively (*ten/zehn, sitten/sitzen*). The shift of *ik* "I" to *ich* is indicated by line 1, which shows that the shift of *k* to *ch* after a vowel in this particular word spread unusually far. Line 3, which indicates the shift of *Dorp* "village" to *Dorf* (*cf.* archaic English "thorp"), shows that shifted *p* after *r* and *l* spread less far north than did shifted *p, t, k* after a vowel. And line 4, indicating the shift of *dat* "that" to *das,* shows that the shift of *t* to *s* after a vowel spread still less far north in this word (and in a few others: *it/es* "it," *wat/was* "what"). The striking way in which these lines "fan out" in the west (in the area along the Rhine River) has led to their being called the "Rhenish fan."

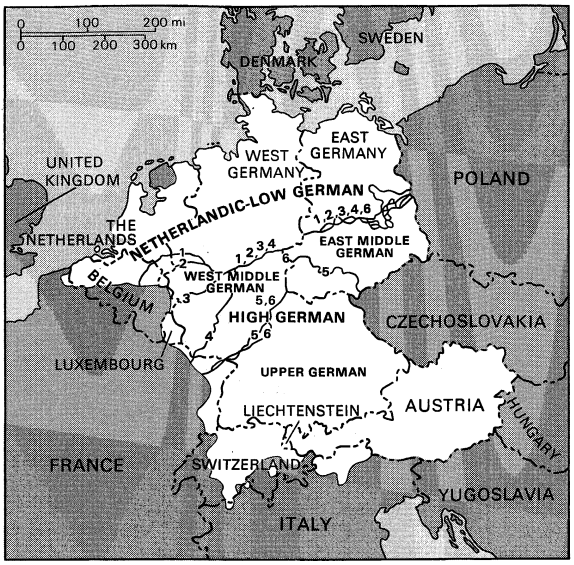


Figure 12: The Netherlandic–German dialect divisions. Numbers refer to isoglosses described in text.

The shift of *p* when doubled or at the beginning of a word occurred in a much smaller area. Line 5, showing the shift of *Appel* "apple" to *Apfel,* lies wholly within the High German speech area and is customarily used to subdivide it into Middle German (*Appel*) and Upper German (*Apfel*) areas. Line 6, which indicates the shift of *Pund* "pound" to *Pfund,* follows much the same course as does line 5 in the west, but it then runs north to join the *maken/machen* line; it is customarily used to distinguish West Middle German (*Appel, Pund*) from East Middle German (*Appel, Fund*—the latter being more common than Upper German *Pfund*).

Beginning of German

As noted above (in the section on the branches of Germanic), during the early centuries of the Christian era there was only one "Germanic" language, with little more than minor dialect differences. Only after the consonant shift just described is there justification in speaking of a "German" (*i.e.,* High German) language distinct from the other Germanic languages. The fact that many early loans from Latin spread throughout all of Germanic makes it clear that the various dialects of early Germanic were mutually intelligible and that there was easy communication among them. At the same time, the modern German forms of these early loans show that they must have been borrowed before the consonant shift, because they show its effects. Examples include Latin *pondō,* English "pound," but German *Pfund;* Latin *piper,* English "pepper," but German *Pfeffer;* Latin *tegula,* English "tile," but German *Ziegel;* Latin (*via*) *strāta* "paved (road)," English "street," but German *Strasse;* Latin *catillus,* English "kettle," but German *Kessel;* and Latin *coquus,* English "cook," but German *Koch.*

Toward the end of the 4th century there began the great migrations (German *Völkerwanderung*) of Germanic tribes, resulting in an expansion of the Germanic-speaking territory. Angles, Saxons, and Jutes crossed the Channel to England; Franks moved southwest into northern France and south into southern Germany; and Alamanni, Bavarians, and Langobardi moved south into southern Germany, Switzerland, Austria, and northern Italy. At the same time, the area east of the Elbe and Saale rivers was largely vacated by Germanic speakers, and Slavic speakers moved in.

Old period (*c.* 750–1050)

In the southern area settled by Franks, Alamanni, and Bavarians, the first Old High German written records began to appear during the 2nd half of the 8th century. Their language is best described as a collection of monastery dialects; there is a certain uniformity in the writings of any given monastery, but little for the area as a whole. The first documents are translations into German of Latin word lists. Later documents include prose translations of St. Isidore of Seville (made *c.* 800) and of Tatian (*c.* 830), as well as a new verse form with end rhyme (Otfrid, *c.* 870). This literature reached its highest point in the able translations and interpretations of the Swiss monastery teacher Notker Labeo (died 1022). From the north (the Old Low German or Old Saxon speech area), the most extensive documents preserved are a life of Christ in alliterative verse (*Heliand, c.* 830) and a fragment of a similar Genesis translation.

In this period there were many borrowings from Latin, nearly all connected with Christianization of the Germans. Because they were made after the consonant shift, they do not show its effects. Examples of these borrowings include *predigōn* (modern German *predigen* "to preach"), from Latin *praedicāre; tempal* (modern German *Tempel* "temple"), from Latin *templum;* and *spiagal* (modern German *Spiegel* "mirror"), from Latin *speculum.* On the other hand, borrowings of this period reflect sound changes that had occurred in popular Latin, such as the change of Latin *c* before *e* from a *k* sound to *ts* in *cella* "cell" and *crucem* "cross," Old High German *zella, krūzi,* modern German *Zelle, Kreuz* (the letter *z* in the German and Old High German examples represents the sound of *ts*); or the change of Latin medial *-b-* to *-v-* in *tabula* "table," borrowed into Old High German as *tavala,* modern German *Tafel.*

Middle period (*c.* 1050–1350)

Several developments justify the usual assumption of a new period, the language of which is called Middle High German, beginning around 1050. First, there were changes in the language itself, among which were the unvoicing of final *b, d,* and *g* (*cf.* Old High German *grab* "grave," *rad* "wheel," and *tag* "day" with Middle High German *grap, rat,* and *tac;* in modern German these words are again spelled *Grab, Rad,* and *Tag* but are pronounced with final *p, t,* and *k*) and the reduction of the vowels of unstressed syllables to a *ə* sound, usually spelled *e* (*e.g.,* in the plural of the word for "day," the Old High German nominative-accusative form was *taga,* the genitive was *tago,* and the dative was *tagun,* but for these Middle High German had *tage, tage,* and *tagen,* respectively, and modern German has *Tage, Tage,* and *Tagen*). Second, there were great changes in the geographical in which German was spoken. In the west the Franks of northern France had become romanized, and the French–German language border had assumed approximately its present location; in the east, on the other hand, German began to spread into Slavic territory, a process that was to continue for many centuries and to be reversed only at the end of World War II. Third, writing became independent of the monasteries, and the number of written documents soon increased greatly in both north and south. In the south, especially, a remarkable literature developed that included courtly epic and *Minnesang.* There is clear evidence of a trend toward a standard Middle High German literary language, though it seems to have had no influence on ordinary speech. Because this literature was based largely on French models, many French words were borrowed into German.

Early modern period (*c.* 1350–1650)

Four events—the growth of trade, the rise of a middle class, the invention of printing, and the Reformation—had great influence on the development of the language. In the north, because of the prosperity of the Hanseatic League, a standard Low German written language began to develop, though it never reached full growth and probably had little influence on everyday speech. In the south the dialects that had arisen in the recently settled East Middle German area were relatively uniform and contained elements from both West Middle German and Upper German. Gradually these East Middle German dialects came to be used as the offical languages of the chancelleries of the area, including that of Saxony; and on this latter Martin Luther based the language of his widely read Bible translation (1522–34). The growth of this type of German, which developed gradually into modern Standard German, was aided by the fact that printers preferred it as a means of making their books appeal to the widest possible audience.

Three striking vowel changes are characteristic of this period. In the southeast, as early as the 12th century, the long vowels *ī, ū,* and *ǖ* came to be diphthongized to *ei, ou,* and *öü;* this is called the "New High German diphthongization." By the 15th century these new diphthongs had spread to East Middle German, and in the standard language they merged with the old diphthongs *ei, ou,* and *öü.* Examples include Middle High German *mîn* "my," *hûs* "house," and *hiuser* "houses" with the monophthongs *ī, ū,* and *ǖ,* in contrast to *ein* "a," *troum* "dream," and *tröume* "dreams" with the diphthongs *ei, ou,* and *öü,* but modern Standard German *mein, Haus,* and *Häuser* appear with the same diphthongs (*ai, au,* and *oi*) as *ein, Traum,* and *Träume.* By a specifically Middle German development, the diphthongs *iə, uə,* and *üə,* still preserved in the southern dialects, were monophthongized to long *ī, ū,* and *ǖ*; this is the "New High German monophthongization." Examples include Middle High German *tief* "deep," *vuoz* "foot," and *vüeze* "feet" with the diphthongs *iə, uə,* and *üə,* contrasted to modern Standard German *tief, Fuss,* and *Füsse* with the monophthongs *ī, ū,* and *ǖ.* Short vowels remained short in closed syllables before long consonants but were lengthened in open syllables before a short consonant plus an unstressed vowel. This is called "lengthening in open syllables."

Modern period (since 1650)

The outstanding developments of the modern period have been the increasing standardization of High German and its increasing acceptance as the supradialectal form of the language. In writing, it is almost the only form used (except for small amounts of dialect literature); in speech, it is the first or second language of nearly the entire population.

Although Standard German is clearly based on the East Middle German dialects, it is not identical with any one of them; it has accepted and standardized many forms from other areas, notably the Upper German sound *pf* (*Pfund, Apfel*) and also large numbers of individual words in the form of other dialect areas. Since it is the only type of German taught in schools, its spoken form is based to a large extent on its written form; and the spoken form that carries the greatest prestige (that of stage, screen, radio, and so on) uses a largely Low German pronunciation of this written form. As a result, the spoken form of modern Standard German has often been aptly described as "High German with Low German sounds."

Characteristics of modern Standard German. German has the following consonants, given here in phonetic symbols because the spelling often varies: stops, *p, b, t, d, k, g;* fricatives, *f, v, ç~x;* sibilants, *s, z, š, ž;* nasals, *m, n, ŋ;* liquids, *l, r;* glides, *h, j.* German *ç~x,* spelled *ch,* is the voiceless velar fricative *x* after *a, ā, o, ō, u, ū,* and *au* but is the voiceless palatal fricative *ç* in other phonetic environments. The German sound *ž* occurs only in loanwords.

In the orthography, German *w* always indicates a *v* sound; German *v* spells an *f* sound in native words but a *v* sound in loanwords. German *sp* and *st* spell the sounds *sp* and *st* in most positions, but they spell *šp* (*shp*) and *št* (*sht*) at the beginnings of words or word stems. In other positions the *š* (*sh*) sound is spelled *sch—e.g., Schiff* "ship." Medial *ss* marks a preceding vowel as short, medial β marks it as long; medial *ss,* however, changes to β at the end of a word and before a consonant. German *z* always indicates the sound *ts.* The spelling *tz* marks a preceding vowel as short, and the spelling *z* marks it as long.

Voiced *b, d, g, v,* and *z* do not occur at the ends of words, at the ends of parts of compound words, before suffixes beginning with a consonant, or before endings in *s* or *t.* In these positions they are replaced in pronunciation (though not in spelling) by the corresponding voiceless consonants, namely *p, t, k, f,* and *s.* For example, the *g* in *Tage* "days" is pronounced as English *g,* and the *g* in *Tag* "day" is pronounced as English *k.*

The German vowel system is given in Table 22 in phonetic symbols.

Though the spelling does not always indicate the difference between short and long vowels, the following devices are used more or less consistently: (1) A vowel is al-

Table 22: Vowel System of German

short and lax vowels	long and tense vowels	diphthongs	unstressed vowel
i ü u	ī ǖ ū		
e ö o	ē ȫ ō	oi	ə
a	ā	ai au	

ways short if followed by a double consonant letter—*e.g., still* "still," *wenn* "if," *Rasse* "race," *offen* "open," *Hütte* "hut"—in contrast to the long vowels of *Stil* "style," *wen* "whom," *Straße* "street," *Ofen* "oven," *Hüte* "hats." (2) A vowel is always long if followed by an (unpronounced) *h—e.g., ihnen* "to them," *stehlen* "to steal," *Kahn* "barge," *wohnen* "to dwell," *Ruhm* "fame"—in contrast to the short vowels of *innen* "inside," *stellen* "to place," *kann* "can," *Wonne* "bliss," *dumm* "dumb." (3) A vowel is always long if written double—*e.g., Beet* "(flower)bed," *Staat* "state," *Boot* "boat"—in contrast to the short vowels of *Bett* "bed (for sleeping)," *Stadt* "city," *Gott* "god"; *ie* counts as the doubled spelling of *i—e.g.,* long *ī* in *Miete* "rent" but short *i* in *Mitte* "middle." (4) A vowel (except unstressed *e*) is always long when it stands at the end of a word.

The "plain" vowels—*a, o, u, ā, ō, ū, au*—often alternate with the "umlaut" vowels—*e, ö, ü, ē, ȫ, ǖ, oi,* respectively—as in the following examples with plain vowels in the singular but umlauted vowels in the plural: *Gast* "guest," *Gäste; Gott* "god," *Götter; Mutter* "mother," *Mütter.* As these examples show, the vowel sounds *e, ē,* and *oi* are spelled *ä, ä,* and *äu* when they are the umlaut of *a, ā,* and *au* sounds. *Gast–Gäste, Vater–Väter, Braut–Bräute.* Otherwise they are generally spelled *e, eh,* or *ee* (*beten* "to pray," *geht* "goes," *Beet* ["flower]bed"), and *eu* (*Leute* "people").

The sound *ai* is generally spelled *ei: Seite* "side," *nein* "no," though in a few words *ai: Saite* "string (of an instrument)," *Kaiser* "kaiser." The unstressed schwa sound (*ə*), as the *a* in English "sofa," is spelled *e: beginnen/bəgínən/* "to begin," *geredet/gərḗdət/*"spoken." (Wi.G.M.)

Yiddish. Although there were about 11,000,000 speakers of Yiddish before World War II, approximately half of them were killed in the Nazi holocaust. There are perhaps 4,000,000 Yiddish speakers today, including native speakers but excluding those who use it as a second language. Most speakers live in the United States, Latin America, Israel, and the Soviet Union. They are served by an active press, including 11 daily newspapers.

International nature of Yiddish

History. Yiddish, although Germanic, is not a typical Germanic language; it includes not only Germanic features but also elements from Romance, Hebrew-Aramaic, and Slavic languages. A cursory examination of the German component of Yiddish indicates that no Yiddish dialect stands in a one-to-one relationship to any German dialect. The language had its beginnings in the 10th century when Jews from northern France and northern Italy settled in the Rhineland. These early Jewish settlements were dislocated by the Crusades and later by the persecutions that followed in the wake of the Black Plague. The subsequent move to Slavic territory had enormous influence on the development of the language.

Onomastic evidence (evidence from recorded proper names) for Yiddish is known from 1096, and glosses in biblical commentaries may be several decades older. The earliest known connected text is a rhymed couplet inscribed in a Hebrew holiday prayer book from Worms and bearing the date 1272–73. The earliest extensive manuscript, known as the Cambridge Yiddish Codex, is explicitly dated November 9, 1382. It excites the interest of Germanicists for its version of "Ducus Horant" (a poem from the Hildesage of the Gudrun epic known from the Ambross Manuscript written by Hans Reid, 1502/4–1515), which antedates the earliest extant manuscript of the Hildesage by at least 130 years. The documentary history of Yiddish is unbroken thereafter to the present day. Unique evidence for spoken Yiddish is incorporated in an extensive body of rabbinical Responsa (published rabbinical opinions on matters of religious law) beginning in the 15th century. Testimony before the rabbinical court, recorded verbatim, provides unusual insight into the colloquial language.

Scholars divide the history of Yiddish into four periods: Earliest Yiddish, to 1250; Old Yiddish, 1250–1500; Middle Yiddish, 1500–1750; and Modern Yiddish, 1750 to the present. The earliest literary tradition had a Western Yiddish dialectal base; writing in this literary dialect continued into the Modern Yiddish period long after the major population centres had shifted to the East. The establishment of the modern literary language on an Eastern Yiddish base occurred only in the early 19th century. At the same time a new style in the language of Yiddish Bible translation emerged, free from the constraints of the original Hebrew syntax and of the stricture against the use of Yiddish words of Hebrew-Aramaic origin in translating from Hebrew. The continuous contact of Yiddish speakers with Hebrew-Aramaic texts and, in the European language area, with one or another Germanic or Slavic language have been important factors in the development of the language.

Characteristics. Because of the conditions under which Yiddish developed (*i.e.,* the numerous contacts it has had with other languages), it is of great interest to scholars.

Yiddish uses all the letters of the Hebrew alphabet, including traditional word-final variants, which have only recently been reintroduced into the orthography of Soviet Yiddish. Several letters occur only in words of Hebrew-Aramaic origin, which retain their traditional spelling except in the Yiddish of the Soviet Union.

The vowel system of Standard Yiddish consists of the simple vowels *i, e, a, o,* and *u* and the diphthongs *ej, aj, oj.* Under Slavic influence a palatal series of consonants has emerged. Unlike German, *x* corresponding to German *ch* has no palatal variant, the ŋ sound (the *ng* in English "sing") is simply a positional variant of *n,* there is no glottal stop (a sound made by closure of the vocal cords), and word final voicing is distinctive (phonemic—*i.e.,* it carries a change in meaning). Words of Hebrew-Aramaic and Slavic origin have introduced a rich variety of consonant clusters that do not appear in German. Intonation contours, apparently related to the chant with which the Talmud is studied, convey syntactic-semantic distinctions independently.

Yiddish grammar

Case inflections, preserved only in the singular, appear in noun modifiers but only rarely in nouns themselves. The dative and accusative cases have merged in the masculine; the nominative and accusative cases have merged in the feminine and neuter. All prepositions govern the dative case. The system of noun plural formation, basically of German origin, is enriched by word elements of Hebrew origin. Many nouns differ from their German cognates both in gender and plural form. A well-developed system of diminution uses word elements largely of German origin but on a Slavic grammatical model. A semantically significant distinction between inflected and uninflected predicate adjectives has emerged, while the difference between weak and strong adjectives, a characteristic of other Germanic languages, has effectively disappeared. The verb is inflected only in the present indicative. Other tenses and moods are expressed by means of auxiliary words. In normal word order the inflected verb immediately follows the subject; any remaining part of the verb phrase occurs as close to the inflected verb as possible. The special word order of the German subordinate clause is unknown, and verb initial constructions generally express consecutiveness rather than interrogation.

In the vocabulary, words and word elements borrowed from a number of different languages co-occur and often combine freely in a manner unfamiliar to the languages from which they derive. Furthermore, when words borrowed from different languages are partially alike, one of them may be analyzed and inflected in terms historically appropriate to the other, thereby yielding blends of complex etymology. In addition, a highly productive system of prefixing yields verbs that are German in form but derive their meanings from an underlying Slavic model.

Dialects. The basic dialectal division is between Western Yiddish, which occurs largely within the German language area, and Eastern Yiddish in the Slavic-speaking areas. Eastern Yiddish is traditionally subdivided into Northeastern Yiddish and Southern Yiddish, the latter consisting of Central Yiddish and Southeastern Yiddish. The phonological criteria on which this division is based are typically reflected in the variants of the phrase "to buy meat": Western Yiddish *kāfn flāš,* Central Yiddish *kojfn flajš,* Southeastern Yiddish *kojfn flejš,* Northeastern Yiddish *kejfn flejš.* Other phonological and many lexical differences reinforce the distinctness of Western Yiddish.

In the East, Central Yiddish is further distinguished by a full set of contrasts in vowel length, while the varieties of Southeastern Yiddish have made changes in vowel quality that have led to the types *hont* "hand," *huz* "house," and *rign* "rain." Northeastern Yiddish is characterized by the loss of the neuter gender. Standard Yiddish adheres more closely to Northeastern Yiddish in its sound system, and more closely to Southern Yiddish in its grammatical patterns. (M.I.H.)

NORTH GERMANIC (THE SCANDINAVIAN LANGUAGES)

Runes

History. About 125 inscriptions dated from AD 200 to 600, carved in the older runic alphabet (futhark), are chronologically and linguistically the oldest evidence of any Germanic language. Most of them are from Scandinavia, but enough have been found in southeastern Europe to suggest that the use of runes was also familiar to other Germanic tribes. Most inscriptions are brief, marking ownership or manufacture, as on the Gallehus Horns (Denmark; *c.* AD 400): *Ek Hlewagastiz Holtijaz horna tawido* "I, Hlewagastiz, son of Holti, made [this] horn." A number of inscriptions are memorials to the dead, while others are magical in content. The earliest were carved on loose wooden or metal objects, while later ones were also chiselled in stone.

The inscriptions retain the unstressed vowels that were descended from Germanic and Indo-European but were lost in the later Germanic languages; *e.g.*, the *i*'s in *Hlewagastiz* and *tawido* (Old Norse would have been **Hlégestr* and **táða*) or the *a*'s in *Hlewagastiz, Holtijaz,* and *horna* (Old Norse **Høltir, horn*). The scantiness of the material (fewer than 300 words) makes it impossible to be sure of the relationship of this language to Germanic and its daughter languages. It is traditionally known as Proto-Scandinavian but shows few if any distinctively North Germanic features and may reflect a stage, sometimes called Northwest Germanic, prior to the splitting of North and West Germanic (but after the separation of Gothic). Only after the departure of the Angles and Jutes for England and the establishment of the Eider River in south Jutland as a border between Scandinavians and Germans is it reasonable to speak of a clearly Scandinavian or North Germanic dialect.

Viking expansion

Common Scandinavian: 600–1050. Inscriptions from the 7th century show North Germanic as a distinct, fairly uniform, and recognizable dialect. Information about the language is derived from runic inscriptions, which became more abundant after the creation of the short runic futhark about AD 800, from names and loanwords in foreign texts and from reconstructions based on placenames and later dialects. The expansion of Nordic peoples in the Viking Age (*c.* 750–1050) led to the establishment of Scandinavian speech in Iceland, Greenland, the Faeroes, Shetlands, Orkneys, Hebrides, and the Isle of Man, as well as parts of Ireland, Scotland, England, France (Normandy), and Russia. Scandinavian languages later disappeared in all these territories except the Faeroes and Iceland through absorption or extinction of the Scandinavian-speaking population.

During the period of expansion, all Scandinavians could communicate without difficulty and thought of their language as one (sometimes called "Danish" in opposition to "German"), but the differing orientations of the various kingdoms in the Viking Age led to a number of dialectal differences. It is possible to distinguish a more conservative West Scandinavian area (Norway and her colonies, especially Iceland) from a more innovative East Scandinavian (Denmark and Sweden), the former oriented to the Atlantic, the latter to the Baltic. There were no firm borders, however, and the sea lanes formed active channels of contact. An example of a linguistic difference setting off the eastern dialect area is the monophthongization of the Common Scandinavian diphthongs *ei, au,* and *øy* to *ē* and *ø* (*e.g., steinn* "stone" became *stēn, lauss* "loose" became *løs,* and *høyra* "hear" became *hø̄ra*). The diphthongs remained on the island of Gotland and in most North Swedish dialects, however, while they were lost in some East Norwegian dialects. The pronoun *ek* "I" became *jak* in East Scandinavian (modern Danish *jeg,* Swedish *jag*) but remained *ek* in West Scandinavian (New Norwegian and Faeroese *eg,* Icelandic *ég*); in East Norwegian it later became *jak* (dialects *je, jæ,* Dano-Norwegian *jeg*) but remained *ek* (dialects *a, æ*) in Jutland.

Old Scandinavian: 1050–1450. The establishment of the Christian church in its Roman Catholic form during the 10th and 11th centuries had considerable linguistic significance. It helped to consolidate the existing kingdoms, brought the North into the sphere of classical and medieval European culture, and introduced the writing on parchment of Latin letters. Runic writing continued in use for epigraphic purposes and messages for the general population (several thousand inscriptions are extant, from 11th-century Sweden, especially, and also all the way from Russia to Greenland). For more sustained literary efforts, the Latin alphabet was used—at first only for Latin writings but soon for native writings as well. The oldest preserved manuscripts date from *c.* 1150 in Norway and Iceland and *c.* 1250 in Denmark and Sweden. The first important works to be written down were the previously oral laws; these were followed by translations of Latin and French works, among them sermons, saints' legends, epics, and romances. Some of these may have stimulated the extraordinary flowering of native literature, especially in Iceland. One can hardly speak of distinct languages in this period, although it is customary to distinguish Old Icelandic, Old Norwegian, Old Swedish, Old Danish, and Old Gutnish (or Guthnic, spoken in Gotland) on the basis of quite minor differences in the writing traditions. Some of these were merely scribal habits resulting from local usage, but others did reflect the growing separation of the kingdoms and the centralization within each. Literary Old Icelandic is often presented in a normalized textbook form under the name of Old Norse.

Roman influence

Culture words like *caupō* "merchant" (giving Old Norse *kaupa* "buy") and *vinum* "wine" (Old Norse *vín*) had been filtering into the North from the Roman Empire for a long time. But the first great wave of such words came from the medieval church and its translations, often with the other Germanic languages as intermediaries because the first missionaries were English and German. Some religious terms were borrowed from other Germanic languages; among these are Old Norse *helviti* "hell" from Old Saxon *helliwiti* or Old English *hellewite,* and Old Norse *sál* "soul" from Old English *sāwol.* East Scandinavian borrowed the Old Saxon word *siala,* from which come later Danish *sjæl* and Swedish *själ.* In the secular field the most profound influence on Scandinavian was that exerted by Middle Low German because of the commercial dominance of the Hanseatic League and the political influence of the North German states on the royal houses of Denmark and Sweden between 1250 and 1450. The major commercial cities of Scandinavia had large Low German-speaking populations, and the wide use of their language resulted in a stock of loanwords and grammatical formatives comparable in extent to that which French left behind in English after the Norman Conquest.

Reformation and Renaissance: 1450–1550. The many local dialects that exist today developed in the late Middle Ages, when the bulk of the population was rural and tied to its local village or parish, with few opportunities to travel. The people of the cities developed new forms of urban speech, coloured by surrounding rural dialects, by foreign contacts, and by the written languages. The chanceries in which documents of government were produced began to be influential in shaping written norms that were no longer local but nationwide. The Reformation came from Germany and brought with it High German influence through Martin Luther's translation of the Bible, which was quickly translated into Swedish (1541), Danish (1550), and Icelandic (1584). That it was not translated into Norwegian was one of the major reasons that no separate Norwegian literary language arose. Until the 19th century there was no distinct written Norwegian but only a Norwegian variety of Danish. With the invention of printing and the growth of literacy, all speakers of Scandinavian dialects gradually learned to read (and eventually write) the new standard languages.

The modern languages. The six standard languages of today, in the order of their emergence as languages of culture and prestige, are Danish, Swedish, Icelandic, Faeroese, New Norwegian (Nynorsk), and Dano-Norwegian (Bokmål).

Danish

The norms of the first printed books in Danish continued the norm of the royal chancery in Copenhagen, which was not based on any particular dialect and probably reflected a state of the language closer to 1350 than 1550. Because of the influence of the written language, many speech forms used even by the aristocracy at that time were eliminated or branded as vulgar. Danish is clearly the Scandinavian language that has undergone the greatest amount of change away from the Common Scandinavian norm. In the 18th century a mildly puristic reform led to the replacement of many French loans by their native equivalents (*e.g., imagination* by *indbildning; cf.* German *Einbildung*), and, in the 18th and 19th centuries, Danish became the vehicle of a classical literature. There are regional differences in the cultivated speech norm, but upper-class Copenhagen speech probably has the highest prestige. A spelling reform in 1958 eliminated the capitalization of nouns and introduced the letter *å* for *aa,* thereby bringing the spelling closer to that of Norwegian and Swedish. Danish is spoken by most of the more than 5,000,000 inhabitants of Denmark and in a few communities south of the German border; it is taught in the schools of the Faeroe Islands, Iceland, and Greenland.

Swedish

Before the Swedish revolt of Gustav Vasa in 1525, Danish influence on the Swedish language had been strong; the new government, however, made vigorous efforts to eliminate this. The written norm was based on one that had developed in the manuscripts of central Sweden, extending from the Vadstena monastery in east Götaland to Stockholm and Uppsala. In relation to the speech of the area, many of its features were conservative (*e.g.,* silent *-t* and *-d* in words like *huset* "the house" and *kastad* "thrown"). The written language was cultivated energetically as a symbol of national strength, and in 1786 the Swedish Academy was established by King Gustav III. The language expanded its area at the expense of Danish and Norwegian by the conquest of southern and western provinces in the 17th century. After Sweden lost Finland in 1809, the role of Swedish was gradually reduced in that country. Since independence (1917), Finland has accepted Swedish as one of its official languages and has taught it in its schools, but only about 6.3 percent of its population uses it. Except for small Lappish and Finnish minorities, the entire population of Sweden (more than 8,000,000) has Swedish as its daily language, and there is a rich and distinguished literature.

Icelandic

Important factors in the survival of Icelandic during the period of Danish rule were its continued use for literary purposes, the geographical remoteness of Iceland, a scattered population, and the great linguistic differences between Danish and Icelandic. In the period when the Scandinavian languages in continental Europe became essentially uninflected, Icelandic preserved Old Scandinavian grammar almost intact. The native Bible became a basis for the further development of Icelandic. Nevertheless, the circumstances of the language were highly restricted until self-government developed in the 19th century, and Icelandic was rediscovered by Scandinavian scholars. A firm orthography along etymological lines was gradually established, and the policy of not adopting foreign words was confirmed, so that Icelandic today offers a strikingly different appearance from the other Scandinavian languages.

Faeroese

Prior to modern times literary activity in the Faeroe Islands was minimal, but the local dialects continued to develop, though Danish was the official language. The Danish language scholar Rasmus Rask, who wrote the first Faeroese grammar (1811), described the language as a dialect of Icelandic, but it is actually an independent language, intermediate between West Norwegian and Icelandic but containing many Danish loanwords. Traditional dance ballads were written down after 1773 before the establishment in 1846 of an independent orthography. This orthography is etymologizing and unphonetic and gives Faeroese a strong Icelandic appearance. The establishment of home rule in 1948 led to the introduction of Faeroese as the primary language taught in the schools. The language is now spoken by about 43,000 people.

New Norwegian

Old Norwegian writing traditions gradually died out in the 15th century, after the union of Norway with Denmark and the removal of the central government to Copenhagen. After independence was achieved in 1814, the linguistic union with Danish persisted, but the ideology of National Romanticism stimulated a search for a national standard language. In 1853 a young, self-taught linguist of rural stock, Ivar Aasen, constructed a language norm from the spoken dialects that would continue the Old Norwegian tradition and, hopefully, might eventually replace Danish. After long research and experimentation, he presented this New Norwegian norm (often called Landsmål, but now officially Nynorsk) in a grammar, a dictionary, and in numerous literary texts. New Norwegian was officially recognized as a second national language in 1885. Today all Norwegians learn to read and write it, but only a fifth of the school population and an even smaller percentage of the writers actually use it as their primary language. It has been cultivated by many excellent authors and has a quality of poetic earthiness that appeals even to nonusers. Its norm has changed considerably since Aasen's time in the direction of spoken East Norwegian or written Dano-Norwegian.

Dano-Norwegian

Most Norwegian literature in the 19th century was written in a superficially Danish norm, but it was given Norwegian pronunciation and had many un-Danish words and constructions. The spoken norm was a compromise Dano-Norwegian that had grown up in the urban bourgeois environment. In the 1840s Knud Knudsen formulated a policy of gradual reform that would bring the written norm closer to that spoken norm and, thereby, create a distinctively Norwegian language without the radical disruption envisaged by the supporters of Aasen's New Norwegian. This solution was supported by most of the new writers in the powerful literary movement of the late 19th century. The official reforms of 1907, 1917, and 1938 broke with the Danish writing tradition and adopted native pronunciation and grammar as its normative base; the resultant language form was called Riksmål, later officially Bokmål. Controversial efforts to bring Dano-Norwegian and New Norwegian together into an amalgamated Pan-Norwegian (Samnorsk) have not yet led to any definite result. In its current form Dano-Norwegian is the predominant language of Norway's population of 4,000,000, except in western Norway and among the small Lappish minority in the North. It is the language usually taught abroad as "Norwegian."

Dialects

The teaching of the standard languages in the schools and the high levels of literacy have tended to spread the urban norms of speaking. Nevertheless, very diverse dialects, partially unintelligible to outsiders, are spoken in many rural communities; some of them are used occasionally for the writing down of local traditions or for giving local colour in plays and novels. Dialect institutes for their study exist in each country. Boundaries between dialect areas are gradual and do not always coincide with national borders, so that the following traditional divisions are somewhat arbitrary: in Denmark, West (Jutland), Central (Fyn, Sjœlland), and East (Bornholm); in Sweden, South (especially Skåne), Götaland, Svealand, North (Norrland), Gotland, East (Finland); in Norway, East (Lowland, Midland), Trönder (around Trondheim), North (Nordland), West. In the Faeroese language there are minor dialectal differences between the southern and northern islands; minor dialectal differences occur in Icelandic as well, but there are no clearly defined regional dialects. In the larger cities there is a range of social dialects from the everyday speech of the working classes (often similar to nearby rural speech) to the more cultivated forms of middle- and upper-class speech, including the highly formal style of courts and legislatures. Speakers of Danish, Norwegian, and Swedish normally use their own languages in communicating with one another; Norwegian and Swedish have a common phonetic base, and Norwegian and Danish share many vocabulary items.

Characteristics. *Common and distinctly Scandinavian characteristics.* North Germanic differs from West Ger-

Differences between North and West Germanic

manic (but not East Germanic) in having *ggj* and *ggv* for medial *jj* and *ww*, respectively (Old Norse *tveggja* "two," *hoggva* "hew"), *-t* for *-e* in the 2nd person singular of the strong preterite (Old Norse *namt* "you took"; *cf.* Old English *nœme*), and a reflexive possessive *sin*.

North Germanic differs from East Germanic (but not West Germanic) in that original *ē* becomes *ā* (Old Norse *máni* "moon") and original *z* becomes *r* (Old Norse *meiri* "more"); furthermore, there is a new demonstrative pronoun *þessi* "this" (Danish, Swedish, and Norwegian *denne*), back vowels are mutated to front vowels by the influence of a following *i* or *j* ("*i*-umlaut"—*a* and *ā* become *œ* and *œ̄*, *o* and *ō* become *ø* and *ø̄* [*ø* represents umlauted *o*], *u* and *ū* become *y* and *ȳ* [*y* represents umlauted *u*], *au* becomes *ey* or *øy*), and the number of unstressed vowels is reduced to three (*a, i, u*).

North Germanic differs from both West Germanic and East Germanic in the following ways: rounding of unrounded vowels by following *u* or *w* ("*u*-umlaut"—*a* and *ā* become *ǫ* and *ǭ* [*ǫ* represents a low back rounded vowel], *e* becomes *ø, i* becomes *y, ei* becomes *ey* or *øy*); loss of initial *j* and *w* in some positions (Old Norse *ungr* "young," *ár* "year," *Óðinn* "Wodan," *ull* "wool"); loss of final nasals (Old Norse *frá* "from," *fara* "fare, go"; *cf.* Old English *faran,* German *fahren*); diphthongization ("breaking") of short *e* to *ja* or *jǫ* (Old Norse *jafn* "even," *jǫrd* "earth"). It has new pronouns for the 3rd person singular (Old Norse *hann* "he," *hon* "she"); attaches the reflexive pronoun (*sik*) to the verb to make a new mediopassive in *-sk, -st,* or *-s* (*finna sik* "find oneself" became Old Norse *finnast* "be found, exist," Danish *findes*); attaches the demonstrative *inn* "that" to nouns as a definite article (Old Norse *fótrinn* "the foot," Norwegian and Swedish *foten,* Danish *foden*), except in West Jutland (possibly a later development); and uses *-t* as marker of the neuter in pronouns and adjectives (Old Norse *hvítt* "white" from *hvít-, eitt* "one" from *ein-*). Furthermore, North Germanic employed *es* (which changed to *er*) and later *sum* as an indeclinable relative pronoun; and it lost some Germanic prefixes, such as *ga-* (German *ge-*), and contains a considerable number of words such as *hestr* "horse," *fœr* or *fár* "sheep," *gríss* "pig," *gólf* "floor," and *ostr* "cheese" that do not occur in East or West Germanic.

Orthography. The five basic vowel symbols of the Latin alphabet are supplemented by a number of special symbols, mostly for umlauted vowels: thus, there is *y* (pronounced as German *ü*), *æ* (used in Danish, Norwegian, Icelandic, and Faeroese) and the corresponding *ä* (used in Swedish), *ø* (in Danish, Norwegian, and Faeroese) and the corresponding *ö* (in Swedish and Icelandic), and *å* (also written *aa,* used in Danish, Swedish, and Norwegian).

Their present-day values are not identical; Icelandic *æ* is pronounced as the diphthong sound *ai* (as the *i* in English "ice"). Icelandic also uses accents on vowels that were long in Old Norse but are now mostly diphthongs (*á, é, í, ó, ú,* and *ý*); Faeroese has the same system except for *é*. The consonant symbols are the usual Latin ones, except that *þ* (thorn) and *ð* (edh) are used in Icelandic for voiceless and voiced *th* (*ð* in Faeroese has a different value). Loanwords containing *c, q, w, x,* and *z* have generally been naturalized by substituting, respectively, *k* or *s, kv, v, ks,* and *s* (*e.g., kontakt* "contact" but Norwegian *sigar* "cigar" versus Danish and Swedish *cigar*).

Influence of borrowing on stress

Phonology. Stress is on the first syllable in native words, with sporadic exceptions for compounds. Stress on a later syllable reflects borrowing from other languages, except in Icelandic, which has stress on the first syllable of all words.

Pitch is usually high on the stressed syllable, falling at the end of a statement, rising for a yes-no question. An exception is East Norwegian and some Swedish dialects, in which the stressed syllable is low and the pitch is often rising at the end of statements. In most of Norway and Sweden and in scattered Danish dialects, there is a special word tone, by which old monosyllables have one kind of pitch while old polysyllables have another. The first pitch type is usually high or low pitch on the stressed syllable, like that in other Germanic languages, while the second is more complex and varies from region to region. In Danish the tones have been replaced by glottalization in instances in which Norwegian and Swedish have the first type.

Vowels are short before two or more consonants (with some exceptions) or when unstressed. Doubled consonants after short stressed vowels are pronounced long, except in Danish, which also does not double consonants in final position.

The Common Scandinavian vowel system contained 10 vowels, each of which could be long, short, or nasalized: front unround (*i, e,* and *æ*), front round (*y, ø,* and *ø*), back round (*u, o,* and *ø̨*), and back unround (*a*). There were three falling diphthongs: front unround (*ęi*), front round (*øy*), and back round (*au*). While most of these are still present in some dialects, there have been many changes. The nasalized vowels disappeared, though they were still present in Icelandic around 1150. Diphthongs became long vowels in Danish and Swedish in the 10th century. Short low umlauted vowels coalesced with neighbouring vowels (*æ* became *e, ø̨* became *ø, ǫ* became *o/ø*). Long *ā* (Old Norse *á*) was rounded to *å* (pronunciation similar to the *o* in English "order"; in Icelandic and West Norwegian, pronunciation is like the *ow* in "now"). In Norwegian and Swedish the rounded vowels were shifted upwards and forwards, giving "overrounded" *o* and *u* that resemble *u* and *y* respectively. The unstressed vowels *a, i,* and *u* have remained in Icelandic and Faeroese but have been partially merged in New Norwegian and Swedish (written *a, e, o*), completely merged as *ə* (the schwa sound, as *a* in English "sofa") in Danish and Dano-Norwegian, and lost in Jutland and Trönder dialects. High round vowels (*y, ȳ, øy*) have been merged with the unround vowels in Icelandic and Faeroese (and in scattered dialects elsewhere) but are still distinguished in writing. Long vowels have been diphthongized not only in many dialects (*e.g.,* Jutland, Skåne, and West Norwegian) but also in standard Icelandic and Faeroese (Icelandic *é,* pronounced [je], *ó* [ou], *á* [au], *æ* [ai]; Faeroese *í* [ui], *æ* [æa], and so on). (Symbols in brackets are phonetic symbols designating actual pronunciation.) A quantity shift took place in the late Middle Ages, in which short vowels were lengthened before single consonants and long vowels were shortened before clusters, sometimes with qualitative changes that affected different dialects differently; thus, in Swedish *veta* "know" *i* became *e* (though all the other languages have *i*).

The Common Scandinavian consonant system contained voiceless stops (*p, t, k*), voiced stops (*b, d, g*), voiceless-voiced spirants (*f/ƀ, þ/ð, x/g*), nasals (*m, n*), a sibilant (*s*), liquids (*l, r*), and glides (*w, j*). The chief changes were as follows: Short voiceless stops became voiced after vowels in Danish and neighbouring dialects and then partially opened to become spirants or glides (*tapa* became *tabe* "lose," *ūt* became *ud* "out," *kakur* became *kager* "cakes"). Velar stops (*k, g, sk*) were palatalized before front vowels to merge with *kj, gj,* and *skj,* as still occurs in Icelandic (and Jutland dialect); in Faeroese, Norwegian, Swedish, and many Danish dialects, these were fronted to *tj, dj, stj* or even opened to spirants [ç, j, š], while in Danish they reverted to *k, g,* and *sk.* Voiced *f* [ƀ] merged with *w* to become *v,* though it is still written *f* in Icelandic; in Danish both *f* and *w* have become pronounced as *w* after vowels. Voiceless *þ* became *t* (occasionally *h* in Faeroese) and voiced *þ* [*ð*] became *d,* except in Icelandic. Voiceless *x* became *h* initially before vowels, but was lost elsewhere; voiced *x* [*g*] became *g,* except in Icelandic (in Danish it has become either [j] or [w] after vowels). The *r* sound was assimilated to following dental sounds (*l, n, s, t, d*) to make a series of retroflex consonants (*ḷ, ṇ, ś, ṭ, ḍ,* pronounced with the tip of the tongue curled up towards the hard palate) in many Swedish and Norwegian dialects, including those of Oslo and Stockholm. In much the same area a "dark" *l* was merged with *rð* to make a new "thick *l*," defined as a "cacuminal flap" not acceptable in standard speech. The *r* sound became a uvular *r* in Danish and in the dialects of the nearest parts of Sweden and Norway in the last century or two.

Morphology. The Common Scandinavian system of inflections in nouns, adjectives, pronouns, and verbs is almost totally preserved in Icelandic, if allowance is made for some sound changes (*e.g., -r* becomes *-ur* as in *situr* "sits," and *-t* becomes *ð* as in *húsið* "the house"). In Faeroese and New Norwegian the genitive case is replaced in speech by prepositional phrases or compounds. Declensions in Faeroese have been simplified in the plurals of nouns (all end in *-r*), verb plurals (*-a* in present, *-u* in preterite), verb singulars (*-i* in weak present for all persons), and the subjunctive (*-i* in the present, same ending as indicative in the preterite). In the remaining languages all case forms except the genitive (which invariably ends in *-s*) have been merged, as have markers of person and number in the verbs. This simplification began in Danish and spread to Norwegian and Swedish between 1200 and 1500, perhaps under the influence of Low German.

Major similarities

The present-day system of Danish, Dano-Norwegian, New Norwegian, and Swedish is basically identical. Nouns have singular and plural forms, to which the definite article may be suffixed; the plural suffixes vary, reflecting earlier stem, gender, and umlaut classes. Adjectives have neuter singulars marked by *-t,* plurals marked by a vowel (*-e* or *-a*), and weak forms used after determiners, usually identical in form with the plurals; the comparatives are marked by *r* and superlatives by the cluster *st.* Adverbs derived from adjectives are identical in form with the neuter singular forms of the adjectives. Personal pronouns occur in three persons and in both singular and plural. In part, they still distinguish nominative and accusative (*e.g.,* Swedish *jag* "I"—*mig* "me"). There are polite pronouns of address that are either identical with the 2nd person plural (Swedish *ni,* Icelandic *þér,* Faeroese *tygum,* and New Norwegian *de* or *dykk*) or the 3rd person plural (Danish or Dano-Norwegian *De*); in Icelandic and Faeroese old duals have taken over the function of plurals (Icelandic *við* "we," *þið* "you"; Faeroese *vit* "we," *tit* "you"). Each personal pronoun has a corresponding possessive pronoun, the 3rd person being identical with the genitive of the pronoun and invariable. The possessive pronouns for the other persons and the reflexive *sin* are inflected for gender and number like most other pronouns and articles. Verbs inflect for tense only, with *-r* as the usual present marker (New Norwegian does not have an ending to indicate present tense in the strong verbs), while the preterites have stem-vowel ablaut changes in the strong verbs and a dental suffix in the weak verbs. Nonfinite forms of the verb have invariable suffixes (*-a* or *-e* for the infinitive, *-ande* or *-ende* for present participles, and *-at* or *-et* for perfect participles), except that Swedish and New Norwegian mark gender when the perfect participle is used adjectivally.

Major differences

New Norwegian, like Icelandic and Faeroese, and, in part, Dano-Norwegian preserve masculine, feminine, and neuter genders; Danish and Swedish combine masculine and feminine into a common (non-neuter) gender. Swedish and New Norwegian (in part) preserve non-neuter plurals in *-ar, -er,* and *-or,* which merged as *-er* in Dano-Norwegian; in Danish these have become *-e,* while a new plural in *-er* has arisen, primarily for loanwords. The preterite of first class weak verbs (Old Norse *-aði*) ends in *-a* in New Norwegian, *-et* in Dano-Norwegian, *-ede* in Danish, and *-ade* in Swedish (usually pronounced *-a*). In Norwegian and Swedish a new class of weak verbs with preterite ending *-dde* has arisen, including stems ending in *-d* or long vowels (Swedish *födde* "bore," *bodde* "lived"). The present tense form of strong verbs is umlauted in New Norwegian (as in Icelandic and Faeroese); it is monosyllabic in New Norwegian, has high or low pitch on the stressed syllable in Dano-Norwegian and Swedish, and glottalization in Danish (New Norwegian *kjem;* Dano-Norwegian, Swedish *kommer,* pronounced *'kåmmər;* Danish *kommer,* pronounced *kåmʔər.* New Norwegian has *-st* in the mediopassive (like Icelandic and Faeroese); Dano-Norwegian, Swedish, and Danish have *-s.*

Normal word order in sentences

Syntax. The reduction of morphological complexity has been accompanied by the emergence of a more rigid order of sentence elements. Normal order is subject–finite verb–indirect object–direct object. The verb must precede the subject in yes–no questions, or when any part of the predicate is put first. Contrary to German practice, the verb keeps its normal place in subordinate clauses, except that negatives and other lightly stressed adverbs usually precede the finite verb (except in Icelandic). Complex verb phrases are formed with modal auxiliaries (*e.g., kan* "can") and infinitives or with the perfect auxiliaries *ha(ve)* "have" and *få* "get" (Icelandic *geta*) and the perfect participle. Instead of such durative aspect markers as the English progressive (*e.g.,* "is talking"), verbs indicating position are combined with the main verb (*e.g.,* Dano-Norwegian *han sitter* [*står, går, ligger*] *og prater* "he is sitting [standing, walking, lying] and talking."). Icelandic has special constructions for present and perfect aspects (*er að ganga* "is going" or *er buinn að ganga* "is through going").

Major differences in the Norwegian languages, Swedish, and Danish are few: (1) New Norwegian and Swedish use the nominative after a copula (*Det er eg/jag* "It is I"), Dano-Norwegian and Danish, the accusative (*Det er meg/mig* "It is me"). (2) A complex passive is formed either with Old Scandinavian *verða* (Swedish *varda,* New Norwegian *verta*) or Low German *bliven* (Danish *blive,* Dano-Norwegian *bli*) and the perfect participle. (3) Swedish supplements the polite pronoun of address with a pronominal use of titles: *Önskar professoren kaffe?* "Do you [professor] wish coffee?" (4) The reflexive pronoun *sin* is used with singular or plural subjects, except in Danish, in which it is used only with singular subjects. (5) A definite article is indicated by a form before the adjective and a suffix after the noun ("double definite"), except in Icelandic and Danish (*e.g.,* in Norwegian and Swedish *det store* [*stora*] *huset* "the big house," both *det* and *-et* in *huset* mean "the," in Danish the suffix *-et* is not used: *det store hus*). (6) A possessive may follow its noun in Icelandic, Faeroese, and Norwegian but not in Danish or Swedish (Icelandic *hesturinn minn* "my horse," literally, "horse mine," Swedish *min häst* "my horse"). (7) The numeral "one" is used (in unstressed form) as an indefinite article (*i.e.,* as "a," "an"), except in Icelandic, which has no indefinite article. (8) Swedish omits the auxiliary *hava* "have" in subordinate clauses (*Huset jag sett . . .* "The house I [have] seen . . . ").

Vocabulary. The everyday stock of Scandinavian words, including most of the high frequency words, is Indo-European and Germanic in its core. Of the 200,000 or more entries in the large dictionaries of each language, the vast majority are either compounds and derivatives of the simpler words or else borrowings from other languages—mostly of a scientific and cultural nature. At the present time the chief source of loanwords is English.

Borrowings

Icelandic preserved the creative powers of the older language by making it a policy not to accept new words in unassimilated form. Whenever possible, new compounds and derivatives have been created to avoid the borrowing of foreign terms. To some extent Faeroese and New Norwegian have followed the same policy but without the success of Icelandic. Danish, Swedish, and Dano-Norwegian have adopted numerous German words, along with their prefixes and suffixes; *e.g.,* Danish and Norwegian *betale* and Swedish *betala* "pay" from Low German *betalen (cf.* Icelandic and Faeroese *gjalda, borga).* A knowledge of German is very helpful in learning to read Norwegian, Swedish, and Danish, but this is less true for Icelandic and Faeroese, which baffle even fellow Scandinavians.

The borrowings of Danish, Swedish, and Norwegian reflect the varied contacts discussed above. Their vocabulary consists of a native core, a German middle layer (with words like Danish *skrædder* "tailor"; *cf.* Icelandic and Faeroese *klæðskeri,* literally "cloth-cutter"), and an international outer layer (with words like *psykologi* "psychology"; *cf.* Icelandic and Faeroese *sálfræði,* literally, "soul science"). While there are some differences among the languages in the exact composition of these layers, there is also considerable agreement. Differences occur especially in words of local origin (slang, humour, endearments, abuse) and in borrowings of different origin; *e.g.,* Norwegian *etasje*/Swedish *våning*/Danish *sal* "story" (in a hotel), from French *étage,* Middle Low German *woninge,* and Old Scandinavian *salr* (but with its meaning from North German *Saal*). (Ei.H.)

English language

English is widely spoken in all six continents and has had a strong effect in many regions in which it is not the principal language spoken. The portmanteau words Franglais, Russlish, and Japlish, for example, have been invented by resentful purists to describe the numerous expressions, in vogue among the young, resulting from the infiltration of English into French, Russian, and Japanese, respectively.

The widely diffused English-speaking community is fairly stable in the British Isles, North America, and Australasia. In Africa, the Indian subcontinent, and Southeast Asia its future remains uncertain and unpredictable. People who speak English fall into three groups: those who have inherited it as their native language; those who have acquired it as a second language within a society that is largely bilingual; and those who are driven by necessity to use it for some practical purpose—administrative, professional, or educational. One person in seven of the world's entire population now belongs to one of these three groups.

ORIGINS AND BASIC CHARACTERISTICS

Parent tongue

English belongs to the Indo-European family of languages and is therefore related to most other languages spoken in Europe and western Asia from Iceland to India. The parent tongue, called Proto-Indo-European, was spoken about 5,000 years ago by nomads believed to have roamed the southeast European plains. Germanic, one of the language groups descended from this ancestral speech, is usually divided by scholars into three regional groups: East (Burgundian, Vandal, and Gothic, all extinct), North (Icelandic, Faeroese, Norwegian, Swedish, Danish), and West (German, Netherlandic [Dutch and Flemish], Frisian, English). Though closely related to English, German remains far more conservative than English in its retention of a fairly elaborate system of inflections. Frisian, spoken by the inhabitants of the Dutch province of Friesland and the islands off the west coast of Schleswig, is the language most nearly related to Modern English. Icelandic, which has changed little over the last thousand years, is the living language most nearly resembling Old English in grammatical structure.

Modern English is analytic (*i.e.,* relatively uninflected), whereas Proto-Indo-European, the ancestral tongue of most of the modern European languages (*e.g.,* German, French, Russian, Greek), was synthetic, or inflected. During the course of thousands of years, English words have been slowly simplified from the inflected variable forms found in Sanskrit, Greek, Latin, Russian, and German, toward invariable forms, as in Chinese and Vietnamese. The German and Chinese words for "man" are exemplary. German has five forms: *Mann, Mannes, Manne, Männer, Männern.* Chinese has one form: *jen.* English stands in between, with four forms: man, man's, men, men's. In English only nouns, pronouns, and verbs are inflected. Adjectives have no inflections aside from the determiners "this, these" and "that, those." (The endings *-er, -est,* denoting degrees of comparison, are better regarded as noninflectional suffixes.) English is the only European language to employ uninflected adjectives; *e.g.,* "the tall man," "the tall woman," compared to Spanish *el hombre alto* and *la mujer alta.* As for verbs, if the Modern English word ride is compared with the corresponding words in Old English and Modern German, it will be found that English now has only five forms (ride, rides, rode, riding, ridden), whereas Old English *ridan* had 13, and Modern German *reiten* has 16 forms.

In addition to this simplicity of inflections, English has two other basic characteristics: flexibility of function and openness of vocabulary.

Loss of inflection

Flexibility of function has grown over the last five centuries as a consequence of the loss of inflections. Words formerly distinguished as nouns or verbs by differences in their forms are now often used as both nouns and verbs. One can speak, for example, of "planning a table" or "tabling a plan," "booking a place" or "placing a book," "lifting a thumb" or "thumbing a lift." In the other Indo-European languages, apart from rare exceptions in Scandinavian, nouns and verbs are never identical because of the necessity of separate noun and verb endings. In English, forms for traditional pronouns, adjectives, and adverbs can also function as nouns; adjectives and adverbs as verbs; and nouns, pronouns, and adverbs as adjectives. One speaks in English of the Frankfurt Book Fair, but in German one must add the suffix *-er* to the place-name and put attributive and noun together as a compound, Frankfurter Buchmesse. In French one has no choice but to construct a phrase involving the use of two prepositions: Foire du Livre de Francfort. In English it is now possible to employ a plural noun as adjunct (modifier), as in "wages board" and "sports editor"; or even a conjunctional group, as in "prices and incomes policy" and "parks and gardens committee."

Openness of vocabulary implies both free admission of words from other languages and the ready creation of compounds and derivatives. English adopts (without change) or adapts (with slight change) any word really needed to name some new object or to denote some new process. Like French, Spanish, and Russian, English frequently forms scientific terms from Classical Greek word elements.

English possesses a system of orthography that does not always accurately reflect the pronunciation of words; this is discussed below in the section *Orthography.*

CHARACTERISTICS OF MODERN ENGLISH

Phonology. British Received Pronunciation (RP), by definition, the usual speech of educated people living in London and southeastern England, is one of the many forms of standard speech. Other pronunciations, although not standard, are entirely acceptable in their own right on conversational levels.

The chief differences between British Received Pronunciation, as defined above, and a variety of American English, such as Inland Northern (the speech form of western New England and its derivatives, often popularly referred to as General American), are in the pronunciation of certain individual vowels and diphthongs. Inland Northern American vowels sometimes have semiconsonantal final glides (*i.e.,* sounds resembling initial *w,* for example, or initial *y*). Aside from the final glides, this American dialect shows four divergences from British English: (1) the words cod, box, dock, hot, and not are pronounced with a short (or half-long) low front sound as in British "bard" shortened (the terms front, back, low, and high refer to the position of the tongue); (2) words such as bud, but, cut, and rung are pronounced with a central vowel as in the unstressed final syllable of "sofa"; (3) before the fricative sounds *s, f,* and *θ* (the last of these is the *th* sound in "thin") the long low back vowel *a,* as in British "bath," is pronounced as a short front vowel *a,* as in British "bad"; (4) high back vowels following the alveolar sounds *t* and *d* and the nasal sound *n* in words such as tulips, dew, and news are pronounced without a glide as in British English; indeed, the words sound like the British "two lips," "do," and "nooze" in "snooze." (In several American dialects, however, these glides do occur.)

The consonants

The 24 consonant sounds comprise six stops (plosives): *p, b, t, d, k, g;* the fricatives *f, v, θ* (as in "thin"), *ð* (as in "then"), *s, z, ʃ* (as in "ship"), ʒ (as in "pleasure"), and *h;* two affricatives: *tʃ* (as in "church") and *dʒ* (as the *j* in "jam"); the nasals *m, n,* ŋ (the sound that occurs at the end of words such as "young"); the lateral *l;* the vibrant or retroflex *r;* and the semivowels *j* (often spelled *y*) and *w.* These remain fairly stable, but Inland Northern American differs from British English in two respects: (1) *r* following vowels is preserved in words such as "door," "flower," and "harmony," whereas it is lost in British; (2) *t* between vowels is voiced, so that "metal" and "matter" sound very much like British "medal" and "madder," although the pronunciation of this *t* is softer and less aspirated, or breathy, than the *d* of British English.

Like Russian, English is a strongly stressed language. Four degrees of stress may be differentiated: primary, secondary, tertiary, and weak, which may be indicated, respectively, by acute (´), circumflex (ˆ), and grave (`) accent marks and by the breve (◡). Thus, "Têll mè thĕ trúth" (the whole truth, and nothing but the truth) may be contrasted with "Têll mé thĕ trûth" (whatever you may tell other people); "blác̆k bîrd" (any bird black in colour) may be contrasted with "bláckbìrd" (that particular bird *Turdus merula*). The verbs "permít" and "recórd" (henceforth only primary stresses are marked) may be contrasted with their corresponding nouns "pérmit" and "récord." A feeling for antepenultimate (third syllable from the end) primary stress, revealed in such five-syllable words as equanímity, longitúdinal, notoríety, opportúnity, parsimónious, pertinácity, and vegetárian, causes stress to shift when extra syllables are added, as in "histórical," a derivative of "hístory" and "theatricálity," a derivative of "theátrical." Vowel qualities are also changed here and in such word groups as périod, periódical, periodícity; phótograph, photógraphy, photográphical. French stress may be sustained in many borrowed words; *e.g.,* bizárre, critíque, duréss, hotél, prestíge, and techníque.

Pitch, or musical tone, determined by the rate of vibration of the vocal cords, may be level, falling, rising, or falling–rising. In counting "one," "two," "three," "four," one naturally gives level pitch to each of these cardinal numerals. But if a person says "I want two, not one," he naturally gives "two" falling pitch and "one" falling–rising. In the question "One?" rising pitch is used. Word tone is called pitch, and sentence tone is referred to as intonation. The end-of-sentence cadence is important for meaning, and it therefore varies least. Three main end-of-sentence intonations can be distinguished: falling, rising, and falling–rising. Falling intonation is used in completed statements, direct commands, and sometimes in general questions unanswerable by "yes" or "no"; *e.g.,* "I have nothing to add." "Keep to the right." "Who told you that?" Rising intonation is frequently used in open-ended statements made with some reservation, in polite requests, and in particular questions answerable by "yes" or "no": "I have nothing more to say at the moment." "Let me know how you get on." "Are you sure?" The third type of end-of-sentence intonation, first falling and then rising pitch, is used in sentences that imply concessions or contrasts: "Some people do like them" (but others do not). "Don't say I didn't warn you" (because that is just what I'm now doing). Intonation is on the whole less singsong in American than in British English, and there is a narrower range of pitch. American speech may seem more monotonous but at the same time may sometimes be clearer and more readily intelligible. Everywhere English is spoken, regional dialects display distinctive patterns of intonation.

Intonation in American and British English

Morphology. *Inflection.* Modern English nouns, pronouns, and verbs are inflected. Adjectives, adverbs, prepositions, conjunctions, and interjections are invariable.

Most English nouns have plural inflection in *(-e)s,* but this form shows variations in pronunciation in the words cats (with a final *s* sound), dogs (with a final *z* sound), and horses (with a final *iz* sound), as also in the 3rd person singular present-tense forms of verbs: cuts (*s*), jogs (*z*), and forces (*iz*). Seven nouns have mutated (umlauted) plurals: man, men; woman, women; tooth, teeth; foot, feet; goose, geese; mouse, mice; louse, lice. Three have plurals in *-en:* ox, oxen; child, children; brother, brethren. Some remain unchanged; *e.g.,* deer, sheep, moose, grouse. Five of the seven personal pronouns have distinctive forms for subject and object.

The forms of verbs are not complex. Only the substantive verb ("to be") has eight forms: be, am, is, are, was, were, being, been. Strong verbs have five forms: ride, rides, rode, riding, ridden. Regular or weak verbs customarily have four: walk, walks, walked, walking. Some that end in a *t* or *d* have three forms only: cut, cuts, cutting. Of these three-form verbs, 16 are in frequent use.

In addition to the above inflections, English employs two other main morphological (structural) processes—affixation and composition—and two subsidiary ones—back-formation and blend.

Types of affixes

Affixation. Affixes, word elements attached to words, may either precede, as prefixes (do, undo; way, subway), or follow, as suffixes (do, doer; way, wayward). They may be native (*over*do, wayward*ness*), Greek (*hyper*bole, thes*is*), or Latin (*super*sede, pedi*ment*). Modern technologists greatly favour the neo-Hellenic prefixes *macro-*"long, large," *micro-* "small," *para-* "alongside," *poly-* "many," and the Latin *mini-,* with its antonym *maxi-*. Greek and Latin affixes have become so fully acclimatized that they can occur together in one and the same word, as, indeed, in "ac-climat-ize-d," just used, consisting of a Latin prefix plus a Greek stem plus a Greek suffix plus an English inflection. Suffixes are bound more closely than prefixes to the stems or root elements of words. Consider, for instance, the wide variety of agent suffixes in the nouns act*or,* artis*an,* dot*ard,* engin*eer,* financ*ier,* hire*ling,* magistr*ate,* merch*ant,* scient*ist,* secret*ary,* son*gster,* stud*ent,* and work*er.* Suffixes may come to be attached to stems quite fortuitously, but, once attached, they are likely to be permanent. At the same time, one suffix can perform many functions. The suffix *-er* denotes the doer of the action in the words worker, driver, and hunter; the instrument in chopper, harvester, and roller; and the dweller in Icelander, Londoner, and Trobriander. It refers to things or actions associated with the basic concept in the words breather, "pause to take breath"; diner, "dining car on a train"; and fiver, "five-pound note." In the terms disclaimer, misnomer, and rejoinder (all from French) the suffix denotes one single instance of the action expressed by the verb. Usage may prove capricious. Whereas a writer is a person, a typewriter is a machine. For some time a computer was both, but now, with the invention and extensive use of electronic apparatus, the word is no longer used of persons.

Composition. Composition, or compounding, is concerned with free forms. The primary compounds "already," "cloverleaf," and "gentleman" show the collocation of two free forms. They differ from word groups or phrases in phonology, stress, or juncture or by a combination of two or more of these. Thus, "already" differs from "all ready" in stress and juncture, "cloverleaf" from "clover leaf" in stress, and "gentleman" from "gentle man" in phonology, stress, and juncture. In describing the structure of compound words it is necessary to take into account the relation of components to each other and the relation of the whole compound to its components. These relations diverge widely in, for example, the words cloverleaf, icebreaker, breakwater, blackbird, peace-loving, and paperback. In "cloverleaf" the first component noun is attributive and modifies the second, as also in the terms aircraft, beehive, landmark, lifeline, network, and vineyard. "Icebreaker," however, is a compound made up of noun object plus agent noun, itself consisting of verb plus agent suffix, as also in the words bridgebuilder, landowner, metalworker, minelayer, and timekeeper. The next type consists of verb plus object. It is rare in English, Dutch, and German but frequent in French, Spanish, and Italian. The English "pastime" may be compared, for example, with French *passe-temps,* the Spanish *pasatiempo,* and the Italian *passatempo.* From French comes "passport," meaning "pass (*i.e.,* enter) harbour." From Italian comes "portfolio," meaning "carry leaf." Other words of this type are daredevil, scapegrace, and scarecrow. As for the "blackbird" type, consisting of attributive adjective plus noun, it occurs frequently, as in the terms bluebell, grandson, shorthand, and wildfire. The next type, composed of object noun and a present participle, as in the terms fact-finding, heart-rending (German *herzzerreissend*), life-giving (German *lebenspendend*), painstaking, and time-consuming, occurs rarely. The last type is seen in barefoot, bluebeard, hunchback, leatherneck, redbreast, and scatterbrain.

Compound words

Back-formations and blends. Back-formations and blends are becoming increasingly popular. Back-formation is the reverse of affixation, being the analogical creation of a new word from an existing word falsely assumed to be its derivative. For example, the verb "to edit" has been formed from the noun "editor" on the reverse analogy of the noun "actor" from "to act," and similarly the verbs automate, bulldoze, commute, escalate, liaise, loaf, sightsee, and televise are backformed from the nouns automation, bulldozer, commuter, escalation, liaison, loafer, sightseer, and television. From the single noun "procession" are backformed two verbs with different stresses and meanings: procéss, "to walk in procession," and prócess, "to subject food (and other material) to a special operation."

Two types of blends

Blends fall into two groups: (1) coalescences, such as "bash" from "bang" and "smash"; and (2) telescoped forms, called portmanteau words, such as "motorcade" from "motor cavalcade." In the first group are the words clash, from clack and crash, and geep, offspring of goat and sheep. To the second group belong dormobiles, or dormitory automobiles, and slurbs, or slum suburbs. A travel monologue becomes a travelogue and a telegram sent by cable a cablegram. Aviation electronics becomes avionics; biology electronics, bionics; and nuclear electronics, nucleonics. In cablese a question mark is a quark; in computerese a binary unit is a bit. In astrophysics a quasistellar source of radio energy becomes a quasar, and a pulsating star becomes a pulsar.

Simple shortenings, such as "ad" for "advertisement," have risen in status. They are listed in dictionaries side by side with their full forms. Among such fashionable abbreviations are exam, gym, lab, lib, op, spec, sub, tech, veg, and vet. Compound shortenings, after the pattern of Russian *agitprop* for *agitatsiya propaganda,* are also becoming fashionable. Initial syllables are joined as in the words Fortran, for formula (computer) translation; mascon, for massive (lunar) concentration; and Tacomsat, for Tactical Communications Satellite.

Syntax. Sentences can be classified as (1) simple, containing one clause and predication: "John knows this country"; (2) multiple or compound, containing two or more coordinate clauses: "John has been here before, and he knows this country"; and (3) complex, containing one or more main clauses and one or more subordinate clauses: "John, who has been here before, knows this country" or "Because he has been here before, John knows this country." Simple, declarative, affirmative sentences have two main patterns with five subsidiary patterns within each. Verb and complement together form the predicate. "Complement" is here used to cover both the complement and the object of traditional grammarians (see Table 23).

Table 23: Simple Sentences—First Pattern

subject	verb	complement
1. John	knows	this country
2. Science	is	organized knowledge
3. Elizabeth	becomes	queen
4. The captain	falls	sick
5. Nothing	passes	unobserved

In (1) the complement is the direct object of a transitive verb; in (2) it is a predicative nominal group forming the second component of an equation linked to the first part by the meaningless copula is; in (3) it is a predicative noun linked with the subject by the meaningful copula becomes; in (4) it is a predicative adjective; and in (5) it is a predicative past participle.

In Table 24 each sentence contains four components: subject, verb, and two complements, first and second, or inner and outer. In (6) inner and outer complements consist of indirect object (without preposition) followed by direct object; in (7) these complements are direct object and appositive noun; in (8) direct object and predicative adjective; in (9) direct object and predicative past participle; in (10) direct object and predicative infinitive.

Table 24: Simple Sentences—Second Pattern

subject	verb	inner complement	outer complement
6. John	gives	Mary	a ring
7. The sailors	make	John	captain
8. You	have kept	your record	clean
9. The driver	finds	the road	flooded
10. We	want	you	to know

One can seldom change the word order in these 10 sentences without doing something else—adding or subtracting a word, changing the meaning. There is no better way of appreciating the importance of word position than by scrutinizing the 10 frames illustrated. If, for instance, in (6) one reverses inner and outer complements, one adds "to" and says, "John gives a ring to Mary"; one does not say "John gives a ring Mary." Some verbs, such as "explain" and "say," never omit the preposition "to" before the indirect object: "John's father explained the details to his son." "He said many things to him." If, in (10), the inner and outer complements are reversed (*e.g.,* "We want to know you"), the meaning is changed as well as the structure.

Positions of adjectives

Apart from these fundamental rules of word order, the principles governing the positions of adjectives, adverbs, and prepositions call for brief comment. For attributive adjectives the rule is simple: single words regularly precede the noun, and word groups follow—*e.g.,* "an unforgettable experience" but "an experience never to be forgotten." There is a growing tendency, however, to abandon this principle, to switch groups to front position, and to say "a never to be forgotten experience." In the ordering of multiple epithets, on the other hand, some new principles are seen to be slowly emerging. Attributes denoting permanent qualities stand nearest their head nouns: "long, white beard," "six-lane elevated freeway." The order in multiple attribution tends to be as follows: determiner; quantifier; adjective of quality; adjective of size, shape, or texture; adjective of colour or material; noun adjunct (if any); head noun. Examples include: "that one solid, round, oak dining table," "these many fine, large, black race horses," "those countless memorable, long, bright summer evenings."

Adverbs are more mobile than adjectives. Nevertheless, some tentative principles seem to be at work. Adverbs of frequency tend to come immediately after the substantive verb ("You are often late"), before other verbs ("You never know"), and between auxiliaries and full verbs ("You can never tell"). In this last instance, however, American differs from British usage. Most Americans would place the adverb before the auxiliary and say "You never can tell." (In the title of his play of that name, first performed in 1899, George Bernard Shaw avowedly followed American usage.) Adverbs of time usually occur at the beginning or end of a sentence, seldom in the middle. Particular expressions normally precede more general ones: "Neil Armstrong set foot on the Moon at 4 o'clock in the morning on July 21, 1969." An adverb of place or direction follows a verb with which it is semantically bound: "We arrived home after dark." Other adverbs normally take end positions in the order of manner, place, and time: "Senator Smith summed it all up most adroitly [manner] in Congress [place] last night [time]."

In spite of its etymology (Latin *prae-positio* "before placing"), a preposition may sometimes follow the noun it governs, as in "all the world over," "the clock round," and "the whole place through." "This seems a good place to live in" seems more natural to most speakers than "This seems a good place in which to live." "Have you anything to open this can with?" is now more common than "Have you anything with which to open this can?"

The above are principles rather than rules, and in the end it must be agreed that English syntax lacks regimentation. Its structural laxity makes English an easy language to speak badly. It also makes English prone to ambiguity. "When walking snipe always approach up wind," a shooting manual directs. The writer intends the reader to understand, "When you are walking to flush snipe always approach them up against the wind." "John kept the car in the garage" can mean either (1) "John retained that car you see in the garage, and sold his other one" or (2) "John housed the car in the garage, and not elsewhere." "Flying planes can be dangerous" is ambiguous because it may mean either (1) "Planes that fly can be dangerous" or (2) "It is dangerous to fly planes."

Two ways in which "John gives Mary a ring" can be stated in the passive are: (1) "A ring is given to Mary by John" and (2) "Mary is given a ring by John." Concerning this same action, four types of question can be formulated: (1) "Who gives Mary a ring?" The information sought is the identity of the giver. (2) "Does John give Mary a ring?" The question may be answered by "yes" or "no." (3) "John gives Mary a ring, doesn't he?" Confirmation is sought of the questioner's belief that John does in fact give Mary a ring. (4) "John gives Mary a ring?" This form, differing from the declarative statement only by the question mark in writing, or by rising intonation in speech, calls, like sentences (2) and (3), for a "yes" or "no" answer but suggests doubt on the part of the questioner that the action is taking place.

Origins of English words

Vocabulary. The vocabulary of Modern English is approximately half Germanic (Old English and Scandinavian) and half Italic or Romance (French and Latin), with copious and increasing importations from Greek in science and technology and with considerable borrowings from Dutch, Low German, Italian, Spanish, German, Arabic, and many other languages. Names of basic concepts and things come from Old English or Anglo-Saxon: heaven and earth, love and hate, life and death, beginning and end, day and night, month and year, heat and cold, way and path, meadow and stream. Cardinal numerals come from Old English, as do all the ordinal numerals except "second" (Old English *other,* which still retains its older meaning in "every other day"). "Second" comes from Latin *secundus* "following," through French *second,* related to Latin *sequi* "to follow," as in English "sequence." From Old English come all the personal pronouns (except "they," "their," and "them," which are from Scandinavian), the auxiliary verbs (except the marginal "used," which is from French), most simple prepositions, and all conjunctions.

Numerous nouns would be identical whether they came from Old English or Scandinavian: father, mother, brother (but not sister); man, wife; ground, land, tree, grass; summer, winter; cliff, dale. Many verbs would also be identical, especially monosyllabic verbs—bring, come, get, hear, meet, see, set, sit, spin, stand, think. The same is true of the adjectives full and wise; the colour names gray, green, and white; the disjunctive possessives mine and thine (but not ours and yours); the terms north and west (but not south and east); and the prepositions over and under. Just a few English and Scandinavian doublets coexist in current speech: no and nay, yea and ay, from and fro, rear (*i.e.,* to bring up) and raise, shirt and skirt (both related to the adjective short), less and loose. From Scandinavian, "law" was borrowed early, whence "bylaw," meaning "village law," and "outlaw," meaning "man outside the law." "Husband" (*hus-bondi*) meant "householder," whether single or married, whereas "fellow" (*fe-lagi*) meant one who "lays fee" or shares property with another, and so "partner, shareholder." From Scandinavian come the common nouns axle (tree), band, birth, bloom, crook, dirt, egg, gait, gap, girth, knife, loan, race, rift, root, score, seat, skill, sky, snare, thrift, and window; the adjectives awkward, flat, happy, ill, loose, rotten, rugged, sly, tight, ugly, weak, and wrong; and many verbs, including call, cast, clasp, clip, crave, die, droop, drown, flit, gape, gasp, glitter, life, rake, rid, scare, scowl, skulk, snub, sprint, thrive, thrust, and want.

Borrowing from French

The debt of the English language to French is large. The terms president, representative, legislature, congress, constitution, and parliament are all French. So, too, are duke, marquis, viscount, and baron; but king, queen, lord, lady, earl, and knight are English. City, village, court, palace, manor, mansion, residence, and domicile are French; but town, borough, hall, house, bower, room, and home are English. Comparison between English and French synonyms shows that the former are more human and concrete, the latter more intellectual and abstract; *e.g.,* the terms freedom and liberty, friendship and amity, hatred and enmity, love and affection, likelihood and probability, truth and veracity, lying and mendacity. The superiority of French cooking is duly recognized by the adoption of such culinary terms as boil, broil, fry, grill, roast, souse, and toast. "Breakfast" is English, but "dinner" and "supper" are French. "Hunt" is English, but "chase," "quarry," "scent," and "track" are French. Craftsmen bear names of English origin: baker, builder, fisher (man), hedger, miller, shepherd, shoemaker, wainwright, and weaver, or webber. Names of skilled artisans, however, are French: carpenter, draper, haberdasher, joiner, mason, painter, plumber, and tailor. Many terms relating to dress and fashion, cuisine and viniculture, politics and diplomacy, drama and literature, art and ballet come from French.

In the spheres of science and technology many terms come from Classical Greek through French or directly from Greek. Pioneers in research and development now regard Greek as a kind of inexhaustible quarry from which they can draw linguistic material at will. By prefixing the Greek adverb *tēle* "far away, distant" to the existing compound photography, "light writing," they create the precise term "telephotography" to denote the photographing of distant objects by means of a special lens. By inserting the prefix *micro-* "small" into this same compound, they make the new term "photomicrography," denoting the electronic photographing of bacteria and viruses. Such neo-Hellenic derivatives would probably have been unintelligible to Plato and Aristotle. Many Greek compounds and derivatives have Latin equivalents with slight or considerable differentiations in meaning (see Table 25).

Greek borrowings

Table 25: Equivalent Compounds and Derivatives*

from the Greek	from the Latin
Nouns	
dys-troph-y	mal-nutr-it-ion
hypo-sta-sis	sub-stan-ce
hypo-the-sis	sup-pos-it-ion
meta-morph-o-sis	trans-form-at-ion
meta-phor	trans-fer
meta-the-sis	trans-pos-it-ion
peri-pher-y	circum-fer-en-ce
peri-phra-sis	circum-loc-ut-ion
sym-path-y	com-pass-ion
syn-drom-e	con-curr-en-ce
syn-op-sis	con-spect-us
syn-the-sis	com-pos-it-ion
sy-zyg-y	con-junc-t-ion
Adjectives	
dia-phan-*ous*	trans-par-ent
hyper-aesth-et-ic	super-sens-it-ive
hyper-phys-ic-*al*	super-nat-ur-al
hypo-derm-ic	sub-cut-an-eous
hypo-ge-*al*	sub-terr-an-ean
melan-chol-ic	atra-bil-ious
mono-morph-ic	uni-form
oxy-phyll-*ous*	acut-i-fol-i-ate
peri-pat-et-ic	circum-amb-ul-at-ory
phos-phor-*escent*	lumin-i-fer-ous
poly-glott-*al*	multi-lingu-al
sphen-oid	cunei-form
syn-chron-ic	con-temp-or-ary

*The italicized suffixes *-al*, *-escent*, and *-ous*, attached to some of the Greek adjectives, are of Latin origin.

At first sight it might appear that some of these equivalents, such as "metamorphosis" and "transformation," are sufficiently synonymous to make one or the other redundant. In fact, however, "metamorphosis" is more technical and therefore more restricted than "transformation." In mythology it signifies a magical shape changing; in nature it denotes a postembryonic development such as that of a tadpole into a frog, a cocoon into a silkworm, or a chrysalis into a butterfly. Transformation, on the other hand, means any kind of change from one state to another.

Ever since the 12th century, when merchants from the Netherlands made homes in East Anglia, Dutch words have infiltrated into Midland speech. For centuries a form of Low German was used by seafaring men in North Sea ports. Old nautical terms still in use include buoy, deck, dock, freebooter, hoist, leak, pump, skipper, and yacht. The Dutch in New Amsterdam (later New York) and adjacent settlements gave the words boss, cookie, dope, snoop, and waffle to American speech. The Dutch in Cape Province gave the terms apartheid, commandeer, commando, spoor, and trek to South African speech.

German contributions

The contribution of High German has been on a different level. In the 18th and 19th centuries it lay in technicalities of geology and mineralogy and in abstractions relating to literature, philosophy, and psychology. In the 20th century this contribution has sometimes been indirect. "Unclear" and "meaningful" echoed German *unklar* and *bedeutungsvoll,* or *sinnvoll.* "Ring road" (a British term applied to roads encircling cities or parts of cities) translated *Ringstrasse;* "round trip," *Rundfahrt;* and "the turn of the century," *die Jahrhundertwende.* The terms "classless society," "inferiority complex," and "wishful thinking" echoed *die klassenlöse Gesellschaft, der Minderwertigkeitskomplex,* and *das Wunschdenken.*

Along with the rest of the Western world, English has accepted Italian as the language of music. The names of voices, parts, performers, instruments, forms of composition, and technical directions are all Italian. Many of the latter—allegro, andante, cantabile, crescendo, diminuendo, legato, maestoso, obbligato, pizzicato, staccato, and vibrato—are also used metaphorically. In architecture, the terms belvedere, corridor, cupola, grotto, pedestal, pergola, piazza, pilaster, and rotunda are accepted; in literature, burlesque, canto, extravaganza, stanza, and many more are used.

From Spanish, English has acquired the words armada, cannibal, cigar, galleon, guerrilla, matador, mosquito, quadroon, tornado, and vanilla, some of these loanwords going back to the 16th century, when sea dogs encountered hidalgos on the high seas. Many names of animals and plants have entered English from indigenous languages through Spanish: "potato" through Spanish *patata* from Taino *batata,* and "tomato" through Spanish *tomate* from Nahuatl *tomatl.* Other words have entered from Latin America by way of Texas, New Mexico, Arizona, and California; *e.g.,* such words as canyon, cigar, estancia, lasso, mustang, pueblo, and rodeo. Some have gathered new connotations: bonanza, originally denoting "goodness," came through miners' slang to mean "spectacular windfall, prosperity"; mañana, "tomorrow," acquired an undertone of mysterious unpredictability.

From Arabic through European Spanish, through French from Spanish, through Latin, or occasionally through Greek, English has obtained the terms alchemy, alcohol, alembic, algebra, alkali, almanac, arsenal, assassin, attar, azimuth, cipher, elixir, mosque, nadir, naphtha, sugar, syrup, zenith, and zero. From Egyptian Arabic, English has recently borrowed the term loofah (also spelled luffa). From Hebrew, directly or by way of Vulgate Latin, come the terms amen, cherub, hallelujah, manna, messiah, pharisee, rabbi, sabbath, and seraph; jubilee, leviathan, and shibboleth; and, more recently, kosher, and kibbutz.

English has freely adopted and adapted words from many other languages, acquiring them sometimes directly and sometimes by devious routes. Each word has its own history. The following lists indicate the origins of a number of English words: Welsh—flannel, coracle, cromlech, eisteddfod; Cornish—gull, brill, dolmen; Gaelic and Irish—shamrock, brogue, leprechaun, ogham, Tory, galore, blarney, hooligan, clan, claymore, bog, plaid, slogan, sporran, cairn, whisky, pibroch; Breton—menhir, penguin; Norwegian—ski, ombudsman; Finnish—sauna; Russian—kvass, ruble, tsar, verst, mammoth, ukase, astrakhan, vodka, samovar, tundra (from Lapp), troika, pogrom, duma, soviet, bolshevik, intelligentsia (from Latin through Polish), borscht, balalaika, sputnik, soyuz, salyut, lunokhod; Polish—mazurka; Czech—robot; Hungarian—goulash, paprika; Portuguese—marmalade, flamingo, molasses, veranda, port (wine), dodo; Basque—bizarre; Turkish—janissary, turban, coffee, kiosk, caviar, pasha, odalisque, fez, bosh; Hindi—nabob, guru, sahib, maharajah, mahatma, pundit, punch (drink), juggernaut, cushy, jungle, thug, cheetah, shampoo, chit, dungaree, pucka, gymkhana, mantra, loot, pajamas, dinghy, polo; Persian—paradise, divan, purdah, lilac, bazaar, shah, caravan, chess, salamander, taffeta, shawl, khaki; Tamil—pariah, curry, catamaran, mulligatawny; Chinese—tea (Amoy), sampan; Japanese—shogun, kimono, mikado, tycoon, hara-kiri, gobang, judo, jujitsu, bushido, samurai, banzai, tsunami, satsuma, No (the dance drama), karate, Kabuki; Malay—ketchup, sago, bamboo, junk, amuck, orangutan, compound (fenced area), raffia; Polynesian—taboo, tattoo; Hawaiian—ukulele; African languages—chimpanzee, goober, mumbo jumbo, voodoo; Eskimo—kayak, igloo, anorak, mukluk; Algonkian—totem; Nahuatl—mescal; languages of the Caribbean—hammock, hurricane, tobacco, maize, iguana; Aboriginal Australian—kangaroo, corroboree, wallaby, wombat, boomerang, paramatta, budgerigar.

The alphabet

Orthography. The Latin alphabet originally had 20 letters, the present English alphabet minus *J, K, V, W, Y,* and Z. The Romans themselves added *K* for use in abbreviations and *Y* and *Z* in words transcribed from Greek. After its adoption by the English, this 23-letter alphabet developed *W* as a ligatured doubling of *U* and later *J* and *V* as consonantal variants of *I* and *U.* The resultant alphabet of 26 letters has both uppercase, or capital, and lowercase, or small, letters.

English spelling is based for the most part on that of the 15th century, but pronunciation has changed considerably since then, especially that of long vowels and diphthongs. The extensive change in the pronunciation of vowels, known as the Great Vowel Shift, affected all of Geoffrey Chaucer's seven long vowels, and for centuries spelling remained untidy. If the meaning of the message was clear, the spelling of individual words seemed unimportant. In the 17th century during the English Civil War, compositors adopted fixed spellings for practical reasons, and in the order-loving 18th century uniformity became more and more fashionable. Since Samuel Johnson's *Dictionary of the English Language* (1755), orthography has remained fairly stable. Numerous tacit changes, such as "music" for "musick" (*c.* 1880) and "fantasy" for "phantasy" (*c.* 1920), have been accepted, but spelling has nevertheless continued to be in part unphonetic. Attempts have been made at reform. Indeed, every century has had its reformers since the 13th, when an Augustinian canon named Orm devised his own method of differentiating short vowels from long by doubling the succeeding consonants or, when this was not feasible, by marking short vowels with a superimposed breve mark (˘). William Caxton, who set up his wooden printing press at Westminster in 1476, was much concerned with spelling problems throughout his working life. Noah Webster produced his *Spelling Book,* in 1783, as a precursor to the first edition (1828) of his *American Dictionary of the English Language.* The 20th century has produced many zealous reformers. Three systems, supplementary to traditional spelling, are actually in use for different purposes: (1) the Initial Teaching (Augmented Roman) Alphabet (ITA) of 44 letters used by educationists in the teaching of children under seven; (2) the Shaw alphabet of 48 letters, designed in implementation of the will of George Bernard Shaw; and (3) the International Phonetic Alphabet (IPA), constructed on the basis of one symbol for one individual sound and used by many trained linguists. Countless other systems have been worked out from time to time, of which R.E. Zachrisson's "Anglic" (1930) and Axel Wijk's *Regularized English* (1959) may be the best.

Obstacles to spelling reform

Meanwhile, the great publishing houses continue unperturbed because drastic reform remains impracticable, undesirable, and unlikely. This is because there is no longer one criterion of correct pronunciation but several standards throughout the world; regional standards are themselves not static, but changing with each new generation; and, if spelling were changed drastically, all the books in English in the world's public and private libraries would become inaccessible to readers without special study.

HISTORICAL BACKGROUND

Among highlights in the history of the English language, the following stand out most clearly: the settlement in Britain of Jutes, Saxons, and Angles in the 5th and 6th centuries; the arrival of St. Augustine in 597 and the subsequent conversion of England to Latin Christianity; the Viking invasions of the 9th century; the Norman Conquest of 1066; the Statute of Pleading in 1362 (this required that court proceedings be conducted in English); the setting up of Caxton's printing press at Westminster in 1476; the full flowering of the Renaissance in the 16th century; the publishing of the King James Bible in 1611; the completion of Johnson's *Dictionary* of 1755; and the expansion to North America and South Africa in the 17th century and to India, Australia, and New Zealand in the 18th.

Old English. The Jutes, Angles, and Saxons lived in Jutland, Schleswig, and Holstein, respectively, before settling in Britain. According to the Venerable Bede, the first historian of the English people, the first Jutes, Hengist and Horsa, landed at Ebbsfleet in the Isle of Thanet in 449; and the Jutes later settled in Kent, southern Hampshire, and the Isle of Wight. The Saxons occupied the rest of England south of the Thames, as well as modern Middlesex and Essex. The Angles eventually took the remainder of England as far north as the Firth of Forth, including the future Edinburgh and the Scottish Lowlands. In both Latin and Common Germanic the Angles' name was Angli, later mutated in Old English to Engle (nominative) and Engla (genitive). "Engla land" designated the home of all three tribes collectively, and both King Alfred (known as Alfred the Great) and Abbot Aelfric, author and grammarian, subsequently referred to their speech as Englisc. Nevertheless, all the evidence indicates that Jutes, Angles, and Saxons retained their distinctive dialects.

Four Old English dialects

The River Humber was an important boundary, and the Anglian-speaking region developed two speech groups: to the north of the river, Northumbrian, and, to the south, Southumbrian, or Mercian. There were thus four dialects: Northumbrian, Mercian, West Saxon, and Kentish (see Figure 13). In the 8th century, Northumbrian led in literature and culture, but that leadership was destroyed by the Viking invaders, who sacked Lindisfarne, an island near the Northumbrian mainland, in 793. They landed in strength in 865. The first raiders were Danes, but they were later joined by Norwegians from Ireland and the Western Isles who settled in modern Cumberland, Westmorland, northwest Yorkshire, Lancashire, north Cheshire, and the Isle of Man. In the 9th century, as a result of the Norwegian invasions, cultural leadership passed from Northumbria to Wessex. During King Alfred's reign, in the last three decades of the 9th century, Winchester became the chief centre of learning. There the Parker Chronicle (a manuscript of the Anglo-Saxon Chronicle) was written; there the Latin works of the priest and historian Paulus Orosius, St. Augustine, St. Gregory, and the Venerable Bede were translated; and there the native poetry of Northumbria and Mercia was transcribed into the West Saxon dialect. This resulted in West Saxon's becoming "standard Old English"; and later, when Aelfric (*c.* 955–*c.* 1010) wrote his lucid and mature prose at Winchester, Cerne Abbas, and Eynsham, the hegemony of Wessex was strengthened.

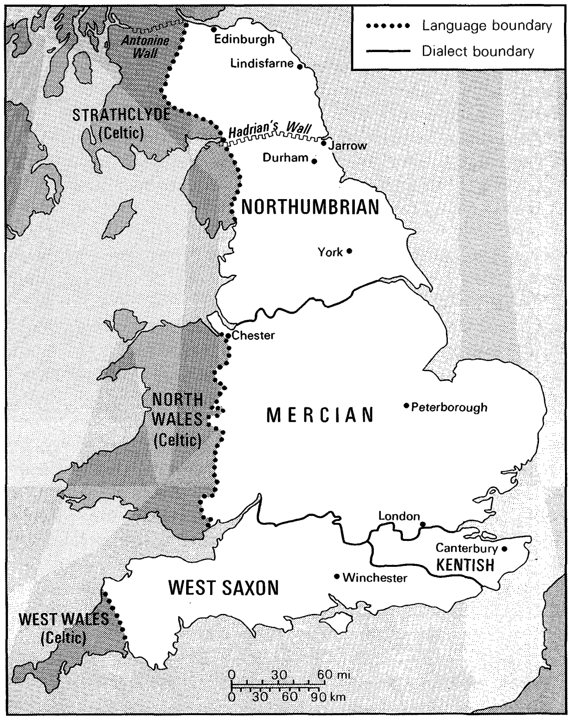


Figure 13: Old English dialects.

In standard Old English, adjectives were inflected as well as nouns, pronouns, and verbs. Nouns were inflected for four cases (nominative, genitive, dative, and accusative) in singular and plural. Five nouns of first kinship—*faeder, mōdor, brōthor, sweostor,* and *dohtor* ("father," "mother," "brother," "sister," and "daughter," respectively)—had their own set of inflections. There were 25 nouns such as *mon, men* ("man," "men") with mutated, or umlauted, stems. Adjectives had strong and weak declensions, the strong showing a mixture of noun and pronoun endings and the weak following the pattern of weak nouns. Personal, possessive, demonstrative, interrogative, indefinite, and relative pronouns had full inflections. The pronouns of the 1st and 2nd persons still had distinctive dual forms:

iċ	"I"	*wit*	"we two"	*wē*	"we"
thū (*þū*)	"thou"	*ġit*	"you two"	*ġē*	"you"

There were two demonstratives: *sē, sēo, thaet,* meaning "that," and *thes, thēos, this,* meaning "this," but no articles, the definite article being expressed by use of the demonstrative for "that" or not expressed at all. Thus, "the good man" was *sē gōda mon* or plain *gōd mon.* The function of the indefinite article was performed by the numeral *ān* "one" in *ān mon* "a man," by the adjective–pronoun *sum* in *sum mon* "a (certain) man," or not expressed, as in *thū eart gōd mon* "you are a good man."

Inflection of verbs

Verbs had two tenses only (present–future and past), three moods (indicative, subjunctive, and imperative), two numbers (singular and plural), and three persons (1st, 2nd, and 3rd). There were two classes of verb stems. (A verb stem is that part of a verb to which inflectional changes—changes indicating tense, mood, number, etc.—are added.) One type of verb stem, called vocalic because an internal vowel shows variations, is exemplified by the verb for "sing": *singan, singth, sang, sungon, gesungen.* The word for "deem" is an example of the other, called consonantal: *dēman, dēmth, dēmde, dēmdon, gedēmed.* Such verbs are called strong and weak, respectively.

All new verbs, whether derived from existing verbs or from nouns, belonged to the consonantal type. Some verbs of great frequency (antecedents of the modern words "be," "shall," "will," "do," "go," "can," "may," and so on) had their own peculiar patterns of inflections.

Grammatical gender persisted throughout the Old English period. Just as Germans now say *der Fuss, die Hand,* and *das Auge* (masculine, feminine, and neuter terms for "the foot," "the hand," and "the eye"), so, for these same structures, Aelfric said *sē fōt, sēo hond,* and *thaet ēaġe,* also masculine, feminine, and neuter. The three words for "woman," *wīfmon, cwene,* and *wīf,* were masculine, feminine, and neuter, respectively. *Hors* "horse," *sċēap* "sheep," and *maeġden* "maiden" were all neuter. *Eorthe* "earth" was feminine, but *lond* "land" was neuter. *Sunne* "sun" was feminine, but *mōna* "moon" was masculine. This simplification of grammatical gender resulted from the fact that the gender of Old English substantives was not always indicated by the ending but rather by the terminations of the adjectives and demonstrative pronouns used with the substantives. When these endings were lost, all outward marks of gender disappeared with them. Thus, the weakening of inflections and loss of gender occurred together. In the North, where inflections weakened earlier, the marks of gender likewise disappeared first. They survived in the South as late as the 14th century.

Because of the greater use of inflections in Old English, word order was freer than today. The sequence of subject, verb, and complement was normal, but when there were outer and inner complements the second was put in the dative case after *to: Sē biscop hālgode Ēadrēd tō cyninge* "The bishop consecreated Edred king." After an introductory adverb or adverbial phrase the verb generally took second place as in modern German: *Nū bydde iċ ān thing* "Now I ask [literally, "ask I"] one thing"; *Thȳ ilcan gēare gesette Aelfrēd cyning Lundenburg* "In that same year Alfred the king occupied London." Impersonal verbs had no subject expressed. Infinitives constructed with auxiliary verbs were placed at the ends of clauses or sentences: *Hīe ne dorston forth bī th ǣre ēa siglan* "They dared not sail beyond that river" (*siglan* is the infinitive); *Iċ wolde thās lytlan bōc āwendan* "I wanted to translate this little book" (*āwendan* is the infinitive). The verb usually came last in a dependent clause—*e.g., āwrītan wile* in *gif hwā thās bōc āwrītan wile* (*gerihte hē hīe be th ǣre bysene*) "If anyone wants to copy this book (let him correct his copy by the original)." Prepositions (or postpositions) frequently followed their objects. Negation was often repeated for emphasis.

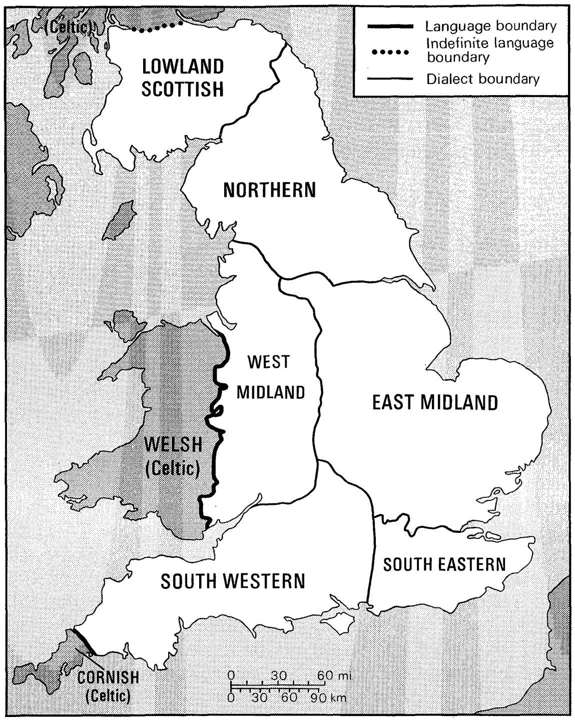


Figure 14: Middle English dialects.

Middle English. One result of the Norman Conquest of 1066 was to place all four Old English dialects more or less on a level. West Saxon lost its supremacy and the centre of culture and learning gradually shifted from Winchester to London. The old Northumbrian dialect became divided into Scottish and Northern, although little is known of either of these divisions before the end of the 13th century (Figure 14). The old Mercian dialect was split into East and West Midland. West Saxon became slightly diminished in area and was more appropriately named the South Western dialect. The Kentish dialect was considerably extended and was called South Eastern accordingly. All five Middle English dialects (Northern, West Midland, East Midland, South Western, and South Eastern) went their own ways and developed their own characteristics. The so-called Katherine Group of writings (1180–1210), associated with Hereford, a town not far from the Welsh border, adhered most closely to native traditions, and there is something to be said for regarding this West Midland dialect, least disturbed by French and Scandinavian intrusions, as a kind of Standard English in the High Middle Ages.

Another outcome of the Norman Conquest was to change the writing of English from the clear and easily readable insular hand of Irish origin to the delicate Carolingian script then in use on the Continent. With the change in appearance came a change in spelling. Norman scribes wrote Old English *y* as *u, ȳ* as *ui, ū* as *ou* (*ow* when final). Thus, *mycel* ("much") appeared as *muchel, fȳr* ("fire") as *fuir, hūs* ("house") as *hous,* and *hū* ("how") as *how.* For the sake of clarity (*i.e.,* legibility) *u* was often written *o* before and after *m, n, u, v,* and *w;* and *i* was sometimes written *y* before and after *m* and *n.* So *sunu* ("son") appeared as *sone* and *him* ("him") as *hym.* Old English *cw* was changed to *qu; hw* to *wh, qu,* or *quh; ċ* to *ch* or *tch; sċ* to *sh; -ċġ-* to *-gg-;* and *-ht* to *ght.* So Old English *cwēn* appeared as *queen; hwaet* as *what, quat,* or *quhat; dīċ* as *ditch; sċip* as *ship; secge* as *segge;* and *miht* as *might.*

Norman changes in spelling

For the first century after the Conquest, most loanwords came from Normandy and Picardy, but with the extension south to the Pyrenees of the Angevin empire of Henry II (reigned 1154–89), other dialects, especially Central French, or Francien, contributed to the speech of the aristocracy. As a result, Modern English acquired the forms canal, catch, leal, real, reward, wage, warden, and warrant from Norman French side by side with the corresponding forms channel, chase, loyal, royal, regard, gage, guardian, and guarantee, from Francien. King John lost Normandy in 1204. With the increasing power of the Capetian kings of Paris, Francien gradually predominated. Meanwhile, Latin stood intact as the language of learning. For three centuries, therefore, the literature of England was trilingual. *Ancrene Riwle,* for instance, a guide or rule (*riwle*) of rare quality for recluses or anchorites (*ancren*), was disseminated in all three languages.

The sounds of the native speech changed slowly. Even in late Old English short vowels had been lengthened before *ld, rd, mb,* and *nd,* and long vowels had been shortened before all other consonant groups and before double consonants. In early Middle English short vowels of whatever origin were lengthened in the open stressed syllables of disyllabic words. An open syllable is one ending in a vowel. Both syllables in Old English *nama* "name," *mete* "meat, food," *nosu* "nose," *wicu* "week," and *duru* "door" were short, and the first syllables, being stressed, were lengthened to *nāme, mēte, nōse, wēke,* and *dōre* in the 13th and 14th centuries. A similar change occurred in 4th-century Latin, in 13th-century German, and at different times in other languages. The popular notion has arisen that final mute *-e* in English makes a preceding vowel long; in fact, it is the lengthening of the vowel that has caused *e* to be lost in pronunciation. On the other hand, Old English long vowels were shortened in the first syllables of trisyllabic words, even when those syllables were open; *e.g., hāligdaeg* "holy day," *ǣrende* "message, errand," *crīstendōm* "Christianity," and *sūtherne* "southern," became *hŏliday* (Northern *hăliday*), *ĕrrende, chrĭstendom,* and *sŭtherne.* This principle still operates in current English. Compare, for example, trisyllabic derivatives such as the words chastity, criminal, fabulous, gradual, gravity, linear, national, ominous, sanity, and tabulate with the simple nouns and adjectives chaste, crime, fable, grade, grave, line, nation, omen, sane, and table.

Final mute *e*

There were significant variations in verb inflections in the Northern, Midland, and Southern dialects as shown in Table 26. The Northern infinitive was already one syllable (*sing* rather than the Old English *singan*), whereas the past participle *-en* inflection of Old English was strictly kept. These apparently contradictory features can be attributed entirely to Scandinavian, in which the final *-n* of the infinitive was lost early in *singa,* and the final *-n* of the past participle was doubled in *sunginn.* The Northern unmutated present participle in *-and* was also of Scandinavian origin. Old English mutated *-ende* (German *-end*) in the present participle had already become *-inde* in late West Saxon (Southern in Table 26), and it was this Southern *-inde* that blended with the *-ing* suffix (German *-ung*) of nouns of action that had already become near-gerunds in such compound nouns as *athswering* "oath swearing" and *writingfether* "writing feather, pen." This blending of present participle and gerund was further helped by the fact that Anglo-Norman and French *-ant* was itself a coalescence of Latin present participles in *-antem, -entem,* and Latin gerunds in *-andum, -endum.* The Northern second person singular *singis* was inherited unchanged from Common Germanic. The final *t* sound in Midland *-est* and Southern *-st* was excrescent, comparable with the final *t* in modern "amidst" and "amongst" from older *amiddes* and *amonges.* The Northern third person singular *singis* had a quite different origin. Like the *singis* of the plural, it resulted almost casually from an inadvertent retraction of the tongue in enunciation from an interdental *-th* sound to postdental *-s.* Today the form "singeth" survives as a

Table 26: Variations in Verb Inflections

	Northern	Midland	Southern
Infinitive	sing	singe(n)	singen
Present participle	singand	singende	singinde
Present singular			
1st person	singe	singe	singe
2nd person	singis	singes(t)	singst
3rd person	singis	singeth-es	singeth
Present plural	singis	singen	singeth
Past participle	sungen	(y)sunge(n)	ysunge

poetic archaism. Shakespeare used both *-eth* and *-s* endings ("It [mercy] blesseth him that gives and him that takes," *The Merchant of Venice*). The Midland present plural inflection *-en* was taken from the subjunctive. The past participle prefix *y-* developed from the Old English perfective prefix *ge-*.

Chaucer's language

Chaucer, who was born and died in London, spoke a dialect that was basically East Midland. Compared with his contemporaries, he was remarkably modern in his use of language. He was in his early 20s when the Statute of Pleading (1362) was passed, by the terms of which all court proceedings were henceforth to be conducted in English, though "enrolled in Latin." Chaucer himself used four languages; he read Latin (Classical and Medieval) and spoke French and Italian on his travels. For his own literary work he deliberately chose English.

Transition from Middle English to Early Modern English. The death of Chaucer at the close of the century (1400) marked the beginning of the period of transition from Middle English to the Early Modern English stage. The Early Modern English period is regarded by many scholars as beginning in about 1500 and terminating with the return of the monarchy (John Dryden's *Astraea Redux*) in 1660. The 15th century witnessed three outstanding developments: the rise of London English, the invention of printing, and the spread of the new learning.

The speech of London

Although the population of London in 1400 was only about 40,000, it was by far the largest city in England. York came second, followed by Bristol, Coventry, Plymouth, and Norwich. The Midlands and East Anglia, the most densely peopled parts of England, supplied London with streams of young immigrants. The speech of the capital was mixed, and it was changing. The seven long vowels of Chaucer's speech had already begun to shift. Incipient diphthongization of high front /i:/ (the *ee* sound in "meet") and high back /u:/ (as in "fool") led to instability in the other five long vowels. (Symbols within slash marks are taken from the International Phonetic Alphabet.) This remarkable event, known as the Great Vowel Shift, changed the whole vowel system of London English. As /i:/ and /u:/ became diphthongized to /ai/ (as in "bide") and /au/ (as in "house") respectively, so the next highest vowels, /e:/ (this sound can be heard in the first part of the diphthong in "name") and /o:/ (a sound that can be heard in the first part of the diphthong in "home"), moved up to take their places, and so on. The whole process is summarized in Table 27. Every one of the sounds appearing in this table can still be heard somewhere in living English dialects.

Table 27: Vowel Shifts in London English

Chaucer's spelling	Chaucer's pronunciation*	Shakespeare's pronunciation	present pronunciation*	present spelling
lyf	li:f	leif	laif	life
deed	de:d	di:d	di:d	deed
deel	dɛ:l	de:l	di:l	deal
name	na:mə	nɛ:m	neim	name
hoom	hɔ:m	ho:m	houm	home
mone	mo:n	mu:n	mu:n	moon
hous	hu:s	hous	haus	house

*Expressed in the International Phonetic Alphabet.

When Caxton started printing at Westminster in the late summer of 1476, he was painfully aware of the uncertain state of the English language. In his prologues and epilogues to his translations he made some revealing observations on the problems that he had encountered as translator and editor. At this time, sentence structures were being gradually modified, but many remained untidy. For the first time, nonprofessional scribes, including women, were writing at length.

The revival of classical learning was one aspect of that Renaissance, or spiritual rebirth, that arose in Italy and spread to France and England. It evoked a new interest in Greek on the part of learned men such as William Grocyn and Thomas Linacre, Sir Thomas More and Desiderius Erasmus. John Colet, dean of St. Paul's in the first quarter of the 16th century, startled his congregation by expounding the Pauline Epistles of the New Testament as living letters. The deans who had preceded him had known no Greek, because they had found in Latin all that they required. Only a few medieval churchmen, such as Robert Grosseteste, bishop of Lincoln, and the Franciscan Roger Bacon could read Greek with ease. The names of the seven liberal arts of the medieval curricula (the trivium and the quadrivium), it is true, were all Greek—grammar, logic, and rhetoric; arithmetic, geometry, astronomy, and music—but they had come into English by way of French.

Renaissance scholars adopted a liberal attitude to language. They borrowed Latin words through French, or Latin words direct; Greek words through Latin, or Greek words direct. Latin was no longer limited to Church Latin: it embraced all Classical Latin. For a time the whole Latin lexicon became potentially English. Some words, such as consolation and infidel, could have come from either French or Latin. Others, such as the terms abacus, arbitrator, explicit, finis, gratis, imprimis, item, memento, memorandum, neuter, simile, and videlicet, were taken straight from Latin. Words that had already entered the language through French were now borrowed again, so that doublets arose: benison and benediction; blame and blaspheme; chance and cadence; count and compute; dainty and dignity; frail and fragile; poor and pauper; purvey and provide; ray and radius; sever and separate; strait and strict; sure and secure. The Latin adjectives for "kingly" and "lawful" have even given rise to triplets; in the forms real, royal, and regal and leal, loyal, and legal, they were imported first from Anglo-Norman, then from Old French, and last from Latin direct.

After the dawn of the 16th century, English prose moved swiftly toward modernity. In 1525 Lord Berners completed his translation of Jean Froissart's *Chronicle,* and William Tyndale translated the New Testament. One-third of the King James Bible (1611), it has been computed, is worded exactly as Tyndale left it; and between 1525 and 1611 lay the Tudor Golden Age, with its culmination in Shakespeare. Too many writers, to be sure, used "inkhorn terms," newly-coined, ephemeral words, and too many vacillated between Latin and English. Sir Thomas More actually wrote his *Utopia* in Latin. It was translated into French during his lifetime but not into English until 1551, some years after his death. Francis Bacon published *De dignitate et augmentis scientiarum (On the Dignity and Advancement of Learning,* an expansion of his earlier *Advancement of Learning)* in Latin in 1623. William Harvey announced his epoch-making discovery of the circulation of the blood in his Latin *De Motu Cordis et Sanguinis in Animalibus* (1628; *On the Motion of the Heart and Blood in Animals*). John Milton composed polemical treatises in the language of Cicero. As Oliver Cromwell's secretary, he corresponded in Latin with foreign states. His younger contemporary Sir Isaac Newton lived long enough to bridge the gap. He wrote his *Principia* (1687) in Latin but his *Opticks* (1704) in English.

Plans to regulate English

Restoration period. With the restoration of the monarchy in 1660, men again looked to France. John Dryden admired the Académie Française and greatly deplored that the English had "not so much as a tolerable dictionary, or a grammar; so that our language is in a manner barbarous" as compared with elegant French. After the passionate controversies of the Civil War, this was an age of cool scientific nationalism. In 1662 the Royal Society of London for the Promotion of Natural Knowledge received its charter. Its first members, much concerned with language, appointed a committee of 22 "to improve the English tongue particularly for philosophic purposes." It included Dryden, the diarist John Evelyn, Bishop Thomas Sprat, and the poet Edmund Waller. Sprat pleaded for "a close, naked, natural way of speaking; positive expressions; clear senses, a native easiness; bringing all things as near the mathematical plainness" as possible. The committee, however, achieved no tangible result, and failed in its attempt to found an authoritative arbiter over the English tongue. A second attempt was made in 1712, when Jonathan Swift addressed an open letter to Robert Harley, earl of Oxford, then Lord Treasurer, making "A Proposal for Correcting, Improving, and Ascertaining [fixing] the English Tongue."

This letter received some popular support, but its aims were frustrated by a turn in political fortunes. Queen Anne died in 1714. The Earl of Oxford and his fellow Tories, including Swift, lost power. No organized attempt to found a language academy on French lines has ever been made since.

With Dryden and Swift the English language reached its full maturity. Their failure to found an academy was partly counterbalanced by Samuel Johnson in his *Dictionary* (published in 1755) and by Robert Lowth in his *Grammar* (published in 1761).

Johnson's *Dictionary*

Age of Johnson. In the making of his *Dictionary,* Johnson took the best conversation of contemporary London and the normal usage of reputable writers after Sir Philip Sidney (1554–86) as his criteria. He exemplified the meanings of words by illustrative quotations. Johnson admitted that "he had flattered himself for a while" with "the prospect of fixing our language" but that thereby "he had indulged expectation which neither reason nor experience could justify." The two-folio work of 1755 was followed in 1756 by a shortened, one-volume version that was widely used far into the 20th century. Revised and enlarged editions of the unabbreviated version were made by Archdeacon Henry John Todd in 1818 and by Robert Gordon Latham in 1866.

It was unfortunate that Joseph Priestley, Robert Lowth, James Buchanan, and other 18th-century grammarians (Priestley was perhaps better known as a scientist and theologian) took a narrower view than Johnson on linguistic growth and development. They spent too much time condemning such current "improprieties" as "I had rather not," "you better go," "between you and I," "it is me," "who is this for?", "between four walls," "a third alternative," "the largest of the two," "more perfect," and "quite unique." Without explanatory comment they banned "you was" outright, although it was in widespread use among educated people (on that ground it was later defended by Noah Webster). "You was" had, in fact, taken the place of both "thou wast" and "thou wert" as a useful singular equivalent of the accepted plural "you were."

As the century wore on, grammarians became more numerous and aggressive. They set themselves up as arbiters of correct usage. They compiled manuals that were not only descriptive (stating what people do say) and prescriptive (stating what they should say) but also proscriptive (stating what they should not say). They regarded Latin as a language superior to English and claimed that Latin embodied universally valid canons of logic. This view was well maintained by Lindley Murray, a native of Pennsylvania who settled in England in the very year (1784) of Johnson's death. Murray's *English Grammar* appeared in 1795, became immensely popular, and went into numerous editions. It was followed by an *English Reader* (1799) and an *English Spelling Book* (1804), long favourite textbooks in both Old and New England.

Proposals for a new dictionary

19th and 20th centuries. In 1857 Richard Chenevix Trench, dean of St. Paul's, lectured to the Philological Society on the theme, "On some Deficiencies in our English Dictionaries." His proposals for a new dictionary were implemented in 1859, when Samuel Taylor Coleridge's grandnephew, Herbert Coleridge, set to work as first editor. He was succeeded by a lawyer named Frederick James Furnivall, who in 1864 founded the Early English Text Society with a view to making all the earlier literature available to historical lexicographers in competent editions. Furnivall was subsequently succeeded as editor by James A.H. Murray, who published the first fascicle of *A New English Dictionary on Historical Principles* in 1884. Later Murray was joined successively by three editors: Henry Bradley, William Alexander Craigie, and Charles Talbut Onions. Aside from its *Supplements,* the dictionary itself fills 12 volumes, has over 15,000 pages, and contains 414,825 words, illustrated by 1,827,306 citations. It is a dictionary of the British Commonwealth and the United States, a fact symbolized by the presentation of first copies in the spring of 1928 to King George V and Pres. Calvin Coolidge. It exhibits the histories and meanings of all words known to have been in use since 1150. From 1150 to 1500 all five Middle English dialects, as has been seen, were of equal status. They are therefore all included. After 1500, however, dialectal expressions are not admitted, nor are scientific and technical terms not in general use. Otherwise, the written vocabulary is comprehensive. A revised edition of this dictionary, known as *The Oxford English Dictionary,* was published in 1933.

VARIETIES OF ENGLISH

British English. The abbreviation RP (Received Pronunciation) denotes the speech of educated people living in London and the southeast of England and of other people elsewhere who speak in this way. If the qualifier educated be assumed, RP is then a regional (geographical) dialect, as contrasted with London Cockney, which is a class (social) dialect. RP is not intrinsically superior to other varieties of English; it is itself only one particular regional dialect that has, through the accidents of history, achieved more extensive use than others. Although acquiring its unique status without the aid of any established authority, it may have been fostered by the public schools (Winchester, Eton, Harrow, Rugby, and so on) and the ancient universities (Oxford and Cambridge). Other varieties of English are well preserved in spite of the levelling influences of film, television, and radio. In the Northern dialect RP /aː/ (the first vowel sound in "father") is still pronounced /æ/ (a sound like the *a* in "fat") in words such as laugh, fast, and path; this pronunciation has been carried across the Atlantic into American English.

Features of the Northern dialect

In the words run, rung, and tongue, the received-standard pronunciation of the vowel is /∧/, like the *u* in "but"; in the Northern dialect it is /u/, like the *oo* in "book." In the words bind, find, and grind, the received standard pronunciation of the vowel sound is /ai/, like that in "bide"; in Northern, it is /i/, like the sound in "feet." The vowel sound in the words go, home, and know in the Northern dialect is /ɔː/, approximately the sound in "law" in some American English dialects. In parts of Northumberland, RP "it" is still pronounced "hit," as in Old English. In various Northern dialects the definite article "the" is heard as *t, th,* or *d.* In those dialects in which it becomes both *t* and *th, t* is used before consonants and *th* before vowels. Thus, one hears "t'book" but "th'apple." When, however, the definite article is reduced to *t* and the following word begins with *t* or *d,* as in "t'tail" or "t'dog," it is replaced by a slight pause as in the RP articulation of the first *t* in "hat trick." The RP /tʃ/, the sound of the *ch* in "church," becomes *k,* as in "thack," ("thatch, roof") and "kirk" ("church"). In many Northern dialects strong verbs retain the old past-tense singular forms band, brak, fand, spak for RP forms bound, broke, found, and spoke. Strong verbs also retain the past participle inflection *-en* as in "comen," "shutten," "sitten," and "getten" or "gotten" for RP "come," "shut," "sat," and "got."

In some Midland dialects the diphthongs in "throat" and "stone" have been kept apart, whereas in RP they have fallen together. In Cheshire, Derby, Stafford, and Warwick, RP "singing" is pronounced with a *g* sounded after the velar nasal sound (as in RP "finger"). In Norfolk one hears "skellington" and "solintary" for "skeleton" and "solitary," showing an intrusive *n* just as does "messenger" in RP from French *messager,* "passenger" from French *passager,* and "nightingale" from Old English *nihtegala.* Other East Anglian words show consonantal metathesis (switch position), as in "singify," and substitution of one liquid or nasal for another, as in "chimbly" for "chimney," and "synnable" for "syllable." "Hantle" for "handful" shows syncope (disappearance) of an unstressed vowel, partial assimilation of *d* to *t* before voiceless *f,* and subsequent loss of *f* in a triple consonant group.

South Western dialects

In South Western dialects, initial *f* and *s* are often voiced, becoming *v* and *z.* Two words with initial *v* have found their way into RP: "vat" from "fat" and "vixen" from "fixen" (female fox). Another South Western feature is the development of a *d* between *l* or *n* and *r,* as in "parlder" for "parlour" and "carnder" for "corner." The bilabial semivowel *w* has developed before *o* in "wold" for "old," and in "wom" for "home," illustrating a similar development in RP by which Old English *ān* has become "one," and Old English *hāl* has come to be spelled "whole," as compared with Northern *hale.* In South Western dialects "yat" comes from the old singular *geat,* whereas RP "gate" comes from the plural *gatu.* Likewise, "clee" comes from the old nominative *clea,* whereas RP "claw" comes from the oblique cases. The verbs keel and kemb have developed regularly from Old English *cēlan* "to make cool" and *kemban* "to use a comb," whereas the corresponding RP verbs cool and comb come from the adjective and the noun, respectively.

In Wales, people often speak a clear and measured form of English with a musical intonation inherited from ancestral Celtic. They tend to aspirate both plosives (stops) and fricative consonants very forcibly; thus, "true" is pronounced with an audible puff of breath after the initial *t.*

Lowland Scottish was once a part of Northern English, but two dialects began to diverge in the 14th century. Today Lowland Scots trill their *r*'s, shorten vowels, and simplify diphthongs. A few Scottish words, such as bairn, brae, canny, dour, and pawky, have made their way into RP. Lowland Scottish is not to be confused with Scottish Gaelic, a Celtic language still spoken by about 90,700 people (almost all bilingual) mostly in the Highlands and the Western Isles. Thanks to Robert Burns and Sir Walter Scott, many Scottish Gaelic words have been preserved in English literature.

Irish pronunciation

Northern Ireland has dialects related in part to Lowland Scottish and in part to the southern Irish dialect of English. Irish pronunciation is conservative and is clearer and more easily intelligible than many other dialects. Its literature has reached worldwide audiences, whether written by Englishmen born in Ireland, such as Jonathan Swift, Laurence Sterne, Oliver Goldsmith, Sir Richard Steele, Edmund Burke, Oscar Wilde, and George Bernard Shaw, or by authentic Irish, such as James Joyce, William Butler Yeats, and John Millington Synge. The influence of Irish Gaelic on the speech of Dublin is most evident in the syntax of drama and in the survival of such picturesque expressions as "We are after finishing," "It's sorry you will be," and "James do be cutting corn every day."

American and Canadian English. The dialect regions of the United States are most clearly marked along the Atlantic littoral, where the earlier settlements were made. Three dialects can be defined: Northern, Midland, and Southern. Each has its subdialects.

The Northern dialect is spoken in New England. Its six chief subdialects comprise northeastern New England (Maine, New Hampshire, and eastern Vermont), southeastern New England (eastern Massachusetts, eastern Connecticut, and Rhode Island), southwestern New England (western Massachusetts and western Connecticut), the inland north (western Vermont and upstate New York), the Hudson Valley, and metropolitan New York (Figure 15).

The Midland dialect is spoken in the coastal region from Point Pleasant, in New Jersey, to Dover, in Delaware. Its seven major subdialects comprise the Delaware Valley, the Susquehanna Valley, the Upper Ohio Valley, northern West Virginia, the Upper Potomac and Shenandoah, southern West Virginia and eastern Kentucky, western Carolina, and eastern Tennessee.

The Southern dialect area covers the coastal region from Delaware to South Carolina. Its five chief subdialects comprise the Delmarva Peninsula, the Virginia Piedmont, northeastern North Carolina (Albemarle Sound and Neuse Valley), Cape Fear and Pee Dee valleys, and the South Carolina Low Country, around Charleston.

These boundaries, based on those of the *Linguistic Atlas of the United States and Canada,* are highly tentative. To some extent these regions preserve the traditional speech of southeastern and southern England, where most of the early colonists were born. The first settlers who came to Virginia (1607) and Massachusetts (1620) soon learned to adapt old words to new uses, but they were content to borrow names from the local Indian languages for unknown trees, such as hickory and persimmon, and for unfamiliar animals, such as raccoons and woodchucks. Later they took words from foreign settlers: "chowder" and "prairie" from the French, "scow" and "sleigh" from the Dutch. They made new compounds, such as "backwoods" and "bullfrog," and gave new meanings to such words as "lumber" (which in British English denotes disused furniture,

Prepared by Mrs. Raven I. McDavid, Jr., for *The Structure of American English,* by W. Nelson Francis; copyright © 1958, The Ronald Press Company, New York.
Atlantic Seaboard dialect areas after Hans Kurath, *A Word Geography of the Eastern United States;* copyright © 1949, University of Michigan Press, Ann Arbor

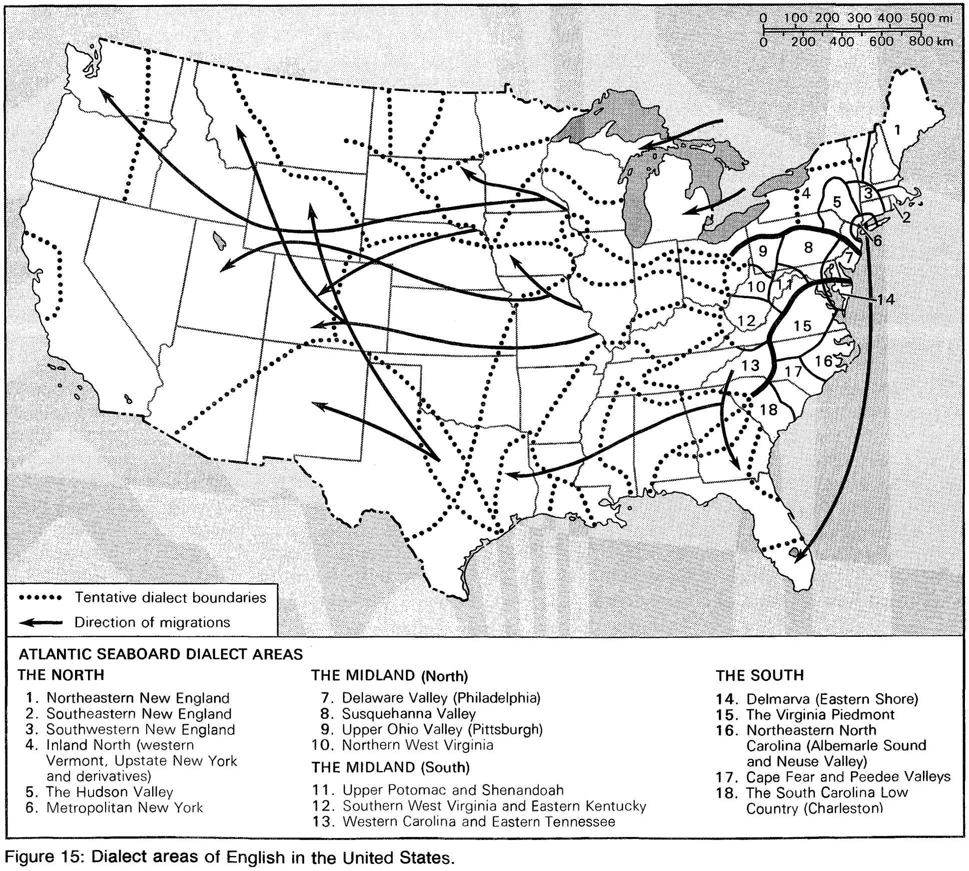

0 100 200 300 400 500 mi
0 200 400 600 800 km
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
•••••• Tentative dialect boundaries
◄—— Direction of migrations
ATLANTIC SEABOARD DIALECT AREAS
THE NORTH
1. Northeastern New England
2. Southeastern New England
3. Southwestern New England
4. Inland North (western Vermont, Upstate New York and derivatives)
5. The Hudson Valley
6. Metropolitan New York
THE MIDLAND (North)
7. Delaware Valley (Philadelphia)
8. Susquehanna Valley
9. Upper Ohio Valley (Pittsburgh)
10. Northern West Virginia
THE MIDLAND (South)
11. Upper Potomac and Shenandoah
12. Southern West Virginia and Eastern Kentucky
13. Western Carolina and Eastern Tennessee
THE SOUTH
14. Delmarva (Eastern Shore)
15. The Virginia Piedmont
16. Northeastern North Carolina (Albemarle Sound and Neuse Valley)
17. Cape Fear and Peedee Valleys
18. The South Carolina Low Country (Charleston)

Figure 15: Dialect areas of English in the United States.

or junk) and "corn" (which in British English signifies any grain, especially wheat).

The Gullah dialect

Before the Declaration of Independence (1776), two-thirds of the immigrants had come from England, but after that date they arrived in large numbers from Ireland. The potato famine of 1845 drove 1,500,000 Irish to seek homes in the New World, and the European revolutions of 1848 drove as many Germans to settle in Pennsylvania and the Middle West. After the close of the American Civil War, millions of Scandinavians, Slavs, and Italians crossed the ocean and eventually settled mostly in the North Central and Upper Midwest states. In some areas of South Carolina and Georgia the American Negroes who had been imported to work the rice and cotton plantations developed a contact language called Gullah, or Geechee, that made use of many structural and lexical features of their native languages. This remarkable variety of English is comparable to such "contact languages" as Sranan (Taki-Taki) and Melanesian Pidgin. The speech of the Atlantic Seaboard shows far greater differences in pronunciation, grammar, and vocabulary than that of any area in the North Central States, the Upper Midwest, the Rocky Mountains, or the Pacific Coast. Today, urbanization, quick transport, and television have tended to level out some dialectal differences in the United States.

The boundary with Canada nowhere corresponds to any boundary between dialects, and the influence of United States English is strong, being felt least in the Maritime Provinces and Newfoundland. Nevertheless, in spite of the effect of this proximity to the United States, British influences are still potent in some of the larger cities; Scottish influences are well sustained in Ontario. Canada remains bilingual. One-fourth of its people, living mostly in the province of Quebec, have French as their mother tongue. Those provinces in which French is spoken as a mother tongue by 10 percent or more of the population are described as "federal bilingual districts" in the Official Languages Bill of 1968.

Australian and New Zealand English. Unlike Canada, Australia has few speakers of European languages other than English within its borders. There are still many Aboriginal languages, though they are spoken by only a few hundred speakers each and their continued existence is threatened. More than 80 percent of the population is British. By the mid-20th century, with rapid decline of its Aboriginal tongues, English was without rivals in Australia.

During colonial times the new settlers had to find names for a fauna and flora (*e.g.,* banksia, iron bark, whee whee) different from anything previously known to them: trees that shed bark instead of leaves and cherries with external stones. The words brush, bush, creek, paddock, and scrub acquired wider senses, whereas the terms brook, dale, field, forest, and meadow were seldom used. A creek leading out of a river and entering it again downstream was called an anastomizing branch (a term from anatomy), or an anabranch, whereas a creek coming to a dead end was called by its native name, a billabong. The giant kingfisher with its raucous bray was long referred to as a laughing jackass, later as a bushman's clock, but now it is a kookaburra. Cattle so intractable that only roping could control them were said to be ropable, a term now used as a synonym for "angry" or "extremely annoyed."

"Rouse-about" and "roust-about"

A deadbeat was a penniless "sundowner" at the very end of his tether, and a no-hoper was an incompetent fellow, hopeless and helpless. An offsider (strictly, the offside driver of a bullock team) was any assistant or partner. A rouseabout was first an odd-job man on a sheep station and then any kind of handyman. He was, in fact, the "down-under" counterpart of the wharf labourer, or roustabout, on the Mississippi River. Both words originated in Cornwall, and many other terms, now exclusively Australian, came ultimately from British dialects. "Dinkum," for instance, meaning "true, authentic, genuine," echoed the "fair dinkum," or fair deal, of Lincolnshire dialect. "Fossicking" about for surface gold, and then rummaging about in general, perpetuated the term fossick ("to elicit information, ferret out the facts") from the Cornish dialect of English. To "barrack," or jeer noisily, recalled Irish "barrack" ("to brag, boast"), whereas "skerrick" in the phrase "not a skerrick left" was obviously identical with the "skerrick" meaning "small fragment, particle," still heard in English dialects from Westmorland to Hampshire.

Some Australian English terms came from Aboriginal speech: the words boomerang, corroboree (warlike dance and then any large and noisy gathering), dingo (reddish-brown half-domesticated dog), galah (cockatoo), gunyah (bush hut), kangaroo, karri (dark-red eucalyptus tree), nonda (rosaceous tree yielding edible fruit), pokutukawa (evergreen bearing brilliant blossom), wallaby (small marsupial), and wallaroo (large rock kangaroo). Australian English has slower rhythms and flatter intonations than RP. Although there is remarkably little regional variation throughout the entire continent, there is significant social variation. The neutral vowel /ə/ (as the *a* in "sofa") is frequently used, as in London Cockney: "arches" and "archers" are both pronounced [*a*:tʃəz], and the pronunciations of RP "day" and "go" are, respectively, [dəi] and [gəu].

Although New Zealand lies over 1,000 miles away, much of the English spoken there is similar to that of Australia. The blanket term Austral English is sometimes used to cover the language of the whole of Australasia, or Southern Asia, but this term is far from popular with New Zealanders because it makes no reference to New Zealand and gives all the prominence, so they feel, to Australia. Between North and South Islands there are observable differences. For one thing, Maori, which is still a living language (related to Tahitian, Hawaiian, and the other Austronesian [Malayo-Polynesian] languages), has a greater number of speakers and more influence in North Island.

English an alternative official language

The English of India–Pakistan. In 1950 India became a federal republic within the Commonwealth of Nations, and Hindi was declared the first national language. English, it was stated, would "continue to be used for all official purposes until 1965." In 1967, however, by the terms of the English Language Amendment Bill, English was proclaimed "an alternative official or associate language with Hindi until such time as all non-Hindi states had agreed to its being dropped." English is therefore acknowledged to be indispensable. It is the only practicable means of day-to-day communication between the central government at New Delhi and states with non-Hindi speaking populations, especially with the Deccan, or "South," where millions speak Dravidian (non-Indo-European) languages—Telugu, Tamil, Kannada, and Malayalam. English is widely used in business, and, although its use as a medium in higher education is decreasing, it remains the principal language of scientific research.

In 1956 Pakistan became an autonomous republic comprising two states, East and West. Bengali and Urdu were made the national languages of East and West Pakistan, respectively, but English was adopted as a third official language and functioned as the medium of interstate communication. (In 1971 East Pakistan broke away from its western partner and became the independent state of Bangladesh.)

African English. Africa is the most multilingual area in the world, if people are measured against languages. Upon a large number of indigenous languages rests a slowly changing superstructure of world languages (Arabic, English, French, and Portuguese). The problems of language are everywhere linked with political, social, economic, and educational factors.

The Republic of South Africa, the oldest British settlement in the continent, resembles Canada in having two recognized European languages within its borders: English and Afrikaans, or Cape Dutch. Both British and Dutch traders followed in the wake of 15th-century Portuguese explorers and have lived in widely varying war-and-peace relationships ever since. Although the Union of South Africa, comprising Cape Province, Transvaal, Natal, and Orange Free State, was for more than a half century (1910–61) a member of the British Empire and Commonwealth, its four prime ministers (Botha, Smuts, Hertzog, and Malan) were all Dutchmen. In the early 1980s Afrikaners outnumbered Britishers by three to two. The Afrikaans language began to diverge seriously from European Dutch in the late 18th century and has gradually come to be recognized as a separate language. Although the English spoken in South Africa differs in some respects from standard British English, its speakers do not regard the language as a separate one. They have naturally come to use many Afrikanerisms, such as *kloof, kopje, krans, veld,* and *vlei,* to denote features of the landscape and occasionally employ African names to designate local animals and plants. The words trek and commando, notorious in South African history, have acquired almost worldwide currency.

Afrikanerisms in South African English

Elsewhere in Africa, English helps to answer the needs of wider communication. It functions as an official language of administration in Botswana, Lesotho, and Swaziland and in Zimbabwe, Zambia, Malaŵi, Uganda, and Kenya. It is the language of instruction at Makerere University in Kampala, Uganda; at the University of Nairobi, Kenya; and at the University of Dar es Salaam, in Tanzania.

The West African states of The Gambia, Sierra Leone, Ghana, and Nigeria, independent members of the Commonwealth, have English as their official language. They are all multilingual. The official language of Liberia is also English, although its tribal communities constitute four different linguistic groups. Its leading citizens regard themselves as Americo-Liberians, being descendants of those freed blacks whose first contingents arrived in West Africa in 1822. South of the Sahara indigenous languages are extending their domains and are competing healthily and vigorously with French and English.

THE FUTURE OF ENGLISH

Geographically, English is the most widespread language on Earth, and it is second only to Mandarin Chinese in the number of people who speak it. The International Telecommunication Union (ITU) has five official languages: English, French, Spanish, Russian, and Chinese. The influence of these languages upon one another will inevitably increase.

It is reasonable to ask if changes in English can be predicted. There will doubtless be modifications in pronunciation, especially in that of long vowels and diphthongs. In weakly stressed syllables there is already a discernible tendency, operating effectively through radio and television, to restore the full qualities of vowels in these syllables. This tendency may bring British English more into line with American English and may bring them both a little nearer to Spanish and Italian. Further, it may help to narrow the gap between pronunciation and spelling. Other factors will also contribute toward the narrowing of this gap: advanced technological education, computer programming, machine translation, and expanding mass media. Spelling reformers will arise from time to time to liven up proceedings, but, in general, traditional orthography may well hold its own against all comers, perhaps with some regularization. Printing houses, wielding concentrated power through their style directives, will surely find it in their best interests to agree on uniformity of spelling. Encyclopaedic dictionaries—computerized, universal, and subject to continuous revision—may not go on indefinitely recording such variant spellings as "connection" and "connexion," "judgment" and "judgement," "labor" and "labour," "medieval" and "mediaeval," "plow" and "plough," "realise" and "realize," "thru" and "through."

Factors leading toward conformity

Since Tudor days, aside from the verb endings *-est* and *-eth,* inflections have remained stable because they represent the essential minimum. The abandonment of the forms thou and thee may encourage the spread of yous and youse in many areas, but it is not necessarily certain that these forms will win general acceptance. The need for a distinctive plural can be supplied in other ways (*e.g.,* the forms "you all, you fellows, you people"). The distinctions between the words "I" and "me," "he" and "him," "she" and "her," "we" and "us," "they" and "them" seem to many authors to be too important to be set aside, in spite of a growing tendency to use objective forms as emphatic subjective pronouns and to say, for instance, "them and us" instead of "they and we" in contrasting social classes. Otherwise, these distinctive forms may remain stable; they are all monosyllabic, they are in daily use, and they can bear the main stress. Thus they are likely to resist levelling processes.

Considerable changes will continue to be made in the forms and functions of auxiliary verbs, catenative (linking) verbs, phrasal verbs, and verb phrases. Indeed, the constituents of verbs and verb groups are being more subtly modified than those of any other word class. By means of auxiliaries and participles, a highly intricate system of aspects, tenses, and modalities is gradually evolving.

In syntax the movement toward a stricter word order seems to many to be certain to continue. The extension of multiple attributives in nominal groups has probably reached its maximum. It cannot extend further without incurring the risk of ambiguity.

In vocabulary further increases are expected if the present trends continue. Unabbreviated general dictionaries already contain 500,000 entries, but even larger dictionaries, with 750,000 entries, may be required. Coiners of words probably will not confine themselves to Greek and Latin in creating new terms, they are likely to exercise their inventive powers in developing an international technical vocabulary that is increasingly shared by Russian, French, and Spanish and that is slowly emerging as the universal scientific language. (S.P.)

Armenian language

The Armenian language, which forms a separate branch of the western group of Indo-European languages, is the mother tongue of the Turkish Armenians and of the Armenians in the Armenian Soviet Socialist Republic, where it is spoken by 2,850,000 people. In other parts of the Soviet Union, especially in the neighbouring republics of Georgia and Azerbaijan, it is used by some 1,300,000. Armenian emigrants and refugees have taken their language with them all over Asia Minor and the Middle East and from there to many European countries, especially Romania, Poland, and France, and to America, particularly the United States. In all, Armenian is probably spoken by about 5,500,000 people.

History of the language. Armenian was introduced into the mountainous Transcaucasian region (called Greater Armenia by the Greek historians) by invaders coming from the northern Balkans, probably in the latter part of the 2nd millennium BC. These invaders occupied the region on the shores of Lake Van that had previously been the site of the ancient Urartean kingdom. By the 7th century BC the Armenian language seems to have re-

Table 28: The Armenian Alphabet

letter		equivalent	letter		equivalent
capital	lowercase		capital	lowercase	
Ա	ա	a	Մ	մ	m
Բ	բ	b	Յ	յ	y
Գ	գ	g	Ն	ն	n
Դ	դ	d	Շ	շ	sh
Ե	ե	e	Ո	ո	o
Զ	զ	z	Չ	չ	ch‘*
Է	է	ē	Պ	պ	p
Ը	ը	ĕ	Ջ	ջ	j
Թ	թ	t‘*	Ռ	ռ	rh
Ժ	ժ	zh	Ս	ս	s
Ի	ի	i	Վ	վ	v
Լ	լ	l	Տ	տ	t
Խ	խ	kh	Ր	ր	r
Ծ	ծ	ts	Ց	ց	ts‘*
Կ	կ	k	Ւ	ւ	w
Հ	հ	h	Փ	փ	p‘*
Ձ	ձ	dz	Ք	ք	k‘*
Ղ	ղ	gh	Օ	օ	ō
Ճ	ճ	ch	Ֆ	ֆ	f

*The *spiritus asper*, ‘, indicates aspiration.

placed the tongues of the native population. It is tempting to connect the invasion with the downfall of the Hittite empire in Anatolia.

Development of the alphabet

After the introduction of Christianity in the beginning of the 4th century AD, the language was reduced to writing; the alphabet, of 38 letters, was invented, according to tradition, by the bishop Mashtots (Mesrob) in about AD 400 (Table 28). Admirably suited to the phonology of Armenian, it is still used in various forms by Armenians all over the world. The oldest writings in the language date from the 5th century; they are preserved in manuscript form from the 9th century. Grabar, the written language of the 5th century, the golden age of Armenian culture, is traditionally said to be based on the dialect of Tarawn on Lake Van. To what extent the spoken language was split into dialects at that time is not known. The language of the literature from the 5th to the 8th century is remarkably homogeneous, but by the 9th century the influence of the spoken dialects was noticeable, especially in legal and historical texts. Among the Middle Armenian varieties of Grabar, the best known is the 12th- and 13th-century chancellery (court) language of the Armenian kingdom in Cilicia. More or less corrupted versions of Grabar continued as the literary language until the middle of the 19th century.

Modern East Armenian and West Armenian. In the 1800s, the writers Khachatur Abovean (1805–48) and Mikael Nalbandean (1829–66) and other Armenian nationalists made efforts to reach the populace with nationalist propaganda. As a result, a national revival occurred from which a new literary language emerged that was much closer to the spoken language. This is known in two varieties. East Armenian, now the official language of the Armenian Soviet Socialist Republic, is based on the dialect of the Ararat valley and the city of Yerevan; West Armenian has its foundation in the dialect of Istanbul. East Armenian is also spoken in other parts of the Soviet Union, whereas the western variety dominates in the Armenian colonies in the Middle East, Asia Minor, Europe, and America. The differences between these two written forms of Modern Armenian are slight, constituting no barrier to mutual intelligibility.

Dialects

In addition to the two literary languages, there are a great number of dialects, some of which are so different that the speakers cannot understand each other. It is estimated that before World War I some 50 distinct dialects were spoken. Today, the spoken dialects are losing ground in the Soviet Union, under the pressure of the standard written language. Accurate statistics on the extent to which Armenian dialects are used in Turkey and in other parts of the Middle East are not available.

When the scientific study of Armenian started in the 19th century, the language was considered an Iranian dialect, a mistake easily explained by the vast number of Iranian loanwords in the vocabulary. Subsequent studies, however, have convincingly shown Armenian to be an independent member of the Indo-European language family. According to the Greek historian Herodotus, Armenian was a variety of Phrygian, a tongue presumed to be Indo-European. What little is known of the latter is insufficient to support or confirm such a claim.

Sounds of Armenian

Phonology. Phonetic developments in Armenian have radically changed the sound system of the old Indo-European parent language. In particular, the pattern of the plosive consonants—the stops—has been completely reshuffled. In the more conservative central Armenian dialects three series of stops are distinguished (voiced *b, d, g,* which in some dialects are aspirated; unvoiced *p, t, k;* and unvoiced aspirated *ph, th, kh*). In the dialects of the periphery, the three series have been reduced to two (aspirated *ph, th, kh* and unaspirated *p, t, k,* or, as in Istanbul, *b, d, g*). These differences are concealed in the traditional orthography. Sibilants (fricatives) of various types and affricates have emerged through palatalization of the old palatal and labiovelar stops. Thus, Old Armenian *dz* (*z*) and *ts* may go back to the Indo-European palatal stops *ǵh* and *ǵ*, and *dž* (*ž*) to the Indo-European labiovelar $g^{w}h$ before *e* and *i;* Old Armenian *tsʻ, tšʻ,* and *tš* may derive from the Indo-European consonant clusters *sk, ky,* and *gy.* All Armenian dialects distinguish two types of *r,* one strongly trilled, one weakly trilled. Old Armenian also differentiated between two types of *l,* a neutral one and a velarized variety, which is made by moving the back of the tongue nearer to the soft palate at the back of the mouth. The latter type has developed into a voiced velar fricative in the modern dialects.

Grammar. All of the spoken dialects and the two literary languages have maintained a fairly complicated system of noun declension, distinguishing six or seven cases. The plural stem, derived from the singular stem by the addition of the suffix -(*n*)*er,* is declined as a singular, according to the Turkish pattern. Characteristic of the changes in the Old Armenian verbal system is the general replacement of simple present tense forms by periphrastic expressions. These are groups of words, including auxiliaries, that take the place of a single word that is capable of being inflected to show tense or some other feature. The various types of periphrastic forms serve as the basis for the classification of the dialects. In Old and Modern Armenian, the main tense distinction is that between present, aorist (denoting occurrence without reference to completeness, duration, etc.), and periphrastic perfect tenses. The old subjunctive, still extant in classical Armenian, has been lost in the modern language. To express future time, Old Armenian used the subjunctive forms; Modern Armenian employs periphrastic expressions, as is done in the English future forms *I shall go* and *he will work.* Also characteristic of Modern Armenian is the importance of the passive forms of the verb, which are strictly parallel to the active forms, and the emergence of a special negative conjugation with differing forms for verbs in instances like "I read" and "I don't read." Whereas Old Armenian was rather close to ancient Greek in many respects, Modern Armenian is typologically much closer to Turkish. Among the features that illustrate this similarity are the agglutinative system of declension (*i.e.,* the compounding of several linguistic elements of independent meaning into a single word), the use of suffixes to indicate possession, the employing of passive and causative forms for all verbs, and the use of postpositions (grammatical elements placed after the word) instead of prepositions (as used in Old Armenian). The vocabulary of the written languages is purely Armenian, being based almost exclusively on that of Grabar, with very few loanwords from the neighbouring languages. (The Iranian loanwords mentioned above were incorporated into Armenian before the creation of the written language.) (H.K.V.)

Tocharian language

Tocharian (Tokharian), also called Tocharish, is an Indo-European language that was spoken in northern Chinese Turkistan (Tarim Basin) during the latter half of the 1st millennium AD. Documents from about AD 500–700 attest to two dialects: Tocharian A, from the area of Turfan in the east; and Tocharian B, chiefly from the region of Kucha in the west but also from the Turfan area.

Discovery and decipherment. The first Tocharian manuscripts were discovered in the 1890s. The bulk of the Tocharian materials were carried to Berlin by the Prussian expeditions of 1903–04 and 1906–07, which explored the Turfan area, and to Paris by a French expedition of 1906–09, which investigated chiefly in the area of Kucha. Smaller collections are in London, Leningrad, and Japan.

Translation and analysis

Tocharian is written with a north Indian syllabary (a set of characters representing syllables) known as Brāhmī, which was also used in writing Sanskrit manuscripts from the same area. The first successful attempt at grammatical analysis and translation was made by the German scholars Emil Sieg and Wilhelm Siegling in 1908 in an article that also established the presence of the two dialects, provisionally called A and B. The Berlin collection consisted of both dialects, whereas all other manuscripts discovered were in B.

The German name Tocharisch was proposed, and the language was demonstrated to be Indo-European.

Characteristics. Tocharian forms an independent branch of the Indo-European language family not closely related to other neighbouring Indo-European languages (Indo-Aryan and Iranian). Rather, Tocharian shows a closer affinity with the western (*centum*) languages: compare, for example, Tocharian A *känt,* B *kante, känte* "100," and Latin *centum* with Sanskrit *śatám;* A *klyos-,* B *klyaus-* "hear," and Latin *clueo* with Sanskrit *śru-;* A *kus,* B *kuse* "who," and Latin *qui, quod* with Sanskrit *kas.* In phonology, Tocharian differs greatly from the other Indo-European languages in that all of the Indo-European stops of each series fall together, resulting in a system of three (voiceless) stops, *p, t,* and *k.*

The Tocharian verb reflects the Indo-European verbal system both in stem formations and in personal endings. Especially noteworthy is the wide development of the mediopassive form in *r* (as in Italic and Celtic); *e.g.,* Tocharian A *klyoṣtär* "is heard." The 3rd person preterit plural ends in *-r,* similar to Latin and Sanskrit perfect forms and the Hittite preterit. The noun, however, shows little trace of the original Indo-European inflection. Instead, it is built up by the addition of postpositions to the oblique (accusative) form. This type of inflection (agglutination) has been attributed to the influence of non-Indo-European languages (Turkish, Finno-Ugric).

The vocabulary shows a remarkable influx of loanwords—from Turkish, Iranian, and, later, Sanskrit. Chinese has had little influence. Many of the most archaic elements of the Indo-European vocabulary are retained—*e.g.,* A *por,* B *puwar* "fire" (Greek *pyr,* Hittite *paḫḫur*); A and B *ku* "dog" (Greek *kyōn*); A *tkaṃ* "earth" (Greek *chthōn,* Hittite *tekan*), and especially, nouns of relationship: A *pācar, mācar, pracar, ckācar,* B *pācer, mācer, procer, tkācer,* "father," "mother," "brother," and "daughter," respectively.

Literature. Tocharian literature is Buddhistic in content, consisting largely of translations or free adaptations of Jātakas, of Avadānas, and of philosophical, didactic, and canonical works. In dialect B there are also commercial documents, such as monastery records, caravan passes, medical and magical texts, and the like. These are important source materials for the social, economic, and political life of central Asia.

The "Tocharian problem." Since the appearance of Sieg and Siegling's article, the appropriateness of the name Tocharian for the language has been disputed. According to Greek and Latin historical sources, the Tochari (Greek Tócharoi, Latin Tochari) inhabited the basin of the upper Oxus River (modern Amu Darya) in the 2nd century BC and were probably Iranians. Sieg and Siegling's identification of this language as belonging to these people was probably in error.

Identification of the Tocharian-speaking people

There have, of course, been numerous attempts to identify the speakers of Tocharian with this or that people or tribe mentioned in Chinese annals or in other documents dealing with the area in and around the Tarim Basin during the last half of the 1st millennium AD. One such identification that gained some adherents was with the Wu-sun. Neither this nor any other identification, however, would appear to be more than mere speculation.

Furthermore, the name *ārśi* (*ārśiype* "Arśi-country," *ārśi-käntu* "Arśi-language"), once accepted as the native name in dialect A, is probably a loanword through Iranian from Sanskrit *ārya.* The question of the name is, however, of little linguistic importance. Tocharian, even if generally accepted as a misnomer, will probably remain. For dialect A and dialect B, the substitution of Turfanian and Kuchean, or of East Tocharian and West Tocharian, has been suggested.

Of greater importance, at least from the linguistic point of view, is the relationship of the Tocharian language to the other Indo-European languages and the interrelationship of the two dialects themselves. In the former regard, in spite of superficial resemblances to Italic and Celtic (see above), the more fundamental shared features—*e.g.,* common vocabulary, certain verbal categories (*s*-aorists, preservation of the perfect active participle), and possible relics of common phonological developments—would appear to align Tocharian with the more southeastern branches of Indo-European; that is, with Thracian and Phrygian or even with Greek and Armenian. Those features shared with the Baltic and Slavic languages (certain present and preterit formations in particular) might be the result of later contacts.

With regard to the interrelationship of the two dialects, it is possible that dialect A was, at the time of documentation, a dead liturgical language preserved in the Buddhist monasteries in the east, whereas dialect B was a living language in the west (note that commercial or at least nonliturgical documents are found in that dialect). The presence of manuscripts in B mixed with those in A in the monasteries of the east can be accounted for by ascribing the B manuscripts to a new missionary invasion of those monasteries by Buddhist monks from the west. (G.S.L.)

Celtic languages

The Celtic languages form one branch of the Indo-European language family, having in common certain sound shifts and vocabulary items. On both geographical and chronological grounds, the languages fall into two divisions, usually known as Continental Celtic and Insular Celtic.

Continental Celtic. Continental Celtic is the generic name for the languages spoken by the people known to classical writers as Keltoi and Galatae; at various times during a period of roughly 1,000 years (approximately 500 BC–AD 500), they occupied an area that stretched from Gaul to Iberia in the south and Galatia in the east. The great bulk of evidence for Continental Celtic consists of the names of persons, tribes, and places recorded by Greek and Latin writers. Only in Gaul and in northern Italy are inscriptions found, and the interpretation of these is in most cases doubtful. Given the nature of the evidence, knowledge for these languages is confined largely to the sound system and a small part of the vocabulary, and no certain conclusions can be reached as to their historical development or the differences between them.

Evidence for Continental Celtic

Insular Celtic. Insular Celtic refers to the Celtic languages of the British Isles, together with Breton (spoken in Brittany, France). As the name Breton implies, it is an importation from Britain and is not a Continental Celtic dialect. Although there is some scanty evidence from classical sources—mainly place-names—and a small body of inscriptions in the Latin and ogham alphabets from the end of the 4th to the 8th centuries AD, the main source of information on the early stages of these languages is manuscripts written from the 7th century onward in Irish and somewhat later in the British languages.

The Insular languages fall into two groups—Irish and British. Irish (often called Goidelic, from Old Irish *Goídel* "Irishman," or Gaelic, from *Gael,* the modern form of the same word) was the only language spoken in Ireland in the 5th century, the time when historical knowledge of that island begins. The two other members of this group, Scottish Gaelic and Manx, arose from Irish colonizations that began about that time. There were also important Irish-speaking colonies in Wales, but no trace of their language survives apart from a few inscriptions.

British (often called Brythonic, from Welsh Brython "Briton") had almost the same degree of influence on the island of Britain and the Isle of Man. Inscriptions and personal names surviving from Scotland show clearly that there was a non-Indo-European language spoken there, usually called Pictish, which was later replaced by British. There were undoubtedly dialectal differences within the island, but the existing dialects arose from the fragmentation of British by the Irish invasions of Man and what is now Scotland and by the English invasions that began in what is now southern England and finally reached Scotland. Scotland has ever since been partitioned linguistically between English (or "Scots") and Irish (or "Erse"—the Scots form of "Irish"—or "Gaelic"). A British dialect, now labelled Cumbric, lingered on in the western borderlands between England and Scotland until perhaps the 10th century, but almost nothing is known about it. In what is now Wales, British survived as the dominant language until a century or so ago; it is now known as Welsh. Another pocket of British speech survived in Cornwall until the end of the 18th century. It was from this area that emigrants in the

Re-establishment of Celtic on the European mainland

5th and 6th centuries AD had brought Celtic once more to the European mainland by establishing a colony in northwestern France, still called Brittany. It is just possible that there were some traces of the Continental Celtic language (*i.e.,* Gaulish) at that time in this remote area, although Breton is too similar to Cornish (an Insular Celtic tongue) to suggest any serious influence from Gaulish.

HISTORICAL DEVELOPMENT

Common Celtic. The reconstruction of Common Celtic (or Proto-Celtic)—the parent language that yielded the various tongues of Continental Celtic and Insular Celtic—is of necessity very tentative. Whereas Continental Celtic offers plenty of evidence for phonology (the sound system), its records are too scanty to help much with the grammar (morphology or syntax), for which the best available evidence is Old Irish, the most archaic of the Insular languages. The records provide a picture of a language of the same type as Latin or Common Germanic; that is, one that still maintains a considerable part of the structure of the ancestral Indo-European language and has not lost final or medial syllables. Its vowel system differs only slightly from that reconstructed for Indo-European by the French linguist Antoine Meillet. Differences include the occurrence of Celtic *ī for Indo-European *ē (*e.g.,* Gaulish *rix* and Irish *rí,* "king"; compare Latin *rex*) and *ā in place of *ō. (An asterisk [*] before a letter or word indicates that the sound or word is not attested but is a hypothetical, reconstructed form.)

The consonantal system, too, is conservative, although there are some striking features. Among them are the loss of **p* (*e.g.,* Irish *athair* "father"; *cf.* Latin *pater*) and the falling together of the aspirated and unaspirated voiced stops assumed for Indo-European. (A stop is a consonant made with complete momentary stoppage of the breath stream some place in the vocal tract; voiced stops are those produced with the vocal cords vibrating, such as *b, d, g.* An aspirated sound is accompanied by a puff of breath, often written as an *h,* as in *bh, dh, gh;* an unaspirated consonant lacks this accompanying puff of breath.) Thus, Old Irish *dā́n* "what is given" corresponds to Latin *donum* "gift" (from Indo-European **d*), but Old Irish *de-naid* "sucks" corresponds to Latin *fe-* in *fe-mina, fe-llare* (from Indo-European **dh*). This loss of distinctive aspiration occurs with three out of the four voiced stops, a situation close to that of Slavic.

Other considerations, however, show that Celtic belongs to the so-called southern group of the European branch of Indo-European languages, or in another classification, to the same centum group as Latin, whereas Slavic belongs to the satem group. (The centum and satem divisions of Indo-European languages are made according to the treatment of certain sounds, called palatals, that existed in the ancestral Indo-European language.)

Correspondence of Celtic *p* and *q*

The loss of **p* in Celtic was very early; only the place-name Hercynia, preserved in Greek, shows that, in initial position, it became an *h* sound before disappearing. In most of the known Celtic languages, a new *p* sound has arisen as a reflex of the Indo-European *k^w sound. Thus there is Gaulish *pempe,* Welsh *pimp* "five," compared to Old Irish *cóic* and Latin *quinque* "five." The Irish evidence shows that *k^wenk^we must be reconstructed as the form in Common Celtic. The terms P-Celtic and Q-Celtic are sometimes used to describe assumed divisions of Common Celtic; to use one sound shift to distinguish dialects is, however, hardly justified, and the classification will not be used in this article.

The morphology (structure) of nouns and adjectives shows no striking changes from Indo-European. The Irish verb, however, exhibits a remarkable archaism not found in any other recorded Indo-European language. It has recently been demonstrated that the so-called primary and secondary endings of the Indo-European verb, as in the 3rd person singular endings *-(*e*)*t* and -(*e*)*ti,* both occurred in the same tense. The forms with *-*i* were used when the verb had absolute initial position; those without it were used in the normal verbal position at the end of the sentence. This is reflected in the Old Irish forms *beirith* (from **bereti*) "he bears" and *ní beir* (from **beret*) "he does not bear." It cannot be stated with certainty that Continental Celtic had preserved such forms. The Continental Celtic dialects show a few cases of sentences—admittedly imperfectly understood—in which the verb appears to be placed after the subject and before the object, as in modern western European languages. The history of Insular Celtic, however, shows a gradual shift from the older final position of the verb to the initial position, a position that has now become regular in all of the languages.

Relationships and ancient contacts of Celtic. The question of the relationship of Common Celtic to the other Indo-European languages remains open. For some time, it was held that Celtic stood in an especially close relation to the Italic branch; some scholars even spoke of a period when an Italo-Celtic "nation" existed, toward the end of the 2nd millennium BC. The existence of a *q–p* relationship (see above) inside Italic too (*e.g.,* Latin *quattuor* "four," but Oscan *petora*) was thought by some to support this view. Much of this argument is, however, based on accidental resemblances (*e.g.,* the Irish future tense in *f-* and the Latin future in *b-*) or on formations such as the deponent and passive verb forms ending in *-r,* which at one time were known mainly in Italic and Celtic but have since been found in the Hittite and Tocharian languages as well. The undeniable common features between Celtic and Italic, such as the superlative endings of adjectives (Latin *-issimus;* Celtic *-*samos,* *-*isamos*), are hardly sufficient to justify the assumption of a special relationship, and the whole concept of an Italo-Celtic unity has been powerfully criticized by the linguists Carl J.S. Marstrander and Calvert Watkins.

The original home of the Celts cannot be located precisely, but, on the whole, the evidence points to the eastern part of central Europe. There is more evidence for their contacts with other Indo-European peoples. One group of Celts, at least, found themselves neighbours of the Germanic peoples and were often confused with them by classical writers. It can be inferred that the Celts had attained a higher standard of social organization than the Germanic peoples from the existence of words such as Gothic *reiki* and *andbahts* (modern German *Reich, Amt*), apparently borrowed from Celtic **rīgion* "kingdom" and **ambactos* "officer." To the Greeks and Romans, on the other hand, the Celts were inferior in culture; Celtic words in Greek are restricted to those describing Celtic institutions, such as *bardoi* "poets." The borrowings from Celtic into Latin, which derive mainly from the period before the expansion of Roman power, belong to a few restricted categories, such as war (*lancea* "lance"), transport (*carrus* "baggage wagon" and *carpentum* "light carriage"), and agricultural products (*cervesia* "beer"). When the Romans finally conquered Gaul and imposed their language, a number of Celtic words came into Latin, but the Celtic terms were mainly concerned with rural life. These are more common in French dialects than in standard French, which preserves a mere handful, such as *mouton* "sheep," *ruche* "beehive," and *arpent* "land measure."

Contacts with the Greeks and Romans

Early records of Celtic. *Continental Celtic.* Celtic died out very quickly in eastern Europe. In its farthest outpost in Asia Minor, it may be assumed that the *Letter of Paul to the Galatians* was addressed to a people whose culture was already Greek but whose Celtic origins are clear from names preserved by classical authors; *e.g.,* Drunemeton "the very sacred place," formed from two distinctively Celtic elements. When commenting on that Letter, St. Jerome (died AD 419/420) said that the Galatians still spoke a language almost the same as that of the people of Trier. As he had been in both places, his evidence cannot be dismissed offhand, and it may be that a few speakers of Celtic still existed in both areas. Since St. Jerome did not claim to know any Celtic dialect, however, his statement cannot be accepted with certainty. Speaking generally, the history of Continental Celtic comes to an end as that of Insular Celtic begins.

Insular Celtic. The earliest evidence for Insular Celtic consists, like that for Continental Celtic, mainly of names recorded by Greek and Latin authors. In the case of Ireland, these were entirely by hearsay, and many of the Irish place-names recorded by Ptolemy in the 2nd century AD have not yet been identified. From perhaps the 4th century, ogham inscriptions (see WRITING) are found in Ireland, consisting almost entirely of personal names. From the 5th century onward, British names in Latin inscriptions are recorded in Wales, as well as Irish names in both Latin and ogham alphabets in areas of Irish settlement. These scanty records are of value above all in establishing that, up to very nearly the time at which written documents become available, British and Irish had remained similar in structure to Gaulish. Thus, in Britain is found the genitive (possessive form) Catotigirni, which in Old Welsh gives Cattegirn, and in Ireland the genitive Dovatuci exists, which in Old Irish gives Dubthaich. These changes—loss of final syllables and connecting vowels, weakening of consonants between vowels, and so on—are very similar to what was happening to Latin in France at the same time; *e.g.,* Latin *avicellus* and *aqua* finally became *oiseau* and *eau.* There is no satisfactory explanation of why these profound changes should have occurred at this time, nor can the period at which they occurred be fixed precisely, for the engravers of the inscriptions clearly went on using traditional forms long after the sound had changed.

LINGUISTIC CHARACTERISTICS OF THE INSULAR CELTIC TONGUES

The new languages, the only forms of Celtic that are known thoroughly, present a considerable number of unusual features, some of them unknown to other Indo-European languages. Some scholars have argued that these features may have resulted from the presence of a large non-Celtic substratum in the British Isles. Because it is hardly likely that the Celtic invasions of those islands began much before 500 BC or that the invaders exterminated the existing inhabitants, such a possibility cannot be denied. On the other hand, some features once thought to be exotic, such as the initial position of the verb in the sentence, have been convincingly demonstrated to be organic developments from Indo-European. Others, such as the system of counting by 20s, are clearly innovations, but this system is shared by English ("three score and ten"), French (*quatre-vingts* "80"), and Danish, in all of which it is also an innovation, as well as Basque, in which it appears to be old.

Possible influences of substratum languages

Phonological characteristics. The most remarkable phonological feature of Insular Celtic is the development of a double series of consonants in which strongly articulated consonants are distinguished from their weak counterparts. The two series were originally merely phonetic variants, with the strong variety occurring in absolute initial position and in certain consonant clusters and the weak elsewhere. Later, however, the two series became independent, or phonological. In the languages as they first appear in writing, considerable changes have taken place in the phonetic forms of the two series. Both in Irish and in Welsh, Cornish, and Breton, the opposition (contrast) of strong:weak in the voiced stops has been replaced by stop:spirant (*e.g., b:v*). (A spirant, such as *v, f, s,* is produced with local friction and without complete stoppage of the breath stream.) Irish has the same system for the unvoiced stops (*e.g., t:th*), but Welsh, Cornish, and Breton have voicing in this instance (*e.g.,* voiceless *t*:voiced *d*). These changes by themselves are not very different from the weakening of consonants between vowels that occurs in other western European languages (compare Welsh *pader* "prayer," a loanword from Latin *pater* "father," with Spanish *padre* "father," deriving from Latin *patrem*), but, in Insular Celtic, they occurred not only inside the word but also inside the phrase, so that the initial consonant of a word preceded by another word ending in a vowel was weakened. When the final syllables were lost in the evolution to the modern languages, these variations remained, and a system of initial mutations (changes) was thus set up. If, for example, a Goidelic nominative form **sindos kattos koilos* "the thin cat" is reconstructed, this will give Old Irish *in catt coel* after the loss of final syllables, but the genitive **sindī kattī koilī* "of the thin cat" will give *in chaitt choīl* with changed initial consonants. The same sort of change occurred in one Italian dialect: in Tuscan, there occur *porta* "door," *la forta* "the door," *tre porte* "three doors," from Latin *porta, illa porta, tres portae.* In both cases, consonant weakening has spread from word to sentence; there is a common development, but it cannot be claimed that it is distinctively Celtic.

Grammatical characteristics. Another feature of Insular Celtic is its lack of the infinitive form of the verb found in most other Indo-European languages—*e.g.,* English "to do," "to call." The equivalent is the verbal noun, which is a noun closely linked to the verb, though not necessarily derived from the same stem. Being a noun, it can have a following noun in the genitive case, which, in the older languages at least, is subjective or objective according to whether the verb with which it is linked is intransitive or transitive. Thus, from the Old Irish sentence *téit in ben* "the woman goes," the verbal noun phrase *techt inna mná* "the coming of the woman" can be derived, whereas from *marbaid in mnaí* "he kills the woman" can be formed *marbad inna mná* (*lais*) "the killing of the woman (by him)." Among many other functions of the verbal noun is its use, when preceded by the appropriate preposition, with the substantive verb to provide a tense with continuous meaning. Thus, to *téit in ben* there is a parallel *a-tá in ben oc techt* "the woman is at going" (= "the woman is going"), and to *marbaid in mnaí* corresponds *a-tá oc marbad inna mná* "he is killing the woman." The close resemblance of this system to that of modern English, in which it is a comparatively recent development, has been variously explained as the working of a substratum or, more recently, in terms of areal (regional) development.

Verbal nouns

MODERN LANGUAGES OF THE FAMILY

The discussion of the individual languages that follows divides them into the two main groups, beginning with Irish, which is the oldest attested.

Irish. The history of Irish may be divided into four periods: that of the ogham inscriptions, probably AD 300–500; Old Irish, 600–900; Middle Irish, 900–1200; and Modern Irish, 1200 to the present. This division is necessarily arbitrary, and archaizing tendencies confuse the situation, especially during the period 1200–1600, when a highly standardized literary norm was dominant. After 1600, the modern dialects, among them Scottish Gaelic and Manx, begin to appear in writing.

The Latin alphabet was introduced into Ireland by British missionaries in the 5th century and soon began to be used for writing Irish. By the middle of the 6th century, the process of putting into literary form the rich oral tradition of the native learned class was certainly well advanced. The problems of interpreting the early writings are complicated by the fact that the orthography was based on that of Latin, but with a British pronunciation; *e.g.,* Latin *pater* was read as *pader,* the form of the loanword in Modern Welsh, and Old Irish *Pátric* was read as *Pádraig* (as it is spelled in Modern Irish). No new letters were evolved; the weak (less forceful) consonants were distinguished only in instances in which there were Latin spellings that could be utilized (*e.g.,* strong *ll:* weak *l,* strong *rr:* weak *r, nn:n, c:ch, t:th*) or with the help of the *punctum delens* (*s:ṡ, f:ḟ*), a dot that shows that the sound is not pronounced. As a result, many ambiguities remain: *ní beir* can mean either "he does not carry" or "he does not carry it," according to whether the *b-* is read as a *b* sound or a *v* sound. Nor was the Latin alphabet capable of dealing with the new system of consonant quality that appears in Irish alone among the Celtic languages. Thus, from the Celtic nominative singular and plural forms *bardos, bardī* developed Welsh *bardd,* plural *beirdd,* with a vowel alternation like that of English "mouse, mice." In Irish, the forms are *bard, baird;* the *-i-* of *baird* is purely graphic, serving to indicate that the following consonants are both palatalized. (Palatalized consonants are those in which the pronunciation is modified by raising the tongue toward the hard palate.) This palatalization had been purely phonetic as long as the *-ī* that caused it survived, but in Old Irish the palatalization became independent, so that each consonant of Common Celtic evolved into four distinct consonants (*i.e.,* phonemes); for example, from original Common Celtic *b* are derived a *b* sound and a palatalized *b* sound, and a *v* sound and a palatalized *v* sound.

Development of palatalized consonants

Apart from these phonetic developments, Old Irish is striking chiefly for the extraordinary proliferation of particles that appear before the verb and are used in forming compound verbs. For example, the Latin word *suffio* "I fumigate" is translated as *fo-timmdiriut,* composed of *fo* "under," *to* "to," *imb-* "around," *di* "from," and the stem *reth-* "run," with vowel and consonant changes appropriate to the 1st person singular present tense. Such forms, combined with a system of infixed accusative and dative pronouns (*i.e.,* pronouns inserted within a word) and syntactical accent shifts, produced a verbal system almost as complicated as that of Basque, though transparently Indo-European in origin. This system began to break down during the Old Irish period; the process was no doubt accelerated by the Viking raids that began at the end of the 8th century and that disrupted the monastic system, the guardian of the literary norm of Old Irish. Popular forms broke through in the Middle Irish period, though always mixed with archaizing forms; the backward-looking Irish scribes were never content to write down their own vernacular. During the 12th century, many ecclesiastical synods were held with the object of bringing the organization of the Irish Church more closely into line with that of western Europe, and the Anglo-Norman invasion took place in the latter part of the same century. It may have been these far-reaching changes that inspired the Irish literati to undertake a new standardization of their language. From the beginning of the 13th century, there is a rigidly fixed norm, often called Classical Modern Irish, which, for over four centuries, was used as the exclusive literary medium in Ireland and in Gaelic-speaking Scotland (there is no evidence for the Isle of Man).

The Scandinavians were first contained and then absorbed; they contributed a small number of loanwords to Irish, mainly in the field of navigation but also in that of urban life, for they were the first to establish towns in Ireland, though only on the coast. The Anglo-Normans were a more serious problem. After almost complete success in the early period, however, they became largely Gaelicized in custom and language outside the towns they had founded. They contributed a large number of loanwords to Irish in the fields of warfare, architecture, and administration, though many of these were comparatively short-lived. When English took over from Anglo-Norman as the language of administration and English colonies began to be planted in Ireland, English loanwords began to come into Irish. Few of these, however, were recognized in the literary language, and only from the evidence of the modern dialects has it become clear that they were quite numerous.

The decline of Irish

It was not until the beginning of the 17th century that the English power was finally consolidated in Ireland, first by military conquest and later by the planting of English-speaking colonists on a much larger scale than before. From this time onward, the decline of Irish began, with Irish becoming the language of an oppressed people. With no schools to teach the literary language nor any native nobility to support the literati who used it, the dialects appeared for the first time and began to be written in the paper manuscripts that constituted almost the only form of publishing available to those using Irish. By the beginning of the 19th century, it is probable that the population was almost equally divided between Irish speakers, mainly in the western half, and English speakers, mainly in the eastern half. The real imbalance lay in the fact that many of the Irish speakers were bilingual, whereas few of the English speakers were. The first census to record language use was taken in 1851, after the great famine that had struck the western areas with exceptional severity. By this time, the total number of Irish speakers was 1,524,286 (23 percent of the population), but only 319,602 spoke Irish exclusively. The decline of Irish has continued to the present day, in spite of a revival campaign initiated by the Gaelic League in 1893 and made part of official policy after the establishment of the Irish Free State in 1921.

Since then, Irish has been recognized as the first official language of the state; it is a compulsory subject in all of the schools and is a requirement for civil service and some other posts. There are probably more people able to read Irish—perhaps 300,000—than there ever were before. From 1945 onward, a standard written language has evolved, and there is a small but flourishing literary movement. Nearly all of the readers of Irish are English speakers by upbringing, however, and not many of them would claim that Irish had become their main language. In the western areas in which Irish was the traditional speech, there are now fewer than 50,000 people to whom it is a mother tongue, and all but a handful of these have a more or less adequate command of English.

Scottish Gaelic. Some aspects of the modern Scottish Gaelic dialects show that they preserve features lost in the language of Ireland during the Old Irish period; such archaism is characteristic of "colonial" languages. The innovations are, however, more striking than the archaisms. Most remarkable is the loss of the voicing feature (*i.e.,* the vibration of the vocal cords) in the stops. All of the stopped consonants are unvoiced, and the original voiceless stops have become strongly aspirated; for example, the equivalent of Irish *bog* "soft" is [pok], *p* being the voiceless counterpart of *b,* and that of *cat* "cat" is [k$^{\text{h}}$aht], the superscript $^{\text{h}}$ after *k* indicating the aspirated quality. (The brackets indicate that the symbols printed within them are phonetic rather than orthographic.)

Scottish Gaelic was planted on British soil, and the verbal system has been remolded on the lines of the British language, which originally had no future tense. As in Modern Welsh, the inherited present tense has largely future meaning, and present time is mainly expressed by the present-tense form of the substantive verb and the preposition *a(ig)* with the verbal noun. (In Insular Celtic, there are two verbs for "to be," a substantive verb with the meaning, roughly, "to exist," and a linking verb such as "is" in "John is a boy" or "sky is blue.") Thus, from Old Irish *téit in ben* "the woman goes" is derived Scottish Gaelic *thèid a' bhean* "the woman will go," and from Old Irish *a-tá in ben oc techt* "the woman is going" results the Scottish Gaelic form *thà an bhean a' dol* "the woman goes" or "the woman is going."

It is only from the 17th century onward that the development of Scottish Gaelic can be studied, for, up until then, Classical Modern Irish was the literary norm. Indeed, the first book to be printed in Irish was a translation of the Calvinist *Book of Common Order,* published in Edinburgh in 1567, and the Scottish Reformers used the Irish Bible for some time, until it became clear that it was too foreign for the people to understand. A native Scottish standard emerged gradually during the 17th century, as poets ignorant of the Irish norm began to compose in their native dialects. It was not until the 18th century that the orthography became more or less fixed, and, until recent reforms in Ireland, the divergencies between the written languages were comparatively small. It is clear, however, that Scottish Gaelic must now be regarded as a separate language, though the differences between it and Irish are no greater than those between standard German and the Swiss dialects.

Scottish Gaelic as a separate language from Irish

Scottish Gaelic was confronted by northern dialects of English (Scots) from the very beginning; these rapidly penetrated into the east of the country, especially in the area centred on Edinburgh, the capital. The so-called Highland Line, marking the boundary between the two languages, has been steadily receding to the west since medieval times. By 1901, there were 230,806 speakers of the language, including 28,106 who spoke Scottish Gaelic exclusively; 106,466 persons, including nearly all of the monolingual Scottish Gaelic people, lived in the two counties of Inverness and Ross. The decline has continued steadily, and, even in those two counties, Gaelic is rapidly disappearing from the mainland, though it is holding its ground well in the Hebrides. Scottish Gaelic speakers in the early 1980s numbered about 90,700, which shows that the state of Scottish Gaelic survival is in many ways less serious than that of Irish. Because the majority of Gaelic speakers are Protestants who are accustomed to reading the Bible and using the vernacular in their religious services, literacy in Gaelic has been widespread. Furthermore, however low the census figures may be, they give an accurate picture of the number of those to whom Gaelic is a mother tongue because the number of English speakers who have acquired it is negligible. It must be admitted, however, that the recent literary revival finds its audience among the displaced Gaelic speakers of Edinburgh and Glasgow rather than in the Hebrides, where Gaelic is still confined to the home and English is the language of culture. In addition, there were about 500 Gaelic speakers in Nova Scotia in the early 1980s.

Manx. The history of the Isle of Man is imperfectly known. It was first inhabited by British speakers, then colonized from Ireland, and later became part of the Scandinavian Lordship of the Isles until 1266, when the King of Norway ceded both Man and the Hebrides to Scotland. From then on, it became involved in the wars between England and Scotland until 1346, when it passed finally to England. Though an Irish dialect survived as the speech of the majority of the people, these circumstances were not propitious for literary contacts with Ireland, and Manx was apparently not written until the Welsh bishop John Phillips translated the Anglican *Book of Common Prayer* in 1610, using an orthography based on that of English. This orthography makes Manx difficult to understand for readers of Irish and Scottish Gaelic, to whom it is of considerable interest because it represents a dialect entirely free of literary influences. The orthography soon became fixed, and a far-reaching series of later phonetic changes made the written form a highly inaccurate representation of the final stages of the language. Phonologically, it has more in common with the eastern dialects of Irish than with Scottish Gaelic, but its morphology and syntax are much more like those of Scottish Gaelic, probably because of the common British substratum. Its tense system is similar to that of Scottish Gaelic and Welsh, and its use of periphrastic verb forms (*i.e.,* longer forms with several elements) with the auxiliary meaning "to do" goes further than either of these, especially in its final stages.

Extinction of Manx in the early 20th century

In the beginning of the 18th century, English was still not understood by most of the people, but during the 19th century the decline of Manx was rapid, and the census of 1901 showed only 4,419 speakers of the language, all bilingual. Twenty years later, the language had ceased to be used as a normal means of communication, but, until recently, investigators have been able to find old people capable of giving useful information.

British languages. Britain was thoroughly romanized, and it is clear that the British language itself had been much affected by Latin; on the level of vocabulary, such an everyday word as Welsh *pysg* "fish," for example, derives from Latin *piscis.* The vowel system lost independent vowel quantity, the length of vowels becoming determined by the structure of the syllable, a situation that also occurred when the later Latin developed into Romance. Even after the collapse of Roman rule, Latin retained the same prestige among British Christians that it had in the rest of the Western Empire. The Irish monks introduced to the British speakers the custom of writing down the vernacular language at about the end of the 8th century; they adapted the clumsy Irish orthography for that purpose. At this period, the British dialects were very close to one another and can hardly be classed as separate languages, though they soon began to diverge. Like Old Irish, they had lost their final syllables and had undergone many other changes from the state shown by the inscriptions. Notably, the languages show only the merest traces of the declension of the noun, although the verb preserves a full inflectional system (that is, it has a full series of endings). It is clear that no future tense existed in early British, though the separate languages were later to fill this gap by various means.

Welsh. Welsh is the earliest and best attested of the British languages. Although the material is fragmentary until the 12th century, the course of the language can be traced from the end of the 8th century. The earliest evidence may represent the spoken language fairly accurately, but a poetic tradition was soon established, and by the 12th century there was a clear divergence between the archaizing verse and a modernizing prose. The latter was characterized by a predominance of periphrastic verbal-noun constructions at the expense of forms of the finite verb. By this time, too, the forms corresponding to other Celtic and Indo-European present-tense forms had largely acquired future meaning; *e.g.,* Welsh *nid â* "he will not go" (future) contrasts with Irish *ní aig* "he does not drive" (present). The gap thus left was filled, as in Scottish Gaelic and Manx, by a construction involving the substantive verb and the verbal noun; *e.g., y mae'r wraig yn myned* "the woman goes" or "the woman is going" is composed of the verb *mae* "is" and the verbal noun *myned* "going."

By the 14th century, prose and verse styles became more similar, the prose being less colloquial and the verse less archaic. This marks the beginning of modern literary Welsh, which was finally fixed by the Bible translation of 1588. Modern literary Welsh developed at a time when Welsh national identity was beginning to be seriously threatened by the close relations with England that followed on the accession of the Welshman Henry Tudor (Henry VII) to the English throne in 1485. Welsh was being written less and less, and the spoken language was being penetrated by English words. In 1536, the Act of Union deprived Welsh of its official status. By the beginning of the 18th century, the position of the Welsh language had fallen very low, though it was still the vernacular of the vast majority of the people. It was saved by the Methodist revival of the 18th century, which established schools everywhere to teach the people how to read the Welsh Bible and which brought the Bible itself, together with Welsh religious books, into almost every home. The literary language rejected most of the English loanwords that had come into the popular speech, and, by the 19th century, a highly literate Wales was equipped with reading material of every kind in the Welsh language. Meanwhile, however, the popular speech diverged further from the fixed literary norm, which was never spoken except in the pulpit or on the platform. Modern Wales has a literary language that no mother speaks to her child and widely differing dialects that appear in print only to represent dialogue in stories and novels.

Effect of the 18th-century Methodist revival on Welsh

The Industrial Revolution of the 19th century first undermined the dominance of Welsh in Wales: English-speaking workers were brought into the mines and factories in such numbers that they could not be absorbed linguistically. By 1901 English speakers outnumbered Welsh speakers for the first time. Out of a population of 2,012,876, only 929,824 were reported as Welsh-speaking, though 280,985 people spoke Welsh alone. By the early 1980s the number of Welsh speakers had dropped to about 395,000, representing about 14 percent of an increased population. Most of rural Wales, however, is still Welsh speaking, and recent years have seen a great improvement in the official status of Welsh and a considerable increase in its use in the schools; it is certainly the most firmly rooted of the modern languages of Celtic origin.

In addition, there are still about 8,000 Welsh speakers in parts of Patagonia, Argentina, which was colonized by Welsh settlers in 1865. These people maintain cultural contacts with the homeland but are all bilingual in Welsh and Spanish and seem fated to final assimilation.

Breton. Breton disappeared from sight after the early period, and no literary texts are available until the 15th century. These, mainly mystery plays and similar religious material, are written in a standardized language that is by now completely differentiated from Welsh and, to a lesser degree, from Cornish. The divergence between Breton and Cornish is largely a matter of the English loanwords in Cornish and the French loanwords in Breton. The present tense was retained in its original function, whereas a future and conditional were formed from the present and past subjunctive, respectively. Later, the Breton dialects became written and showed considerable divergencies in this form. Not until the 1920s was an attempt at standardization made, and even then it was necessary to adopt two norms. One was called KLT, from the initials of the Breton names of the dioceses of Cornouaille, Léon, and Trĕguier, the dialects of which agree with Welsh and Cornish in having the stress accent on the next to the last syllable. The other norm was the dialect of Vannes in the south, which has the stress accent on the final syllable and many other distinctive features, at least some of which can be explained by its close contacts with French. More recently, two norms have been evolved to cover all four

Breton dialect norms

dialects; one of these is used by most writers, whereas the other is officially recognized by the universities of Brest and Rennes, in both of which Breton is taught.

Up until recently, Breton was the common language of the people in Cornouaille, Léon, Tréguier and Vannes, within the boundaries of the *départements* of Côtes-du-Nord, Finistère, and Morbihan. Breton may still have more speakers than Welsh, but this is quite uncertain because no language statistics exist for France. There is, however, general agreement that very few children today are being brought up speaking Breton. This is at least partly the result of French official policy, which in effect excludes the language from primary and secondary schools, though the poor economic opportunities in Brittany also play a part. The literary movement is, therefore, confined to an intelligentsia of perhaps not much more than 10,000 people, many of whom live outside Brittany. The overwhelming mass of the remainder of Breton speakers are literate only in French, and chances for the survival of Breton seem very poor.

Cornish. Like Breton, Cornish had no literary texts before the 15th century. Those that exist are mainly mystery plays, some of which are almost literal translations from English. Cornish is much closer in structure to Breton than to Welsh, but it has also been heavily influenced by English. At the beginning of the 18th century, there were still a number of areas in which Cornish was spoken, but it died out as a means of communication by the end of the century. (D.Gr.)

Baltic languages

The Baltic languages form a branch of the Indo-European language family and are more closely related to Slavic, Germanic, and Indo-Iranian (in that order) than to the other branches of the family. They comprise modern Lithuanian and Latvian (Lettish), the languages of the Balts inhabiting the eastern coast of the Baltic Sea, as well as the now extinct Old Prussian language, Yotvingian (also spelled Yatvingian, Jotvingian, Jatvingian), Curonian (Kurish), Semigallian, and Selonian (Selian); the speakers of this group are here referred to as the B-Balts. There also existed languages and dialects of the Balts (D-Balts) who lived east of the above-mentioned groups in the areas of the upper reaches of the Dnepr River.

Languages of the group. Because its dialects are more archaic in their forms than those of the other living Indo-European languages, Lithuanian is of particular importance in the study of comparative Indo-European linguistics. The language had 2,760,000 speakers in Lithuania in the early 1980s and several thousand speakers in Belorussia and Poland, and until 1945 there were several thousand Lithuanians in East Prussia as well. More than 675,000 Lithuanians live abroad, mostly in the United States. Lithuanian is sharply divided into dialects whose differences are quite marked. The two major ones are Low (or Western) Lithuanian, with three subdialects, and High (or Eastern) Lithuanian, with four subdialects. The Low dialect is spoken by the Lowlanders, who live in the west and along the Baltic Sea; High Lithuanian is spoken by the Highlanders, who live in the eastern (and greater) part of Lithuania. Standard Lithuanian, formed at the end of the 19th and the beginning of the 20th century, is based on the southern subdialect of West High Lithuanian.

Lithuanian, most archaic living Indo-European language

The language most closely related to Lithuanian is Latvian, spoken by 1,344,000 speakers in Latvia in the early 1980s and about 156,000 abroad, mostly in the United States. Latvian is divided into dialects, the major ones being the Central dialect, Livonian (also called Tahmian, or West Latvian), and High (or East) Latvian. Standard Latvian, established at the end of the 19th and the beginning of the 20th century, is based on the Central dialect.

Extinction of several Baltic languages

By the 16th century the Selonians, Semigallians, and Curonians (Kurs), who lived in areas of Latvia and Lithuania, had completely lost their national identities and were assimilated by the Latvians and the Lithuanians. They left no written records. Nor did the Yotvingians (or Suduvians), who lived in southwest Lithuania and farther to the south (in the territory of the present-day Poland). They became extinct around the 16th–17th century, being assimilated by the Lithuanians in the north and the Slavs in the south. Information on the extinct Baltic languages is extremely scarce (mostly place-names). Only Old Prussian, of all the extinct Baltic languages, left any written records, and they are quite poor. The Prussians lived in East Prussia (*i.e.,* between the lower reaches of the Vistula and Neman [Lithuanian Nemunas] rivers on the Baltic coast). They became extinct (*i.e.,* were assimilated by the Germans) at the beginning of the 18th century.

Linguistically, the Yotvingians were very closely related to the Prussians. They made up one ethnic Baltic group, commonly called the Western Balts, as opposed to the so-called Eastern Balts—the Lithuanians, Latvians, Selonians, Semigallians, and Curonians. The traditional terms Western Balts and Eastern Balts are inaccurate when used for all of the Balts—*i.e.,* including the Balts for whose languages there are no records (the D-Balts). These Balts, who were assimilated by Slavs in the 7th–14th century, lived in the upper reaches of the Dnepr.

Historical survey. Proto-Baltic, the ancestral Baltic language from which the various known languages evolved, developed from the dialects of the northern area of Proto-Indo-European. These dialects also included the Slavic and Germanic protolanguages (and possibly also Tocharian). The quite close historic relationship of the Baltic, Slavic, and Germanic languages is shown by the fact that they alone of all the Indo-European languages have the sound *m* in the dative plural ending (*e.g.,* Lithuanian *vĩlká-m-s* "wolf," Common Slavic **vilko-m-ŭ*, Gothic *wulf-am*). (An asterisk [*] indicates that the following sound or word is unattested and has been reconstructed as a hypothetical linguistic form.) This relationship is suggested not only by the morphology and word formation but also by the vocabulary—*e.g.,* Lithuanian *draũgas* (Latvian *dràugs*) "friend," Old Church Slavonic *drugŭ,* Gothic *driugan* "to fulfill military service"; Lithuanian *vãškas* (Latvian *vasks*) "wax," Russian *vosk,* Old High German *wahs.* Probably the earlier close contact of the Balts and the Slavs with the Germanic tribes broke off around the 2nd millennium BC, when the Balts moved from the south (not, however, losing contact with the Slavs) and settled a large area of the eastern coast of the Baltic Sea and the upper reaches of the Dnepr.

Features common to Baltic, Slavic, and Germanic

Relationship between Baltic and Slavic. Because contact between the Balts and Slavs from the time of Proto-Indo-European was never broken off, it is understandable that Baltic and Slavic should share more linguistic features than any of the other Indo-European languages. Thus, Indo-European **eu* passed to Baltic *jau* and Common Slavic **jau* (which became *ju*)—*e.g.,* Lithuanian *liáudis* "people," Latvian *ļáudis,* Old Church Slavonic *ljudije.* Tonal correspondences are found between Lithuanian and Serbo-Croatian (a Slavic language of Yugoslavia), and there are also similarities in stress; *e.g.,* Lithuanian *dũmai* "smoke" and Russian *dym* have the stress on the root, as do Lithuanian *rañką* "hand" (accusative singular) and Russian *rúku,* while both Lithuanian *rankà* "hand" (nominative singular) and Russian *ruká* are stressed on the second syllable.

Baltic and Slavic have specific morphological features in common. Among them, for example, is the genitive plural form. In Lithuanian, *mũsų* "of us" (= Latvian *mūsu*), evolved from the older form **nūsōn,* which comes from Baltic **nōsōn* and corresponds to the genitive plural form in Common Slavic, **nōsōn,* from which developed Old Church Slavonic *nasŭ* "of us." Baltic also shares some syntactic features with Slavic; *e.g.,* the genitive case is used in place of the accusative with verbs expressing negation (Lithuanian *jis* nieko *nežino* "he does not know anything," Latvian *viņš* nekā *nezin,* Russian *on* ničego *ne znajet*). There are also many lexical items common to Baltic and Slavic. More than 100 words are common in their form and meaning to Baltic and Slavic alone, among them Lithuanian *bėgu* "I run," Latvian *bēgu,* Old Church Slavonic *běgǫ;* Lithuanian *líepa* "linden tree," Latvian *liẽpa,* Old Prussian *lipe,* Old Church Slavonic *lipa;* Lithuanian *rãgas* "horn," Latvian *rags,* Old Prussian *ragis,* Old Church Slavonic *rogŭ.*

In addition to these features common to all the Baltic and Slavic languages, there are certain quite archaic features that Slavic has in common with Lithuanian and Latvian but not with Old Prussian. The most striking example is the genitive singular ending in Lithuanian *viĺk-o* = Latvian *vìlk-a* "of a wolf," which comes from Baltic **-ō,* historically paralleled by the genitive singular ending in Common Slavic **vĭl̃k-ā*. Old Prussian, however, has a different ending for the same inflection (*deiw-as* "of God"). In certain instances the Slavic languages, differing from Lithuanian and Latvian, come closer to Old Prussian; *e.g.*, the Prussian possessive pronouns *mais* "my, mine," *twais* "your, yours," *swais* "one's own" are different from Lithuanian *mãnas, tãvas, sâvas* and from Latvian *mans, tavs, savs* but similar to Old Church Slavonic *mojĭ, tvojĭ, svojĭ*.

It is possible to conclude that there was close contact between the Baltic and Slavic protolanguages at the time when they began to develop as independent groups (*i.e.*, from about the 2nd millennium BC) and that the Proto-Slavic area might have been a part of peripheral Proto-Baltic, although a specific part. That is, Proto-Slavic at that time was in direct contact with both the corresponding dialects of the peripheral Proto-Baltic area (*e.g.*, with Proto-Prussian) and the corresponding dialects of the central Proto-Baltic area. All this shows that the Proto-Slavic area of that time (south of the Pripyat River) was much smaller than the Proto-Baltic area. Proto-Slavic began to develop as a separate linguistic entity in the 2nd millennium BC and was to remain quite unified for a long time to come. Proto-Baltic, however, besides developing into an independent linguistic unit in the 2nd millennium BC, also began gradually to split. Among other things, the size of the Proto-Baltic area had an influence on the development of Proto-Baltic in that it considerably reduced contact between its dialects (see also *Slavic languages*).

Development of the individual Baltic languages. By the middle of the 1st millennium BC, the Proto-Baltic area was already quite sharply split into dialects. From the middle of the 1st millennium AD, the Baltic language area began to become considerably smaller; at that time the greater part of Baltic territory, the eastern part, began to be inhabited by Slavs migrating from the south. The Balts there became gradually assimilated by the Slavs; complete assimilation probably occurred around the 14th century. One of these Baltic tribes, the Galindians (Goljadĭ), is mentioned in a chronicle as late as the 12th century. The protolanguage of the so-called Eastern Balts split into Lithuanian and Latvian (Latgalian) around the 7th century. The other languages of the so-called Eastern Balts became separated probably at the same time. Selonian and Semigallian could have been transitional languages between Lithuanian and Latvian. Only Curonian, which some consider to be a transitional language between East and West Baltic, might have developed somewhat earlier. Moreover, the name of the Curonians occurs in historical sources earlier (AD 853: Latin Cori) than the names of the other tribes of the so-called Eastern Balts.

Split between Lithuanian and Latvian

Old Prussian. In historical sources the Prussians are called Aistians from the 1st century AD (by Tacitus) until the 9th century AD (by the Anglo-Saxon seafarer Wulfstan). They are referred to by their own name (by a Bavarian geographer using the form Bruzi, "Prussians") for the first time in the 9th century AD. About 1230 the Teutonic Order began to plunder the lands of the Prussians and finally conquered the Prussians and the Yotvingians (Suduvians) in 1283. From that time the slow extinction of the two Baltic groups began, with the Germanization of the Prussians being completed at the beginning of the 18th century.

The earliest Old Prussian (and, for that matter, Baltic) written record is a German–Prussian vocabulary—the so-called Elbing vocabulary, compiled about 1300 and extant in a copy dated around 1400. This vocabulary, consisting of 802 Old Prussian words (and the same number of German words), was written in a South Prussian dialect (in Pomesania). Somewhat poorer than the Elbing vocabulary is the vocabulary compiled by Simon Grunau, consisting of 100 Old Prussian (and German) words, written between 1517 and 1526. The most important Old Prussian written records are the three catechisms of the 16th century based on the dialects of Sambia and translated from the German; the first two catechisms, which are very short and anonymous, date from 1545, and the third catechism, or *Enchiridion,* dates from 1561 and was translated by Abelis Vilis (Abel Will), a pastor of the church at Pobeten (Pabečiai; modern Romanovo). The language of all the Old Prussian catechisms is rather poor: the translations are excessively literal, and there are many errors in language and orthography. In spite of this, it is from these Old Prussian catechisms that scholars can learn most about the Old Prussian language.

Lithuanian. Lithuanians are first mentioned in historical sources in AD 1009. Old Russian (more precisely, an East Slavic language based primarily on Belorussian), Latin, and Polish were used in official matters in the Grand Duchy of Lithuania, which was established in the mid-13th century and lasted until the 18th century. Lithuanian writings begin to appear in the 16th century, first in East Prussia (where many Lithuanians lived) and, somewhat later, in the Grand Duchy of Lithuania. In East Prussia, a quite uniform written Lithuanian language, based on the West High Lithuanian dialect, had already been established by the second half of the 17th century. In Lithuania, however, a uniform written Lithuanian came into use only at the end of the 19th and the beginning of the 20th century—*i.e.*, when a standard Lithuanian language, based on the (Southern) West High Lithuanian dialect (spoken in both East Prussia and Lithuania), was established. Martynas Mažvydas (died 1563), who published the first Lithuanian book (a catechism) in Königsberg (Lithuanian Karaliaučius; modern Kaliningrad) in the year 1547, is purported to be the first person to use Lithuanian as a written language. Others, in particular Baltramiejus Vilentas, Jonas Bretkūnas, and the pastor-poet Kristijonas Donelaitis, also took part in the formation and standardization of a written Lithuanian language in the 16th–18th century in East Prussia. Great influence was exerted by the first grammars of Lithuanian, by Danielius Kleinas (1653 and 1654), and the works of Donelaitis (1714–80), the first Lithuanian writer to become well known. In the Grand Duchy of Lithuania the first to use Lithuanian as a written language is held to be Mikalojus Daukša (died 1613), who published a catechism in 1595 and a prayer book (*Postilė*) in 1599. Among later writers who helped to standardize written Lithuanian were Konstantinas Sirvydas, who prepared the first dictionary of Lithuanian (1629), Jonas Jaknavičius (1598–1668), and Saliamonas Slavočinskis (17th century). The works of Daukša and Sirvydas in particular, based on the Middle and East High Lithuanian dialects, did much toward establishing the practice of drawing on the various dialects in the creation of a written Lithuanian. This tradition became weakened in the 18th century but was again revived at the beginning of the 19th, when the formation of a standard Lithuanian was undertaken. The practice became most apparent at the end of the 19th and the beginning of the 20th century, during the establishment of Standard Lithuanian. The process of the mixing and levelling of the Lithuanian dialects started at the beginning of the 20th century because of the influence of a standard language, and it was especially intensified after the creation of the Lithuanian Soviet Socialist Republic in 1940. Standard Lithuanian is the official language of the Lithuanian S.S.R., as it was of the Republic of Lithuania (from 1918).

Establishment of uniform written Lithuanian

Latvian. The Latvian (Latgalian) people achieved a separate identity around the 16th century AD, when they completely assimilated the other Balts, as well as a greater part of the Livs (also called Livonians, Livians), who are of Finnic descent and live on Latvian territory. As a result of the conquering of Latvian territory by the German Knights of the Sword by 1290, close contact between all of the so-called Eastern Balts (the Latvians with the Lithuanians as well) was considerably weakened for a long period of time. The first Latvian book was the *Catechismus Catholicorum* of 1585. In 1638 the first Latvian (–German) dictionary, by Georgius Mancelius, appeared; the first grammar of the Latvian language, by Johann Georg Rehehausen, was published in 1644; and a Latvian translation of the Bible was published in 1685. The Latvian writings of the 16th–

Early Latvian books

18th century are translations of religious works, as are the Lithuanian. The language of these Latvian works, however, is somewhat poorer than that of the Lithuanian writings of the same period. The works of the Latvians Juris Alunāns (1832–64) and Atis Kronvalds (1837–75) exerted a great influence on the development of a standard Latvian language, based on the Central dialect, at the beginning of the 19th century. Standard Latvian was finally established at the end of the 19th and the beginning of the 20th century, and the levelling influence of this standard language on the Latvian dialects began at this time. Standard Latvian is the official language of the Latvian S.S.R.

Characteristics of the Baltic languages. All of the Baltic languages are inflected. Old Prussian is the most archaic of the recorded Baltic languages (although it also has innovations of its own), and it differs considerably from Lithuanian and Latvian.

Old Prussian. In contrast to Lithuanian and Latvian, Old Prussian retained the Baltic diphthong *ei*—Old Prussian *deiws* "God," Lithuanian *diẽvas,* Latvian *dìevs;* Old Prussian *deinan* "day" (accusative singular), Lithuanian *dienà,* Latvian *dìeņa.* In place of Lithuanian *š* and *ž* (from Indo-European **ḱ, *ǵ,* and **ǵh*), however, Old Prussian, like Latvian (as well as Curonian, Semigallian, and Selonian), has *s* and *z*—thus, Old Prussian *assis* "axle," Latvian *ass,* Lithuanian *ašìs;* Old Prussian (*po*)*sinnat* "to confess," Latvian *zināt,* Lithuanian *žinóti* "to know." The cluster $s+j$ (and $z+j$) in Old Prussian, as in Latvian, passed to *š* (and *ž*): Old Prussian *schan* (from **sjan*) "this" (accusative singular feminine), Latvian *šùo* "this," Lithuanian *šią.* In contrast to Lithuanian and Latvian, Old Prussian did not replace the clusters $t+j$ and $d+j$ with affricate sounds (begun with complete stoppage of the breath stream from the lungs and released with incomplete closure and friction): Old Prussian *median* "forest," Lithuanian *medžias,* Latvian *mežs.*

Word stress was free in Old Prussian, as it is in Lithuanian (in contrast to Latvian, in which the stress is predictable and falls on the first syllable). Old Prussian also made use of intonations (tones), the character of which is similar to that of the Latvian (*i.e.,* more archaic than that of Lithuanian intonations). The Proto-Baltic circumflex intonation corresponds to the falling tone in Old Prussian, while the acute intonation corresponds to the rising tone.

Old Prussian morphology

Old Prussian, moreover, had a substantive neuter gender, lost by Lithuanian and Latvian: Old Prussian *assaran* "lake," Lithuanian *ẽžeras,* Latvian *ezer̃s;* Old Prussian *lunkan* "bast," Lithuanian *lùnkas,* Latvian *lūks.* It differs in morphology from Lithuanian and Latvian in more than one instance—*e.g.,* in the genitive singular ending, Old Prussian *deiw-as* "of God" (Lithuanian *diev-o* = Latvian *diev-a*) and, in the dative singular, Old Prussian *tebbei* "to you" (Lithuanian *tavi* = Latvian *tev*), among others. Old Prussian did not have the dual number, only the singular and plural. Nouns were declined according to seven types. There were five cases: nominative, genitive, dative, accusative, and vocative. All verbs had three separate forms in the plural, but not in the singular. The 3rd person was the same in both the singular and the plural. There were three tenses: present, preterite, and future.

In vocabulary Old Prussian is quite similar to Lithuanian and Latvian (closer to Lithuanian than Latvian). It should be emphasized, however, that Old Prussian differs from Lithuanian and Latvian in that it retained a greater number of archaisms than either.

Comparison of Lithuanian and Latvian. The differences between Lithuanian and Latvian can be summarized in very broad terms by saying that Lithuanian is far more archaic than Latvian and that modern written Lithuanian could in many instances serve as a "protolanguage" for it. For example, Lithuanian has quite faithfully preserved the old sound combinations *an, en, in, un* (the same is true of Old Prussian, Curonian, Selonian, and, possibly, Semigallian), while they have passed in every case to *uo, ie, ī, ū* in Latvian; thus, Lithuanian *rankà* (Old Prussian *rancko*) = Latvian *rùoka* "hand," Lithuanian *peñktas* (Old Prussian *penckts*) = Latvian *piekt*(*ai*)*s* "fifth," Lithuanian *pìnti* = Latvian *pīt* "to weave, to twine," and Lithuanian *jùngas* = Latvian *jūgs* "yoke." The diphthongs *ei* (as well as *ai*) and *au* in final position were monophthongized and later shortened in Latvian—*e.g.,* Lithuanian *ved-eĩ* (2nd person singular preterite) = Latvian **ved-ie,* which became *ved-i* "you led"; Lithuanian *med-aũs* = Latvian **med-uos,* which became *med-us* "of honey." Long vowels at the end of polysyllabic words have been shortened in Latvian, and short vowels have been dropped—*e.g.,* Latvian *sak-a* "says" (which derives from **-ā*) = Lithuanian *sãk-o,* Latvian *pel-e* "mouse" (from **-ē*) = Lithuanian *pel-ẽ,* Latvian *vìlk-u* "wolf" (from **-uo*) = Lithuanian *vil̃k-ą,* Latvian *daikts* "thing" (from **-ăs*) = Lithuanian *dáiktas,* and Latvian *nakts* "night" (from **-ĭs*) = Lithuanian *naktìs.* Palatalized *k* and *g,* formed with the blade of the tongue closer to the hard palate than nonpalatalized *k* and *g,* were retained in Lithuanian (as in Old Prussian and Semigallian) but changed to *c* (pronounced like *ts*) and *dz* in Latvian (as in Selonian and Curonian): Lithuanian *ãkys* "eyes" (Old Prussian *ackis*) = Latvian *acis,* and Lithuanian *gérvė* "crane" (Old Prussian *gerwe*) = Latvian *dzer̃ve.* The change of the old clusters $t+j$ and $d+j$ progressed further in Latvian. Most Lithuanian dialects have *č* (as *ch* as in "church") and *dž* (as *j* in "jam"), whereas Latvian has *š* (as *sh* in "shore") and *ž* (as *z* in "azure")—*e.g.,* Lithuanian *trẽčias* = Latvian *trešs* "third"; Lithuanian *bríedžai* = Latvian *briẽži* "elks." Another difference between Lithuanian and Latvian is that, instead of Lithuanian *š* and *ž,* Latvian (like Selonian, Semigallian, Curonian, and Old Prussian) has *s* and *z* sounds—*e.g.,* Lithuanian *širdìs* = Latvian *sirds* "heart"; Lithuanian *žiemà* = Latvian *zìema* "winter." Proto-Latvian (and Prussian) $s+j$ and $z+j$ have passed to *š* and *ž*: Latvian *šũt* "to sew" = Lithuanian *siũti;* Latvian *eža* "of a hedgehog" (from Latvian **ezjā*) = Lithuanian *ežio.* Lithuanian has retained the initial clusters *pj* and *bj,* which in Latvian (and similarly in Slavic) have passed to *pļ* and *bļ*—*e.g.,* Lithuanian *piáuti* (*pi* is pronounced as *pj*) = Latvian *pļaũt* "to cut"; Lithuanian *biaurùs* = Latvian *bļaũrs* "hideous, nasty."

Stress and tone in Lithuanian and Latvian

Lithuanian has a free stress in contrast to Latvian fixed stress, which occurs on the first syllable. Latvian is more archaic than Lithuanian in the intonations inherited from Proto-Baltic: the Proto-Baltic circumflex intonation has preserved its falling character in Latvian (it became rising in Lithuanian), and the Proto-Baltic acute intonation retained its rising character (it is falling in Lithuanian), although in some cases (because of stress retraction) it has been changed to the broken intonation; *e.g.,* Latvian *pìrsts* "finger" = Lithuanian *pir̃štas* (falling in Latvian and rising in Lithuanian from the Proto-Baltic circumflex), Latvian *vãrna* "crow" = Lithuanian *várna* (the rising or extended intonation in Latvian and the falling intonation in Lithuanian from the Proto-Baltic acute intonation), Latvian *zâle* "grass" (the Latvian broken intonation from the Proto-Baltic acute intonation through stress retraction).

There are really no differences in the older morphological features between Lithuanian and Latvian if one does not take into account innovations such as the Latvian debitive verb form (*man ir jāmācās* "I must study" or "it is necessary for me to study") and the Lithuanian frequentative past (*jie eidavo* "they used to go"). Lithuanian and Latvian have two grammatical genders (masculine and feminine) and two numbers (singular and plural), while some Lithuanian dialects also have the dual number. Both Lithuanian and Latvian have seven cases—nominative, genitive, dative, accusative, instrumental, locative, vocative. Standard Lithuanian has five declensions of nouns with 12 inflectional types; Latvian has six declensions with eight inflectional types. Lithuanian adjectives have three declensions, Latvian adjectives have one. The comparison of adjectives in the two languages is different. Both Lithuanian and Latvian have indefinite adjectives (Lithuanian *mãžas,* masculine, *mažà,* feminine, "a small one" = Latvian *mazs, maza*) and definite adjectives (Lithuanian *mažàsis, mažóji* "the small one" = Latvian *mazais, mazā*) with their own specific inflection. The verb in Lithuanian and Latvian has three conjugations (genetically different). There are three persons, the third of which is the same (apparently from the time of Proto-Indo-European) in both the singular and the plural (as well as the dual); for example:

Lithuanian		Latvian
Singular		Singular
1. *kertù*	("I cut, I strike")	1. *certu*
2. *kertì*	("you cut, you strike")	2. *certi*
3. *ker̃ta*	("he cuts, he strikes")	3. *cert*
Plural		Plural
1. *ker̃tame*	("we cut, we strike")	1. *certam*
2. *ker̃tate*	("you cut, you strike")	2. *certat*
3. *ker̃ta*	("they cut, they strike")	3. *cert*

Verb forms

The verb in Lithuanian and Latvian has three tenses (present, preterite, future)—*e.g.,* Lithuanian *kertù,* Latvian *certu* (present); Lithuanian *kirtaũ,* Latvian *cirtu* (preterite); Lithuanian *kir̃siu,* Latvian *ciršu* (future). In contrast to Latvian, Lithuanian also has a frequentative past tense—*e.g., kir̃sdavau* "I used to cut, strike." Lithuanian and Latvian have many compound tense forms, compounded from the forms of the verb *būti* "to be" and participles. There are several moods in both languages, although they are different. The system of participles (active and passive) in Lithuanian and Latvian is quite similar, although complicated—*e.g.,* Lithuanian *kertą̃s,* Latvian *certuošs* (present active); Lithuanian *ker̃tamas,* Latvian *certams* (present passive). Lithuanian and Latvian sentence word order is quite free, and, in general, the syntax of both languages is quite similar.

Words are formed in Lithuanian and Latvian basically by means of suffixes, prefixes, and compounding. The languages are very similar in their early vocabulary, and the differences that do occur tend to be more of a semantic nature—*e.g.,* Lithuanian *mṓša* "husband's sister" = Latvian *māsa* "sister"; Lithuanian *žam̃bas* "corner, angle (acute)" = Latvian *zùobs* "tooth." Some older lexical differences do occur, however (*e.g.,* Lithuanian *kraũjas* = Latvian *asins* "blood"; Lithuanian *sūnùs* = Latvian *dēls* "son"). In the newer vocabulary, there are now many differences between Lithuanian and Latvian.

Loanwords in Baltic. The Baltic languages have loanwords from the Slavic languages (*e.g.,* Old Prussian *curtis* "hunting dog," Lithuanian *kùrtas,* Latvian *kur̃ts* come from Slavic [*cf.* Polish *chart*]; Lithuanian *muĩlas* "soap" [*cf.* Russian *mylo*]; Latvian *suods* "punishment, penalty" [*cf.* Russian *sud*]). There are also a few loanwords from Gothic (*e.g.,* Old Prussian *ylo* "awl," Lithuanian *ýla,* Latvian *īlens*) and possibly from Scandinavian, and many from German, especially in Old Prussian and Latvian, as a consequence of the German colonization of the Prussians, Latvians, and, in part, of the Lithuanians in the 13th century.

Relation between the Baltic and Finnic languages

The Balts first came in close contact with their northern neighbours, the Baltic Finns, about 2000 BC. This contact left traces in both the Baltic and the Finnic languages, perhaps most clearly in the vocabulary. Baltic has very few early loanwords from Finnic, but Finnic has many early loans from Baltic. Latvian, with many loanwords from Livian (Livonian) and Estonian (both Finnic languages), has been more influenced by Finnic than has any other recorded Baltic language.

Orthography. The Lithuanian alphabet is based on the roman (Latin) alphabet. It has 33 letters, several employing diacritical marks (*ą, č, ę, ė, į, š, ų, ū, ž*), and is phonetic (*i.e.,* written as it is pronounced). In linguistic literature ′ is used for falling tones, and ~ for rising tones; the grave accent (`) is used for short, stressed vowels. The Latvian alphabet has 33 letters, 11 with diacritical marks: *ā, č, ē, ģ, ī, ķ, ļ, ņ, š, ū, ž.* A macron (-) over a vowel indicates that it is long. In linguistic literature the following accents are used for the Latvian intonations: ` (falling), ~ (extended, or rising), ^ (broken).

The Old Prussian orthography is almost wholly based on the German orthography of that time and is quite inconsistent. Furthermore, every Old Prussian written record has its own specific orthography. (V.J.M.)

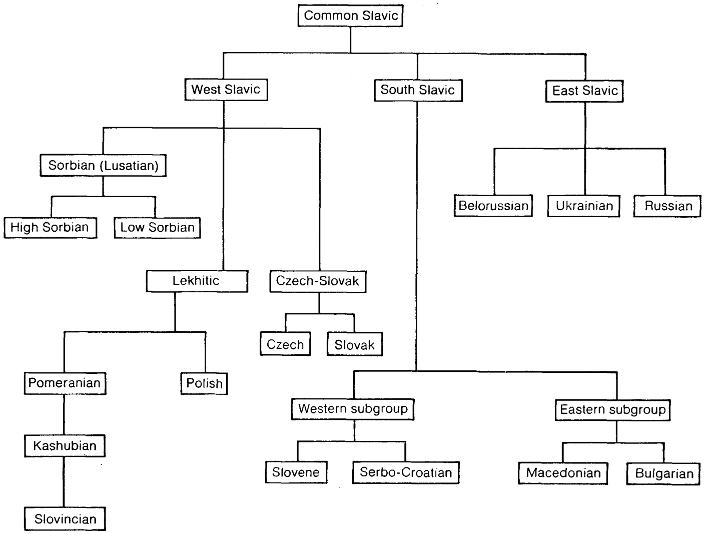
Slavic languages

The Slavic, or Slavonic, languages, constituting a separate branch of the Indo-European language family, are closer to the Baltic languages than to any other Indo-European subgroup, but they share certain linguistic innovations with the other eastern Indo-European languages (such as Indo-Iranian and Armenian) as well. From their original area situated between the Oder and the Dnepr rivers, the Slavic languages have spread to the territory of the Balkans (Bulgarian, Serbo-Croatian), central Europe (Czech and Slovak), eastern Europe (Polish, Ukrainian, Russian), and the northern parts of Asia (Russian). The number of native speakers for the entire branch is about 268,000,000. In addition, Russian is used as a second language by most inhabitants of the Soviet Union. Some of the Slavic languages have been used by writers of worldwide significance (*e.g.,* Russian, Polish, Czech) and the Church Slavonic language is an important means of communication within the Eastern Orthodox Church.

LANGUAGES OF THE FAMILY

Three branches of Slavic

The Slavic language group is divided schematically into three branches: the South Slavic branch, with two subgroups—Serbo-Croatian-Slovene and Bulgarian-Macedonian; the West Slavic branch, with three subgroups—Czech-Slovak, Sorbian, and Lekhitic (Polish and related tongues); and the East Slavic branch, comprising Russian, Ukrainian, and Belorussian.



In the spoken Slavic dialects (as opposed to the sharply differentiated literary languages) the linguistic frontiers are not always apparent. There are several transitional dialects and mixed forms of speech that connect the different languages, with the exception of the area where the South Slavs are separated from the other Slavs by the non-Slavic Romanians, Hungarians, and German-speaking Austrians. But even in this latter domain, some vestiges of the old dialectal continuity (between Slovene and Serbo-Croatian, on the one hand, and Czech and Slovak, on the other) that was later interrupted can be traced; the same traces of the old links are seen in comparing Bulgarian and Russian dialects. Thus the traditional schematic division of the Slavic group into three separate branches is not to be taken as the real model of historical development. It would be more realistic to represent the historical development as a process in which tendencies to differentiate and to reintegrate the cognate dialects have been continuously at work, bringing about the remarkable degree of uniformity in the different dialects. Still it would be an exaggeration to suppose that communication between any two Slavs is possible without any linguistic complications. The myriad differences between the dialects and languages in phonological and phonetic realization as well as in the spheres of semantics and morphology may cause misunderstandings even in the simplest of conversations; and the difficulties are greater in the language of belles lettres, even in the case of closely connected languages. For a Slav to master effectively a second Slavic language demands time and work.

South Slavic. Bulgarian is spoken by more than 8,300,-

000 people in Bulgaria and adjacent areas of other Balkan countries and the Soviet Union. There are two major groups of Bulgarian dialects: an Eastern one that became the basis of the literary language in the middle of the 19th century and a Western one that influenced the literary language. Bulgarian texts prepared before the 16th century were written mostly in an archaic language that preserved some features of Old Bulgarian (10th to 11th centuries) and Middle Bulgarian (beginning in the 12th century).

Although the vocabulary and grammar of the early texts written in the Old Church Slavonic language include some Old Bulgarian features, the language was nevertheless based originally on a Macedonian dialect. Old Church Slavonic was the first Slavic language to be put down in written form, by SS. Cyril and Methodius. The modern Macedonian language, spoken by about 1,500,000 people in Yugoslavia and Greece, was the last Slavic language to attain a standard literary form; during World War II, its central dialects of Prilep and Veles were elevated to this status. The Central Macedonian dialect is closer to Bulgarian, while the Northern dialect shares some features with the Serbo-Croatian language. Modern Macedonian dialects may be considered as links between the Eastern (Bulgarian) subgroup of South Slavic and the Western (Serbo-Croatian) subgroup.

Dialects of Serbo-Croatian

The Western subgroup of South Slavic includes the dialects of Serbo-Croatian, among them those of the Prizren-Timok group, which are close to some North Macedonian and West Bulgarian dialects. The literary Serbo-Croatian language was formed in the first half of the 19th century on the basis of the Shtokavian dialects that extend over the greater part of the Serbo-Croatian territory in Yugoslavia. These dialects are called Shtokavian because they use the form *što* (*shto*) for the interrogative pronoun "what?". They are distinguished from the Chakavian dialects of western Croatia, Istria, the coast of Dalmatia (where a literature in that dialect developed in the 15th century), and of some islands in the Adriatic; in those areas *ča* (*cha*) is the form for "what?". A third main group of Serbo-Croatian dialects, spoken in northwestern Croatia, uses *kaj* rather than *što* or *ča* and is therefore called Kajkavian. In all, more than 18,000,000 people speak Serbo-Croatian.

The Slovene language is spoken by more than 2,000,000 persons in the Socialist Republic of Slovenia in federal Yugoslavia and in the adjacent areas of Italy and Austria. It has some features in common with the Kajkavian dialects of Croatia and includes many dialects with great variations between them. In Slovene (particularly its Western and Northwestern dialects) some traces can be found of old links with the West Slavic languages (Czech and Slovak).

West Slavic. To the West Slavic branch belong Polish and some remnants of other Lekhitic languages (Kashubian and its archaic variant Slovincian), Low and High Sorbian (also called Lusatian or Wendish), Czech, and Slovak. Approximately 40,000,000 people speak Polish in Poland, in some regions of Czechoslovakia and the Soviet Union, and in France, the United States, and Canada. The main Polish dialects are Great Polish (in the northwest), Little Polish (in the southeast), Silesian, and Mazovian. The last dialect shares some features with Kashubian. There are about 210,000 native speakers of Kashubian remaining in Poland on the left bank of the Lower Vistula River. Slovincian belongs to the Northern group of Kashubian dialects, which is distinguished from a Southern group. Kashubian dialects (including Slovincian) are considered to be the remnants of a Pomeranian subgroup that belonged to the Lekhitic group. Lekhitic also included Polabian, which was spoken up to the 17th and 18th centuries by the Slavic population of the Elbe (Labe) River region. (At that time a dictionary and some phrases in the language were written down.)

The Polabian language bordered the Sorbian dialects, which are still spoken by about 140,000 inhabitants of Lower Lusatia and Upper Lusatia in East Germany. There are three main groups of Sorbian dialects: High Sorbian (Upper Sorbian), one of whose dialects in the area of Bautzen (Budyšin) is the basis of the literary language; Low Sorbian (or Lower Sorbian); and East Sorbian, the remnants of which are spoken in the area of Muskau (Mužakow).

Czech is spoken by about 9,800,000 people in the western part of Czechoslovakia (Bohemia, Moravia, and Silesia); its dialects are divided into Bohemian, Moravian, and Silesian groups. The literary language is based on the Central Bohemian dialect of Prague. The Slovak literary language was formed on the basis of a Central Slovak dialect in the middle of the 19th century. Western Slovak dialects are close to Moravian and differ from the Central and the Eastern dialects, which have features in common with Polish and Ukrainian. More than 4,600,000 people speak Slovak; they are located mostly in Slovakia.

Russian, Ukrainian, Belorussian

East Slavic. Russian, Ukrainian, and Belorussian (White-Russian) comprise the East Slavic language group. Russian is the native language of about 139,300,000 people and is widely used by many others in the Soviet Union and in some other eastern European countries. Its main dialects are divided into a Northern Great Russian group, a Southern Great Russian group, and a transitional Central group.

Ukrainian dialects are classified into Northern, Southeastern, Southwestern, and Carpathian divisions (the last group having features in common with Slovak); the literary language is based on the Kiev-Poltava dialect. More than 42,700,000 people speak Ukrainian in the Soviet Ukraine, and there are more than 580,000 Ukrainian speakers in Canada and the United States.

More than 9,600,000 people speak Belorussian in the Belorussian Soviet Socialist Republic. Its main dialectal groups are Northwestern Belorussian, some features of which may be explained by contact with Polish, and Northeastern Belorussian. The dialect of Minsk, which served as a basis for the literary language, lay on the border between these two groups.

HISTORICAL SURVEY

The Slavic protolanguage. Each branch of Slavic originally developed from a dialect of Proto-Slavic, the ancestral parent language of the group, which in turn developed from an earlier language that was also the antecedent of the Proto-Baltic language. Both Slavic and Baltic share with the eastern Indo-European languages (called satem languages) the same change of Indo-European palatialized *ḱ* and *g* sounds (consonants modified by bringing the front of the tongue up to or toward the hard palate) into spirants of the *s* and *z* type (*e.g.*, in Proto-Slavic **s3to* "hundred" contrasts with Latin *centum*, etc.). The Slavic and Baltic branches are distinguished by such innovations as: (1) the change of the old Indo-European syllabic *r, n,* and *m* (which functioned as vowels) to *ir, ur,* and related variants; (2) the same patterns of stress in nouns and verbs; and (3) the same reshuffling of the verbal system to produce two forms of the past tense in *-ā* and *-ē*.

Domain of Proto-Balto-Slavic

Some scholars believe that, after the common Indo-European area had been divided into different dialect zones (approximately after the 3rd millennium BC), a protodialect developed in the Baltic and Slavic areas that had many features peculiar to only these two branches of Indo-European. At the same time this protodialect was connected with certain western Indo-European protodialects called Old European that are identified as the source of a number of river names. The ancient Baltic and Slavic names of rivers (hydronyms), such as the Russian Oka, are of the same type as those found in the central European area. The dialects of the Slavic protolanguage spoken near the Carpathian Mountains in the upper Vistula area may have been part of the intermediate zone situated between the western Indo-European dialects (Germanic, Celtic, Italic, and so on) and the eastern Indo-European ones; in addition to Baltic and Slavic in the north, this intermediate zone included the Indo-European languages of the Balkans (Illyrian, Thracian, Phrygian). The domain of the Proto-Balto-Slavic dialect may have been situated to the east of the Germanic and other Old European dialects, to the north of Ancient Balkanic (including Illyrian), and to the west of Tocharian. The exact geographical borders of the Balto-Slavic domain appear impossible to determine, but they may well have been located in eastern Europe

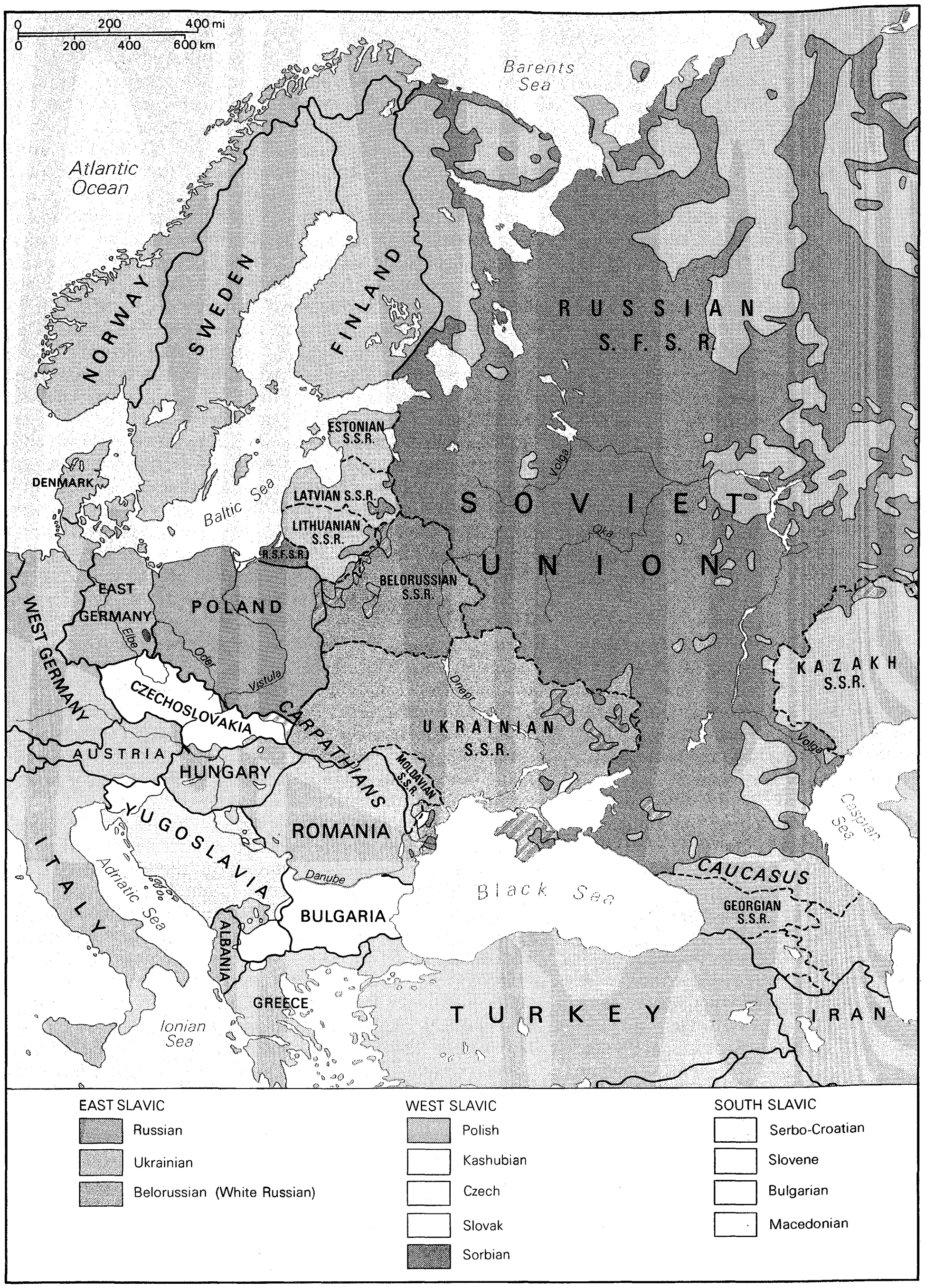


Figure 16: Distribution of the Slavic languages in Europe.

around present-day Lithuania and to the east and south of it. The later diffusion of Slavic languages southward into the Carpathian region may represent the spread of one of the dialects of this Old Baltic domain. The oldest Slavic protolanguage may be described as the result of the historical transformation of the Baltic protolanguage (but not vice-versa).

Until the middle of the 1st millennium AD, the Slavs were known to other people as the inhabitants of the vast territories between the Dnepr and the Vistula. In the 6th century AD the Slavs expanded to the Elbe (Labe) River and the Adriatic Sea and across the Danube River to the Peloponnese. In that period, according to the oldest references of Greek and Latin sources about the Slavs, they were already divided into several groups. The Slavic language, however, was uniform in its phonological and grammatical structure, with important dialectal variations occurring only in the vocabulary. The main phonological difference between the oldest pattern common to Baltic and Slavic and the later one that characterized Slavic alone was that in Slavic all syllables became open (*i.e.*, a syllable could end only in a vowel). Thus, all consonants at the end of a syllable were lost. This led to a reshuffling of most of the inflectional endings.

The next period in Slavic linguistic history began with the loss of the reduced vowels, called *yers* (*jeri*), that resulted from Indo-European short *ĭ* and *ŭ*; this loss caused a wide-ranging change in many words and forms. Although this process was common to all the Slavic dialects, which were still connected with each other at that period, it took place slowly and at different rates in different dialects, beginning in the 10th to the 12th century and expanding from the southwest to the northeast. After the loss of the *yers*, which gave different results in different dialectal groups (see Table 31), the uniformity of the Slavic language area began to disappear, and separate branches and languages emerged. An important clue to the date of the dissolution of Slavic unity is the separate development in different Slavic dialects of the name of the emperor Charlemagne (742–814). This name must have entered into Slavic in the postulated form **korljĭ* ("Karl") before the dissolution took place. Subsequently the proper name became the generic term for "king." The segment *-or-* in the postulated form appears differently in the modern Slavic languages—compare Bulgarian *kral,* Serbo-Croatian *kralj,* Slovene *králj* (*i.e.,* South Slavic *-ra-*), Russian *korol* (*i.e.,* East Slavic *-oro-*), Czech *král,* Polish *król.*

Emergence and early development of the Slavic languages. The separate development of South Slavic was caused by a break in the links between the Balkan and the West Slavic groups that resulted from the settling of the Magyars in Hungary during the 10th century and from the Germanization of the Slavic regions of Bavaria and Austria. Some features common to Slovak and Slovene may possibly have developed before the West-South break.

Disruption between South Slavic and West Slavic groups

The eastward expansion of dialects of Balkan Romanian (a Romance language) led to a break in the connection between the South and the East Slavic groups around the 11th–12th centuries. The history of the Balkan Slavs was closely connected with Byzantium, in contrast to that of the Lekhitic and Sorbian subgroups of the Western Slavs, which was connected with west European culture.

An effort on the part of the Slavs to counteract the influence of the Western Christian Church (which was associated with the German Empire) was the motive behind the introduction of the Old Church Slavonic language into the liturgy in Great Moravia, the first Slavic national state. Founded in the 9th century, Great Moravia united different groups speaking West Slavic dialects. In 863 its prince, Rostislav, invited St. Cyril (Constantine) and his brother St. Methodius to create a national church with a language and writing of its own. Prior to that time some Christian texts in Moravia had been translated into Slavic from Latin (and partly perhaps from Old High German); these have been preserved only in later copies.

The second period in the history of the Old Church Slavonic language (893–1081) occurred in the Bulgarian kingdoms of Symeon (893–927) and Peter (927–969) and in the kingdom of Samuel (997–1014), and was connected with the literary activity of many Bulgarian scholars who translated numerous Greek texts into Slavic and also produced a small number of original works. In the writings of the period of Symeon and Peter, Western (Macedonian) features were substituted for Eastern Bulgarian ones.

Reflection of Old Church Slavonic variants in various manuscripts

Development of the individual Slavic languages. Both the Western (Macedonian) and Eastern (Bulgarian) variants (recensions) of the Old Church Slavonic language are preserved in manuscripts of the 11th century, while the East Slavic (Russian) variant is reflected in the oldest dated Old Church Slavonic manuscript, Ostromir's Evangelium (1056), and in many later texts. The Moravian variant must be reconstructed on the basis of some later texts (such as the Kiev fragments of the beginning of the 11th century) dating from after the break with the Great Moravian tradition. In some documents of the 10th and 11th centuries, the Bohemian variant (sharing some West Slavic peculiarities with Moravian) has been preserved. Several features are common to the Moravian and Bohemian varieties of the Old Church Slavonic language, to the Slovene (Pannonian) variant reflected in the Freising fragments (late 10th century), and to the Croatian Old Church Slavonic tradition that is attested from the 12th century, as well as to the Serbian tradition. All these variants of the Old Church Slavonic language have some peculiarities that are to be explained as the result of the interaction of the original system with that of a local dialect. At about 1000 all Slavic languages were so close to one another that such interaction was possible.

From these local variants of the Old Church Slavonic language that are preserved in the manuscripts of the 10th to the 12th centuries, one should distinguish the later local Church Slavonic languages (Russian, with its different variants; Middle Bulgarian; Serbian, which was replaced by the Russian variant in Serbia in the 18th century; Croatian; and the Romanian variant of Church Slavonic, which was used as a literary and church language of Romania from the 14th to the 18th centuries). From the linguistic point of view, these later Church Slavonic literary languages differ from the earlier varieties in their systems of vowels; the early nasalized vowels were replaced by different later reflexes, and the reduced vowels (*yers*), with the exception of those followed by a syllable containing another *yer,* were generally lost. These changes in the sound pattern were accompanied by changes in vocabulary that were the result of cultural factors.

After the schism between the Eastern (Orthodox) and Western (Roman) Christian churches in the 11th century and the beginning of the Crusades, the Church Slavonic language fell out of use in all West Slavic countries and in the western part of the Balkan Slavic region. The only exception was the renaissance of Croatian Church Slavonic in the 13th century. At the end of the same century, the first Czech verses in the local dialect were written; they were the precursors of the rich poetic literature in the Old Czech language that appeared in the 14th century. The early Czech literary language was marked by the influence of Latin, which had replaced the Bohemian variety of Old Church Slavonic as a literary language.

In the earliest period of its development, the Polish literary language was modelled on the Czech pattern. After the Christianization of Poland, Latin (and later German) loanwords entered the Polish language in their Czech form. The Czech influence is seen in the Polish literary language until the 16th century (the "Golden Age"), when Renaissance tendencies resulted in the creation of texts that had aesthetic merit and were at the same time stylistically close to everyday speech. Later on the Polish literary language was enriched by cross-fertilization with Ukrainian and Belorussian.

In the 16th century in Dalmatia, a rich poetic literature in Croatian was created by poets who were influenced by the Italian Renaissance and who also wrote in Italian and Latin. A Slovene translation of the Bible was published in 1575–84 and Kashubian and Sorbian religious texts were also produced during this period. The comparatively early rise of the West Slavic (and western South Slavic) languages as separate literary vehicles was related to religious and political factors that resulted in the decline of the western variants of the Church Slavonic language.

In contrast, the continuing use of Bulgarian Church Slavonic and different variants of Russian Church Slavonic made it difficult to construct literary languages for Bulgarian and Russian based on everyday speech. Bulgarian texts were written in Bulgarian Church Slavonic until the 16th century. After that the language of the so-called Damaskinar (Damascene) literature was developed, closer to the popular speech; its development, however, was hampered under Turkish rule. Most of the Old East Slavic (Old Russian) literary texts were written in a mixture of Russian Church Slavonic and the Old Russian vernacular language; only a few documents, particularly some parts of the chronicles (annals), are written entirely in Old Russian. The proportion of South Slavic (Church Slavonic) and East Slavic (Old Russian) elements in each text is different depending on its stylistic peculiarities.

Linguistic results of religious schism

In the middle of the 17th century, the old Great Russian variant of the Church Slavonic language in the official Russian Orthodox Church was replaced by a new variant taken from the southwestern East Slavic tradition, a form that incorporated some Ukrainian and Belorussian elements. This development was connected with a split in the Orthodox Russian Church; the Old Believers, who split off from the main body of the church, continued to use the archaic Great Russian variant, while Patriarch Nikon's new variant, based on the southwestern tradition, was adopted by the official church and is used in it to this day. Because the Ukrainian tradition includes many West Slavic elements, this reform, which occurred after the incorporation of the Ukraine into the Russian Empire, was a step in the direction of the Westernization of the Russian language that took place soon after Peter the Great began his attempts to reconstruct and Westernize the whole Russian way of life.

In the 18th and 19th centuries, many waves of loanwords from different Western languages entered the Russian language. While some of the syntactic structures earlier had been formed on Germanic and Latin patterns, many western European semantic characteristics penetrated into the Russian language as a result of the intensive French–Russian bilingualism of the Russian upper classes at the end of the 18th and the beginning of the 19th century. The great Russian literature of the 19th and the early 20th century (up until Tolstoy's death in 1910) created a literary language close to everyday speech, especially to that of the villages. In the official style of Russian, however, Church Slavonic elements still dominate, as can be seen even in general newspaper articles.

The concept of a language that would unite all the Slavs has remained in the back of the Slavic consciousness, not as a real aim but rather as an important symbol. The most interesting attempt to unite different chronological and local Slavic strata was carried out by the Serbian Romantic writer Vuk Stefanović Karadžić. In modern literature one might cite the experiments at unification of Velemir Khlebnikov, a Russian futurist writer, and of the Polish poet Julian Tuwim, who created a Polish–Russian symphony of sounds in some of his poems.

Standardization of the modern Slavic languages. Among the Slavic languages that attained their standard literary form at a later stage in Slavic history than those mentioned above is Ukrainian. It was used in some literary texts in the late 18th century and in turn influenced the language of Gogol, one of the greatest Russian writers of the 19th century, to the extent that the language of some parts of Gogol's early texts may be described as a mixed Russian–Ukrainian dialect. In the 19th century and especially in the first decades of the 20th century, a number of great poets wrote in Ukrainian. The movement toward national liberation led to the introduction of many neologisms into the language, which persisted even after some attempts were made toward the artificial unification of the East Slavic area. After World War I, the Belorussian language became a standard language in the Belorussian Soviet Socialist Republic.

Status of modern Slavic languages

Since World Wars I and II, all of the Slavic languages have acquired the status of either the main language of an independent state (*e.g.,* Bulgarian and Polish) or the language of an autonomous part of a state (*e.g.,* Russian, Ukrainian, and Belorussian in the Soviet Union; Serbo-Croatian, Slovene, and Macedonian in Yugoslavia; Czech and Slovak in Czechoslovakia). Only the minor languages are exceptions: Kashubian is used officially only in some cultural performances, and Low and High Sorbian are now taught in schools in East Germany. The extent of dialectal variation in the different languages ranges from a very great degree in Slovene to a much smaller degree in some areas of such languages as Polish and Russian. Radio and other mass media have been among the main influences leading to linguistic consolidation. Languages like Polish, Czech, and Russian, which have served as a basis for great literatures, have become models for others that are only now being put to literary use (although for such languages as Kashubian and, to some degree, for High and Low Sorbian, the folk literature remains much more important as a model than individual literary works and translations of past centuries).

LINGUISTIC CHARACTERISTICS

A number of distinguishing characteristics set off the Slavic languages from other Indo-European subgroups. On the whole, Slavic auxiliary words tend to be unstressed and to be incorporated into a single phonetic group or phrase with autonomous stressed words. Inflection (*i.e.,* the use of endings, prefixes, and vowel alternations) has persisted as the main method of differentiating grammatical meanings, although to a lesser degree in the nouns than in the verbs because many functions of the noun case endings may also be expressed by prepositions. Verbal categories have retained their complex archaic character. The movable stress pattern common to most South and East Slavic languages has had a profound influence on versification in these languages.

Many linguistic devices found both in the oral tradition and in individual literary works of the different Slavic languages may be traced to common ancestral forms. An exuberant use of diminutives and metaphoric figures marks the Slavic oral tradition. It seems possible to reconstruct a common Proto-Slavic model of the universe as seen through linguistic expression. The main feature of such a model is the recurrence of binary (two-way) contrasts, as is evidenced by such cue words as *bog*Б "god" from "a portion allotted by the gods" and *ne-bog*Б "not having its portion, having bad fortune." Such pairing of opposites bears a striking resemblance to the ancient Iranian dualistic view of the world, a view that evidently influenced the Slavs to a degree not yet fully appreciated.

Influence of other cultures on Slavic vocabularies

As compared with the common Indo-European scheme, the pre-Slavic cultural linguistic heritage seems in some degree simpler, evidently as a result of the loss of direct contact with the Southern civilizations that served as a pattern for pre-Indo-European culture. Later developments were caused largely by western European and Greek (particularly Byzantine Christian) influences and by contacts with eastern cultures, which led to innovations in the vocabularies of the East Slavic and South Slavic languages. In some instances, whole series of designations for objects were borrowed into Russian and other East Slavic languages from eastern sources.

All Slavic languages are synthetic, expressing grammatical meaning through the use of affixes (suffixes and, in verbal forms, also prefixes), vowel alternations partly inherited from Indo-European, and consonant alternations resulting from linguistic processes peculiar to Slavic alone. Although analytical methods of expressing grammatical meanings (through prepositions and other "empty" grammatical words) are present in older strata of the language, they are used to the exclusion of all other means only in the case system of Modern Bulgarian and Macedonian. The tendency toward analytic expression is noticeable in contemporary everyday Russian speech, but the drift of the Slavic languages in this direction (as in the development of the western European languages) has been held back by the stabilization of the language resulting from mass communication and education.

Development of palatals, affricates, and spirants

Phonological characteristics. The Slavic systems of distinctive sounds are rich in consonants, particularly in spirants and affricates. This is especially true in comparison with the protolanguage and with other Indo-European languages. The affricates (which are consonant sounds begun as stops, with complete stoppage of the breathstream, and released as fricatives, with incomplete stoppage) have resulted historically from a succession of different processes of palatalization that have occurred in Slavic and are one of the most characteristic features of Slavic phonology. Palatalization is the process whereby the pronunciation of an originally nonpalatal sound is changed to a palatal sound by touching the hard palate with the tongue; it is also the process whereby a nonpalatal sound is modified by simultaneously moving the tongue up to or toward the hard palate. Originally, palatalization was connected with the adaptation of a consonant to the following vowel within a syllable, specifically, with the adaptation of a consonant to a following front vowel. This adaptation gave rise to "soft" (palatalized) syllables, composed of palatalized consonants followed by front vowels. The *j* sound, as *y* in English "year" (from older nonsyllabic Indo-European *i*), tended to palatalize the preceding consonant either by merging with it or by giving rise to consonant groups such as *bl'* from *bj* (*by*). As palatalized stop consonants (for instance *k', g', t', d'*) became increasingly differentiated from the corresponding nonpalatalized series (*k, g, t, d*), the palatalized stops tended to develop further into affricates (with the subsequent development of voiced affricates into spirants). Thus palatalized *k'* before the ancient front vowels developed into the affricate *č* (as *ch* in English "church"), and palatalized *g'* in the same environment changed to ʒ (as *j* in "judge"), which became the spirant sound *ž* (as *z* in "azure") in all Slavic languages. Before front vowels resulting from ancient diphthongs, palatalized *k'* changed to a *ts* sound, written as *c* (*e.g.,* Old Church Slavonic *cěna* "price," Serbo-Croatian *cijèna,* Russian *cena,* cognate to Lithuanian *káina*), and *g'* changed to a *dz* sound, which later changed to *z* (Old Church Slavonic [*d*]*zelo* "very," Old Czech *zielo,* Belorussian *do zěla,* cognate to Lithuanian *gailas*). The sounds *t'* (from *tj*) and *d'* (from *dj*) changed, respectively, into different stops, affricates, and

Table 29: Development of Proto-Slavic *tj, *dj

Proto-Slavic	Old Church Slavonic	South Slavic				East Slavic	West Slavic		
		Bulgarian	Macedonian	Serbo-Croatian	Slovene	Russian	Polish	Czech	High Sorbian
*tj > *t'	št	št	ḱ	ć	č	č	c	c	
*dj > *d'	žd	ǵ	đ	j	ž	ʒ	z		
*svetja "candle"	svešta	svešt	sveḱa	svijeća	sveča	sveča	świeca	svice	sweca
*medja "bound(ary)"	mežda	mežda	meǵa	međa	meja	meža	mieʒa (miedza)	meze	meza

*Indicates an unattested, reconstructed form.

spirants according to special rules in the separate Slavic dialects (see Table 29).

These processes of assibilation of the palatalized velar (*k', g'*) and dental (*t', d'*) sounds happened repeatedly in the history of the individual Slavic languages. Palatalization (softness) as a distinctive feature of most consonant sounds has been preserved in East Slavic; for example, in Modern Russian palatalized (or soft) *t', d', s', z'* contrast with nonpalatalized (or hard) *t, d, s, z.* (The contrast between the palatalized *k'* and the hard *k* is just now in the process of development.) Some West Slavic languages also have this contrast of palatalized and nonpalatalized consonants, while others do not. Czech, Slovak, and Serbo-Croatian, which have the usual three sets of labial, dental, and velar consonants inherited from the protolanguage, have developed a special, additional series of palatal stops. In all of the Slavic languages, voiced stop consonants (pronounced with vibrating vocal cords) contrast with voiceless stop consonants (pronounced without vibrating vocal cords).

Slavic vowel systems

The tendency to increase the number of different spirants (nonstops) is connected with the processes of palatalization. In the Ukrainian and the Southern Russian dialects and in Belorussian, Czech, Slovak, High Sorbian, and some Slovene dialects there also developed a voiced velar spirant sound, corresponding to the voiceless velar spirant of the Proto-Slavic language. The nasal vowels ę and ǫ that had developed in Proto-Slavic from older combinations of vowels with nasal consonants (still retained in Baltic) have been preserved only in some Lekhitic languages and in some South Slavic dialects, especially those of Slovene (see Table 30). The vowel systems are especially rich in those Slavic languages that have preserved prosodic differences in pitch (tone) and quantity (length)—Serbo-Croatian, Slovene, and Northern Kashubian. The reshaping of the Slavic vowel systems is in large measure a result of the loss of the *yers,* which had different effects in different dialects (see Table 31).

Table 30: Development of Proto-Slavic Nasal Vowels Compared with Baltic

English translation	Lithuanian	Serbo-Croatian	Polish	Russian
"I spin"	sprendžiu	prédēm	przędę	pryadu
"snipe, woodcock"	slánka	šljuka	słanka	sluka (obs.)
"soft"	minkštas	mȅk	miękki	myagky
"wise"	mañdras	mȕdar	mądry	mudry

Prosodic differences in vowel quantity have been preserved in Czech and Slovak, in which new vowels developed as a result of contraction. A fixed stress accent is found in the West Slavic languages as well as the Western and Central Macedonian dialects, in contrast to Proto-Slavic, Serbo-Croatian, Slovene, Bulgarian, and the East Slavic languages. In Czech and Slovak, as well as in Sorbian and Southern Kashubian, stress is fixed on the first syllable of the word, but in Polish, Eastern Slovak, and Southern Macedonian, it falls on the next to the last syllable of the word, while in Western Macedonian it falls on the third syllable from the end. The Slavic languages with a nonfixed placement of stress reflect the Proto-Slavic (and Indo-European) distinction between two types of noun and verb paradigms: (1) the paradigm with movable stress in which the stress (indicated here by ′) falls on the root in some forms and on the inflectional ending in others (*e.g.,* "head" in Russian is *golová* in the nominative case and *gólovu* in the accusative, these forms derive from Proto-Slavic **golvá,* **gólvǫ* [an asterisk indicates an unattested, reconstructed form]) and (2) the paradigm in which the stress is fixed on the stem (*e.g.,* "willow" in Russian is *íva* in the nominative case, *ívu* in the accusative, from **íva,* **ívǫ*).

The seven case forms

Grammatical characteristics. Most Slavic languages reflect the old Proto-Slavic pattern of seven case forms (nominative, genitive, dative, accusative, locative, instrumental, vocative), which occurred in both the singular and the plural. In the dual number the cases that were semantically closest to each other were represented by a single form (nominative–accusative, instrumental–dative, genitive–locative). The dual number is preserved today only in the westernmost area (*i.e.,* in Slovene and Sorbian) and in the archaic Slovincian language. The trend toward the modern, more analytical type of construction using prepositions and away from the synthetic type using case endings exclusively (as in Proto-Slavic and the archaic Slavic languages), is evident in the gradual elimination of the use of the dative and locative forms without prepositions. The end result of this development is seen in Bulgarian and Macedonian, in which noun declension has almost completely disappeared and has been replaced by syntactic combinations using prepositions. In the other South Slavic languages and in the western part of the West Slavic area (Sorbian and Czech), the same tendency to lose some of the distinctions between cases may be observed, but to a lesser degree. In the other West Slavic languages and in East Slavic, on the other hand, the old system of declension by case endings has been preserved in spite of the large number of loanwords and other neologisms that have no case distinctions at all (*e.g.,* borrowed Russian nouns like *kino* "cinema," or acronyms ending in a vowel like *Rayono* "district people education department").

The declension of pronouns has been preserved in all Slavic languages. The old combinations of adjectives with pronouns gave rise to the definite forms of adjectives. These forms still contrast with the indefinite forms in South Slavic, but in the other languages the indefinite forms have either been gradually lost or else have been preserved only to serve a special function, that of predicate noun. In Bularian and Macedonian, as well as in some northern East Slavic dialects, an article is used, placed after the noun (*e.g.,* in Macedonian, *knigata* "book-the"). The noun may be combined with either an article, as in the above example, or with a deictic (pointing) pronominal element—*e.g.,* in Macedonian, *kniga-va* "this book here,"

Table 31: Results of Loss of Yers

English translation	Proto-Slavic	Russian	Bulgarian	Macedonian	Serbo-Croatian	Czech	Polish	High Sorbian	Low Sorbian
"day"	dьnь < *dĭnĭ	den' (d'en')	den	den	dȃn	den	dzień	dzeń	źeń
"dream"	sъnъ < *sŭnŭ	son	sъn	son	sȁn	sen	sen	son	son

*Indicates an unattested, reconstructed form.

Table 32: The Russian Alphabet

Cyrillic letters				equivalent		approximate pronunciation
printed		written		EB preferred	Akademiya Nauk	
capital	lower-case	capital	lower-case			
А	а	*А*	*а*	a		f*a*ther
Б	б	*Б*	*б*	b		*b*a*b*y
В	в	*В*	*в*	v		*v*i*v*id
Г	г	*Г*	*г*	g		*g*o*
Д	д	*Д*	*д д*	d		*d*i*d*
Е	е	*Е*	*е*	e or ye†	e or je	b*e*t or *ye*t
Ё	ё	*Ё*	*ё*	o or yo‡	’o, o, or jo	*yo*re
Ж	ж	*Ж*	*ж*	zh	ž	a*z*ure
З	з	*З*	*з з*	z		*z*one
И	и	*И*	*и*	i‖	i *or* ji	mach*i*ne
Й	й	*Й*	*й*	y‖	j	bo*y*
К	к	*К*	*к*	k		*k*in
Л	л	*Л*	*л*	l		*l*i*l*y
М	м	*М*	*м*	m		*m*ai*m*
Н	н	*Н*	*н*	n		*n*o
О	о	*О*	*о*	o		*o*rder
П	п	*П*	*п*	p		*p*e*pp*er
Р	р	*Р*	*р*	r		e*rr*or (trilled)
С	с	*С*	*с*	s		*s*and
Т	т	*Т*	*т*	t		*t*ie
У	у	*У*	*у*	u		r*u*le
Ф	ф	*Ф*	*ф*	f		*f*i*f*ty
Х	х	*Х*	*х*	kh	ch	Ger. Bu*ch*
Ц	ц	*Ц*	*ц*	ts	c	ca*ts*
Ч	ч	*Ч*	*ч*	ch	č	*ch*in
Ш	ш	*Ш*	*ш*	sh	š	*sh*y
Щ	щ	*Щ*	*щ*	shch	šč	ra*sh ch*oice
Ъ	ъ		*ъ*	omit	”	§
Ы	ы		*ы*	y‖		rhythm
Ь	ь		*ь*	omit	’	¶
Э	э	*Э*	*э*	e		*e*cho
Ю	ю	*Ю*	*ю*	yu	’u or ju	*you*th
Я	я	*Я*	*я*	ya	’a or ja	*ya*rd

*Pronounced as *v* in genitive endings *-ego* and *-ogo*. †*e* after consonant; *ye* initially or after vowel, **ъ**, or **ь**. ‡o after **ж, ч, ш, щ**; yo elsewhere. §Hard sign; hardens preceding consonant or separates syllables. ‖In transliterating the adjectival endings **-ий** and **-ый**, the first vowel should be omitted. ¶Soft sign; softens or palatalizes preceding consonant or separates syllables.

or *kniga-na* "that book there." In addition to three noun genders (masculine, feminine, neuter), many Slavic languages distinguish animate and inanimate noun forms of some cases; and in some West Slavic languages, Ukrainian, and Bulgarian, personal and nonpersonal noun forms are differentiated.

Verb tenses

In the modern Slavic languages the verb is inflected to show present and past tenses. In the early history of the individual Slavic languages, however, a distinction was made between the aorist (originally differentiated as to voice, as in Baltic and Greek) and the imperfect; this distinction is still preserved in modern South Slavic (with the exception of Slovene) and Sorbian. (The aorist is a verb form indicating the occurrence of an action without reference to its completion, repetition, or duration; the imperfect is a verb tense designating a continuing state or an uncompleted action, especially in the past.) Slavic has almost no traces of the Indo-European old perfect tense, but, from combinations with the verbal noun in *-l*, new perfect forms were created that were differentiated from the pluperfects. Later these perfect forms came to be used as past tense forms in different Slavic languages. The most striking feature of the Slavic verb is the existence of paired stems, one of which expresses the perfective (completed) and the other the imperfective (uncompleted) aspects of the same verb—*e.g.,* Russian *dat* "to give" (*i.e.,* "to complete the process of giving"), *davat* "to be in the process of giving."

The present tense form of a perfective verb may be used to express future meaning, which can also be indicated by other means. In most South Slavic languages it is expressed by combinations (originally syntactic) with the verb "to want." The eastern South Slavic languages, Bulgarian and Macedonian, have lost the infinitive forms of the verb as a result of the influence of non-Slavic Balkan languages. These same Slavic languages have developed verb forms to differentiate between an action witnessed by the speaker and one not witnessed (hence only reported) by him.

A striking feature of Slavic syntax is the widespread use of possessive adjectives (*e.g.,* Russian *bozheskaya milost* "the mercy belonging to God") instead of the genitive case of the noun (*milost boga* "the mercy of God"). Word order in the Slavic languages is characterized by a gradual shift of the verb toward the medial position; originally the initial position was more characteristic. Other important features of Slavic syntax are related to this medial positioning of the verb and the consequent occurrence of the verb before the object. For example, grammatical elements are often placed before nouns; today they follow nouns only in some set phrases like Old Church Slavonic *boga radi* "for God's sake," with *radi* following the noun *boga* "God's."

Originally the verb occupied the initial position, which throws light on the origin of the reflexive verbal forms; these may be traced to the Proto-Slavic combination of the verb with a reflexive pronoun that occurred immediately after the verb and was pronounced as part of one accentual unit with the verb. Words other than pronouns that occurred immediately after verbs were also pronounced as a unit with them (these called enclitics).

The rules for the shift of the stress in syntactic combinations with enclitics were identical for verbs and nouns. Depending on the intonation of the word preceding the enclitic, the stress could be shifted either to the enclitic (as in Bulgarian *zimЪs* "this winter") or to the proclitic or preceding unstressed particle or word (as *na* in Serbo-Croatian *ná brijeg* and Russian *ná bereg* "on the shore").

Vocabulary. The original vocabulary of general terms common to Baltic and Slavic is still retained in most of the Slavic languages. In prehistoric times Proto-Slavic borrowed a number of important social and religious terms from Iranian (*e.g., bogЪ* "God," *mirЪ* "peace"). Later, special terms were borrowed by East Slavic and South Slavic from different eastern languages (especially Turkish) as a result of the political domination of the Tatars in Russia and of the Turks in the Balkan area. After the Renaissance, loanwords were taken from classical and western European languages (especially German and French) into all of the Slavic languages. Church Slavonic in its different variants remained the main source of innovations in vocabulary in East Slavic and in some South Slavic languages. The Slavic languages make extensive use of prefixes and suffixes to derive new words and thereby enrich the vocabulary; *e.g.,* Russian *čern-y* "black," *čern-i-t* "to blacken," *o-čern-i-t* "to slander." Several prefixes may be combined to modify the meaning of a verb (*e.g.,* Ukrainian *po-na-vi-pisuvati,* in which three prefixes are added to the verb "to write" to convey the meaning "to write out copiously"; Bulgarian *is-po-razboleyase,* in which the added prefixes intensify the meaning "to develop an illness"). Many derivational suffixes are common to most Slavic languages—*e.g.,* the very productive suffix *-stvo* (as in Russian *khristian-stvo* "Christianity," Ukrainian *pobratim-stvo* "fraternity," Polish *głup-stvo* "foolishness, trifle," Macedonian *golem-stvo* "high status, arrogance"). The archaic type of derivation by compounding, inherited from Indo-European, was particularly productive in Church Slavonic under the stimulus of Greek. Compounding remains one of the methods of creating new terms, especially technical terms (*e.g.,* Russian *vodokhranilishche* "reservoir" from *voda* "water" and *khranilishche* "depository"), but is far less important than affixation. Some Slavic languages typically derive new words by means of a condensed suffixing (*e.g.,* Czech *železnice* "railroad," from *železo* "iron" combined with a noun-forming suffix; *hledisko* "point of view," from

Borrowings from Iranian, eastern languages, and western European languages *hled* "look" combined with a noun-forming suffix), while others tend to use combinations of words (*e.g.*, Russian *železnodorožny* "railroad," from *zeleznaya doroga* "iron road" combined with an adjective-forming suffix).

Writing systems. The first writing system used for Slavic was the Glagolitic system invented by St. Cyril. Quite original in pattern, it reflected accurately the sound system of the Macedonian dialect. The forms of its letters can be traced to several different alphabets, mainly Greek and Semitic ones. Glagolitic was widely used in the first three centuries of Slavic literature but was gradually replaced by the Cyrillic alphabet, created in the 10th century and still used to write all the East Slavic languages, Bulgarian, Macedonian, and Serbian.

Other Slavic languages use the Latin (roman) alphabet. To render the distinctive sounds of a Slavic language, Latin letters are combined or diacritic signs are used (*e.g.*, Polish *sz* for the *sh* sound in "ship," Czech *č* for the *ch* sound in "church"). An orthographic system devised by the Czech religious reformer Jan Hus was adopted into different West Slavic systems of writing, including Czech, Slovak, and Sorbian. Polish spelling was patterned after the Czech spelling of the 14th century. Most of the Slavic writing systems are closely related to the sound patterns of the languages and are constructed to symbolize the distinctive sounds of the language or to render the same morphemes by the same groups of letters despite differences in pronunciation in various forms. Modern Russian spelling reflects a morpheme-based principle. (V.V.I.)

Albanian language

Albanian is an Indo-European language spoken by about 4,400,000 inhabitants of the eastern Adriatic coast in Albania and also in neighbouring Yugoslavia, principally in Kosovo and Makedonija (Macedonia), west of a line from near Leskovac to Lake Ohrid. There are perhaps 300,000 more speakers in isolated villages in southern Italy (Abruzzi, Molise, Basilicata, Puglia, and Calabria), and Sicily, and southern Greece (in Voiotía, Attica, Évvoia, Ândros, and the Pelopónnesos).

The origins of the general name Albanian, which traditionally referred to a restricted area in central Albania, and of the current official name Shqip or Shqipëri, which may well be derived from a term meaning "pronounce clearly, intelligibly," are still disputed. The name Albanian has been found in records since the time of Ptolemy. In Calabrian Albanian the name is Arbresh, in Modern Greek Arvanítis, and in Turkish Arnaut; the name must have been transmitted early through Greek speech.

Table 33: The Serbo-Croatian Alphabet

letters				letters			
Croatian		Serbian		Croatian		Serbian	
capital	lower-case	capital	lower-case	capital	lower-case	capital	lower-case
A	a	А	а	L	l	Л	л
B	b	Б	б	Lj	lj	Љ	љ
C	c	Ц	ц	M	m	М	м
Č	č	Ч	ч	N	n	Н	н
Ć	ć	Ћ	ћ	Nj	nj	Њ	њ
D	d	Д	д	O	o	О	о
Dž*	dž*	Џ	џ	P	p	П	п
Đ†	đ†	Ђ	ђ	R	r	Р	р
E	e	Е	е	S	s	С	с
F	f	Ф	ф	Š	š	Ш	ш
G	g	Г	г	T	t	Т	т
H	h	Х	х	U	u	У	у
I	i	И	и	V	v	В	в
J	j	Ј	ј	Z	z	З	з
K	k	К	к	Ž	ž	Ж	ж

*Alphabetized in *Britannica* as *dz*. †Alternatively, *dj*.

Table 34: The Bulgarian Alphabet

Cyrillic letters				equivalents	approximate pronunciation
printed		written			
capital	lower-case	capital	lower-case		
А	а	*А*	*а*	a	father
Б	б	*Б*	*б*	b	baby
В	в	*В*	*в*	v	vivid
Г	г	*Г*	*г*	g	go
Д	д	*Д*	*д*	d	did
Е	е	*Е*	*е*	e	bet
Ж	ж	*Ж*	*ж*	zh	azure
З	з	*З*	*з*	z	zone
И	и	*И*	*и*	i	machine
Й	й	*Й*	*й*	y	boy
К	к	*К*	*к*	k	kin
Л	л	*Л*	*л*	l	lily
М	м	*М*	*м*	m	maim
Н	н	*Н*	*н*	n	no
О	о	*О*	*о*	o	order
П	п	*П*	*п*	p	pepper
Р	р	*Р*	*р*	r	error (trilled)
С	с	*С*	*с*	s	sand
Т	т	*Т*	*т*	t	tie
У	у	*У*	*у*	u	rule
Ф	ф	*Ф*	*ф*	f	fifty
Х	х	*Х*	*х*	kh	Ger. Buch
Ц	ц	*Ц*	*ц*	ts	cats
Ч	ч	*Ч*	*ч*	ch	chin
Ш	ш	*Ш*	*ш*	sh	shy
Щ	щ	*Щ*	*щ*	sht	Ger. still
Ъ	ъ		*ъ**	ŭ	above
	ь		*ь*		†‡
Ю	ю	*Ю*	*ю*	yu	youth‡
Я	я	*Я*	*я*	ya	yard‡

*ŭ can occur initially in only a few words. †When transliterated, this is represented by an apostrophe. In modern orthography ' is used medially and finally for the *y* consonant before *o* and is pronounced *y*onder. ‡When *ya*, *yu*, and *'o* are preceded by consonants they are themselves pronounced like *a*, *u*, and *o*, respectively, but cause the preceding consonant to be palatalized. The softening is less great in Bulgarian than it is in Russian.

Gheg and Tosk dialects

Dialects. The two principal dialects, Gheg in the north and Tosk in the south, are separated roughly by the Shkumbin River. Gheg and Tosk have been diverging for at least a millennium, and their less extreme forms are mutually intelligible. Gheg has the more marked subvarieties, the most striking of which are the northernmost and eastern types, which include those of the city of Shkodër (Scutari), the neighbouring mountains along the Crna Gora border, Kosovo, Makedonija, and the isolated village of Arbanasi (formerly Borgo Erizzo) on the Croatian coast of Dalmatia outside Zadar. Arbanasi, founded in the early 18th century by refugees from near Bar (formerly Antivari), has about 2,000 speakers.

All of the Albanian dialects spoken in Italian and Greek enclaves are of the Tosk variety, and seem to be related most closely to the dialect of Çamëria in the extreme south of Albania. These dialects resulted from incompletely understood population movements of the 13th and 15th centuries. The Italian enclaves—nearly 50 scattered villages—probably were founded by emigrants from Turkish rule in Greece. A few isolated outlying dialects of south Tosk origin are spoken in Bulgaria and Turkish Thrace but are of unclear date. The language is still in use in Mandritsa, Bulgaria, at the border near Edirne, and in an offshoot of this village surviving in Mândres, near Kilkís in Greece, that dates from the Balkan Wars. A Tosk enclave near Melitopol in the Ukraine appears to be of moderately recent settlement from Bulgaria. The Albanian dialects of Istria, for which a text exists, and of Syrmia (Srem), for which there is none, have become extinct.

History. The official language, written in a standard roman-style orthography adopted in 1909, was based on the south Gheg dialect of Elbasan from the beginning of the Albanian state until World War II, and since has been modelled on Tosk. In Yugoslavia, Albanian speakers in the region of Kosovo in Serbia (officially bilingual in Serbian and Albanian) and in Macedonia speak eastern varieties of Gheg but since 1974 have widely adopted a common orthography with Albania. Before 1909, the little literature that was preserved, was written in local makeshift Italianate or Hellenizing orthographies, or even in Turko-Arabic characters.

The first written records

A few brief written records are preserved from the 15th century, the first being a baptismal formula from 1462. The scattering of books produced in the 16th and 17th centuries originated largely in the Gheg area (often in Scutarene north Gheg) and reflect Roman Catholic missionary activities. Much of the small stream of literature in the 19th century was produced by exiles. Perhaps the earliest purely literary work of any extent is the 18th-century poetry of Gjul Variboba, of the enclave at S.Giorgio, in Calabria. Some literary production continued through the 19th century in the Italian enclaves, but no similar activity is recorded in the Greek areas. All these early historical documents show a language that differs little from the current language. Because these documents from different regions and times exhibit marked dialect peculiarities, however, they often have a value for linguistic study that greatly outweighs their literary merit.

Classification. That Albanian is of clearly Indo-European origin was recognized by the German philologist, Franz Bopp, in 1854; the details of the main correspondences of Albanian with Indo-European languages were elaborated by another German philologist, Gustav Meyer, in the 1880s and 1890s. Further linguistic refinements were presented by the Danish linguist Holger Pedersen and the Austrian Norbert Jokl. The following etymologies illustrate the relationship of Albanian to Indo-European (an asterisk preceding a word denotes an unattested, hypothetical Indo-European parent word, which is written in a conventionalized orthography): *pesë* "five" (from **pénk*^{w}*e*); *zjarm* "fire" (from **g*^{w}*hermos*); *natë* "night" (from **nok*^{w}*t-*); *dhëndër* "son-in-law" (from **ǵemə ter-*); *gjarpër* "snake" (from **sérpŏ̄n-*); *bjer* "bring!" (from **bhere*); *djeg* "I burn" (from **dheg*^{w}*hō*); *kam* "I have" (from **kapmi*); *pata* "I had" (from **pot-*); *pjek* "I roast" (from **pek*^{w}*ō*); *thom, thotë* "I say, he says" (from **k'ēmi*, **k'ēt* . . .).

The verb system includes many archaic traits, such as the retention of distinct active and middle personal endings (as in Greek) and the change of a stem vowel *e* in the present to *o* (from **ē*) in the past tense, a feature shared with the Baltic languages. For example, there is *mbledh* "gathers (transitive)" as well as *mblidhet* "gathers (intransitive), is gathered" in the present tense, and *mblodha* "I gathered" with an *o* in the past. Because of the superficial changes in the phonetic shape of the language over 2,000 years and because of the borrowing of words from neighbouring cultures, the continuity of the Indo-European heritage in Albanian has been underrated.

Albanian shows no obvious close affinity to any other Indo-European language; it is plainly the sole modern survivor of its own subgroup. It seems likely, however, that in very early times the Balto-Slavic group was its nearest of kin. Of ancient languages, both Dacian (or Daco-Mysian) and Illyrian have been tentatively considered its ancestor or nearest relative.

Grammar. The grammatical categories of Albanian are much like those of other European languages. Nouns show overt gender, number, and three or four cases. An unusual feature is that nouns are further inflected obligatorily with suffixes to show definite or indefinite meaning; *e.g., bukë* "bread," *buka* "the bread." Adjectives—except numerals and certain quantifying expressions—and dependent nouns follow the noun they modify; and they are remarkable in requiring a particle preceding them that agrees with the noun. Thus, in *një burrë i madh,* meaning "a big man," *burrë* "man" is modified by *madh* "big," which is preceded by *i,* which agrees with the term for "man"; likewise, in *dy burra të mëdhenj* "two big men," *mëdhenj,* the plural masculine form for "big," follows the noun *burra* "men" and is preceded by a particle *të* that agrees with the noun. Verbs have roughly the number and variety of forms found in French or Italian and are quite irregular in forming their stems. Noun plurals are also notable for the irregularity of a large number of them. When a definite noun or one taken as already known is the direct object of the sentence, a pronoun in the objective case that repeats this information must also be inserted in the verb phrase; *e.g., i-a dhashë librin atij* is literally "him-it I-gave the-book to-him," which in standard English would be "I gave the book to him." In general, the grammar and formal distinctions of Albanian are reminiscent of Modern Greek and the Romance languages, especially of Romanian. The sounds suggest Hungarian or Greek, but Gheg with its nasal vowels strikes the ear as distinctive.

Vocabulary and contacts. Although Albanian has a host of borrowings from its neighbours, it shows exceedingly few evidences of contact with ancient Greek; one such is the Gheg *mokën* (Tosk *mokër*) "millstone," from the Greek *mēkhanḗ.* Obviously close contacts with the Romans gave many Latin loans; *e.g., mik* "friend," from Latin *amicus; këndoj* "sing, read" from *cantāre.* Furthermore, such loanwords in Albanian attest to the similarities in development of the Latin spoken in the Balkans and of Romanian, a Balkan Romance tongue. For example, Latin *palūdem* "swamp" became *padūlem,* and then *pădure* in Romanian and *pyll* in Albanian, both with a modified meaning, "forest."

Early contacts between Albanian and Romanian

Conversely, Romanian also shares some apparently non-Latin indigenous terms with Albanian; *e.g.,* Romanian *brad,* Albanian *bredh* "fir." Thus these two languages reflect special historical contacts of early date. Early communication with the Goths presumably contributed *tirq* "trousers, breeches" (from an old compound "thigh-breech"), while early Slavic contacts gave *gozhdë* "nail." Many Italian, Turkish, Modern Greek, Serbian, and Macedonian-Slav loans can be attributed to cultural contacts of the past 500 years with Venetians, Ottomans, Greeks (to the south), and Slavs (to the east).

A fair number of features—*e.g.,* the formation of the future tense and of the noun phrase—are shared with other languages of the Balkans but are of obscure origin and development; Albanian or its earlier kin could easily be the source for at least some of these. The study of such regional features in the Balkans has become a classic case for research on the phenomena of linguistic diffusion.

(E.P.H.)

URALIC LANGUAGES

The Uralic language family consists of two related groups of languages, the Finno-Ugric and the Samoyedic, both of which are thought to have developed from a common ancestor, called Proto-Uralic, that was spoken 7,000 to 10,000 years ago in the general area of the northern Ural Mountain Range. Over the millennia, both Finno-Ugric and Samoyedic have given rise to more or less divergent subgroups of languages, which nonetheless have retained certain traits from their common source. For example, the degree of similarity between two of the least closely related members of the Finno-Ugric group, Hungarian and Finnish, is comparable to that between English and Russian (which belong to the Indo-European family of languages). The difference between any Finno-Ugric language and any Samoyedic tongue would be even greater. On the other hand, more closely related members of Finno-Ugric, such as Finnish and Estonian, differ in much the same manner as greatly diverse dialects of the same language.

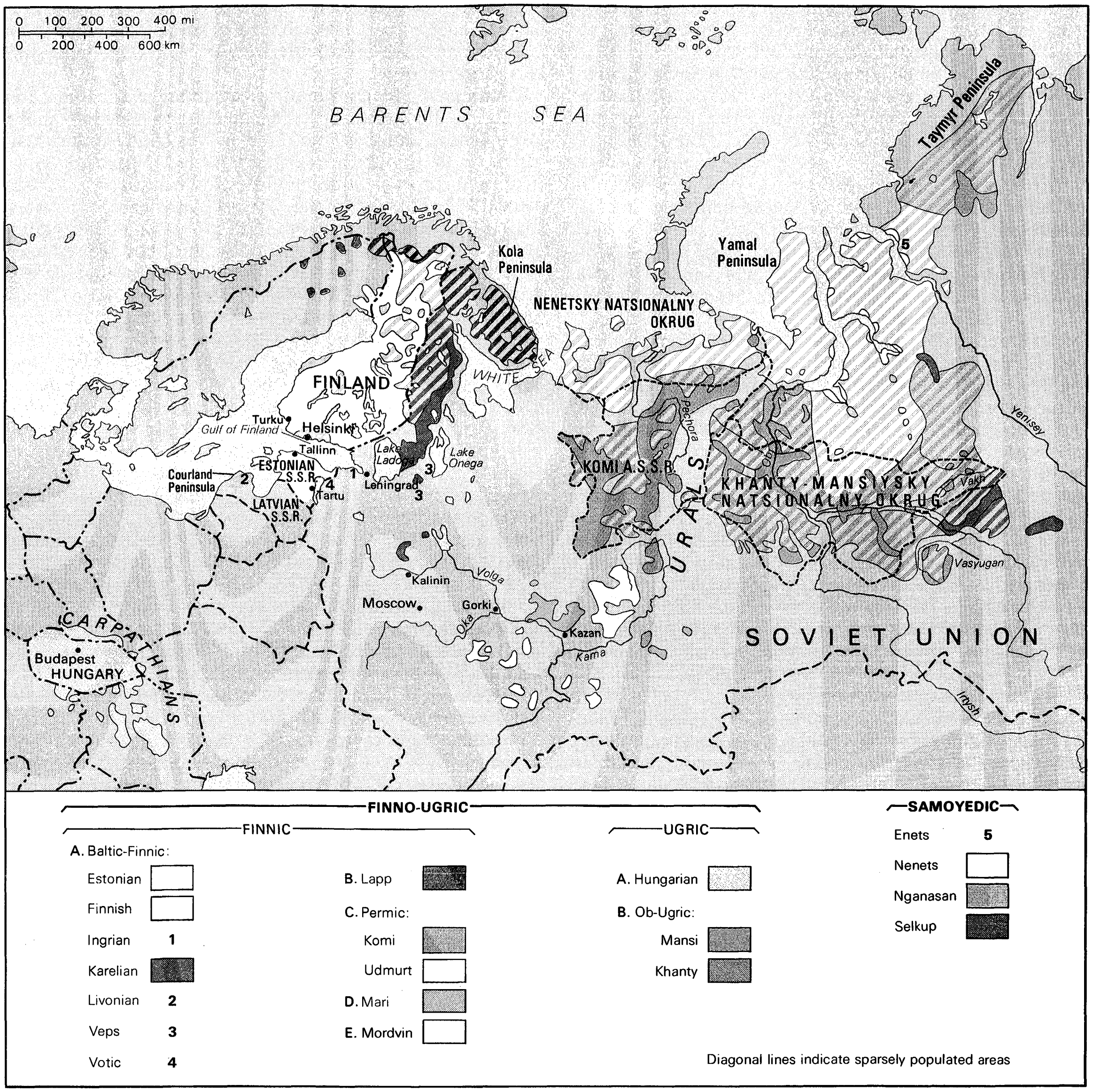


Figure 17: Distribution of the Uralic languages.

Finno-Ugric languages

The Finno-Ugric languages are represented today by some 15 languages scattered over an immense Eurasian territory. In the west they include the European national languages—Hungarian, Finnish, and Estonian—as well as Lapp, the westernmost member of the group, spoken by numerous separate groups across the northern Scandinavian Peninsula from central Norway to the White Sea. The remaining Finno-Ugric languages are located within the Soviet Union, with one major concentration—including Estonian, Livonian, Votic, Karelian, and Veps—along a broad zone extending from the Gulf of Riga to the Kola Peninsula. The Mordvin and Mari (or Cheremis) languages are found in the region of the central Volga; from there extending northward along the river courses west of the Urals are the Permic languages—Udmurt (Votyak) and Komi (Zyryan). East of the Urals, along the Ob River and its tributaries are the easternmost representatives of the Finno-Ugric group—Mansi (Vogul) and Khant (Ostyak).

The largely nomadic Samoyeds are sparsely distributed over an enormous area extending inward from the Arctic shores of the Soviet Union from the White Sea in the west to Khatangsky (Khatanga) Bay in central Siberia in the east. Nenets (Yurak), the westernmost of these languages, reaches eastward to the mouth of the Yenisey River and includes a small insular group on Novaya Zemlya. Speakers of Enets (Yenisey) are located in the region of the upper Yenisey. The lower half of the Taymyr Peninsula is the habitat of the Nganasan (Tavgi), the easternmost of the Uralic groups. The fourth language, Selkup, lies to the south in a region between the central Ob and central Yenisey; its major representation is located between Turukhansk and the Taz River. A fifth Samoyedic language, Kamas (Sayan), originally spoken in the vicinity of the Sayan Mountains, was spoken in the early 1970s by one elderly speaker, then residing in Estonia.

Numbers of speakers

In general, the westernmost members of the Uralic family are spoken by the greatest numbers of speakers. The largest groups are Hungarian, with some 14,000,000 speakers; Finnish, with 5,000,000; and Estonian, with approximately 1,000,000. Among the lesser known Uralic languages of the Soviet Union, several have rather substantial representation: Mordvin, more than 900,000 speakers; Mari, about 570,000; Udmurt, close to 590,000; Komi, about 285,000; and Karelian, with just under 86,000. The approximately 30,000 to 40,000 Lapps are distributed over four countries: Norway, 20,000; Sweden, 10,000; Finland, 3,000; and the Soviet Union, 1,500. Other Finno-Ugric languages with more than 1,000 speakers are Khant (about 14,000), Veps (about 3,000) and Mansi (about 4,000). In the mid-20th century, Votic and Livonian were still maintained by small communities of speakers, but they appeared to be facing extinction in future generations. The entire Samoyedic group consists of about 28,000 speakers. Of these, Nenets claims over 25,000 speakers; Selkup, roughly 2,000; Nganasan, fewer than 700; and Enets, about 300.

The political history of the various Uralic groups largely has been one of resisting encroachment from adjacent European (especially Germanic and Slavic) and Turkic groups and from other Uralic neighbours. Only three groups have succeeded in achieving political independence—Hungary, Finland, and Estonia (the last is now, however, one of the 15 republics of the Soviet Union). Five of the minority groups in the Soviet Union have the status of autonomous republics: Mordvinian A.S.S.R., Mari A.S.S.R., Udmurt A.S.S.R., Komi A.S.S.R., and Karelian A.S.S.R. (formerly a union republic). Four more groups are recognized at the level of local administration under their own national *okruga;* Khant and Mansi (under one area), the Permyak dialect of Komi, and Nenets (under three *okruga*).

Alphabets used for the Uralic languages

The earliest known manuscript in a Uralic language is a Hungarian funeral oration (*Halotti Beszéd*), a short, free translation from Latin, which stems from the turn of the 13th century AD. A 12-word Karelian fragment also dates from the 13th century. Old Permic, the earliest attested form of Komi, received its own alphabet (based on the Greek and Old Slavic symbols) in the 14th century through the missionary efforts of St. Stephen, bishop of Perm. The first Finnish and Estonian texts date from the 16th century, and are in printed form. Lapp was first written in the 17th century. Since then, nearly all the more populous Uralic languages have some kind of written form, but at present there is a native literature only for the above-mentioned languages and for those groups in the Soviet Union that have their own administrative regions. Currently, the Uralic languages within the Russian Soviet Federated Socialist Republic use a modification of the Cyrillic alphabet; the others employ the Latin alphabet, adapted to the peculiar demands of their own sound systems. For example, the important distinction between long and short vowels in Finnish is indicated by doubling the letters for long vowels (*a* versus *aa*), whereas in Hungarian the long vowel is marked by an acute accent (*a* versus *á*).

Racially, the Uralic peoples present an unhomogeneous picture. In general, they may be considered a blend of European and Mongoloid types, with the more western groups (especially the Hungarian, Baltic-Finnic, and Erzya Mordvin groups) being strongly European and those east of the Urals primarily Mongoloid. Although scholars do not agree as to what features, if any, constitute the most archaic Uralic type, recent study indicates that it is possible to speak of a Uralic racial type, an intermediate stage between the European and the Mongoloid, the basic features of which are medium-dark to dark hair and eye colour, relatively small stature, and often a concave bridge of the nose. According to this view, the more archaic Uralic type is best preserved among the Lapp, Mari, and Permic groups.

Attempts to trace the genealogy of the Uralic languages to periods earlier than Proto-Uralic have been hampered by the great changes in the attested languages, which preserve relatively few features upon which to base meaningful claims for a more distant relationship. Most commonly mentioned in this respect is a putative connection with the Altaic language family (including Turkic and Mongolian). This hypothetical language group, called Ural-Altaic, is not considered by most scholars to be soundly based. Although the Uralic and Indo-European languages are not generally thought to be related, more speculative studies have suggested such a connection. Of the various attempts to find outside relationships, only those linking Uralic with Yukaghir, a Paleo-Siberian tongue, appear to have serious support.

Table 35: Names Used to Designate Uralic Groups

English form	native form
Finnish	*suomi*
Karelian	*karjala*
Ingrian	*izhor*
Veps	*vepsä, lüüd* (Old Russian *vesj, chudj*)
Estonian	*eesti* (Old Russian *chudj*)
Votic, Vote	*vadja* (Old Russian *vodj; chudj*)
Livonian	*liiv*
Lapp	*sabme* (Russian *saami;* earlier *lopj*)
Mordvin	*erza, moksha* (no common name)
Mari (Cheremis)	*mari* ("man")
Udmurt (Votyak)	*ud-murt* (*murt* = "man")
Komi (Zyryan)	*komi* (Old Russian *permj*)
Khant (Ostyak)	*khanty* (Old Russian *jugra*)
Mansi (Vogul)	*manshi* (also used to designate the Khanty; Old Russian *jugra*)
Hungarian	*magyar* (Russian *vengr*)
Nenets (Yurak)	*nenets, hasawa* ("man"; Old Russian *samojadj*)
Enets (Yenisey)	*enetj* (related to the name *nenets*)
Nganasan (Tavgi, Avam)	*ŋanasan* (related to the name *nenets*)
Selkup (Ostyak Samoyed)	*shöl-qup* (*shö[l]* = "earth," *qup* = "man")

Because the names designating many of the Uralic peoples have never been standardized, a wide range of appellations is encountered in references to these groups. Earlier designations, especially in the case of the groups in Russia, tended to be taken from derogatory names used by neighbouring peoples; *e.g.,* Cheremis—now Mari. Table 35 indicates the names in use. Standard usage is in the left column, and earlier, Russian-based forms are in parentheses. The name that the group uses for itself and certain other information, such as Russian and Old Russian forms, are in the right column. Several names are identical to the word for man in these languages. (Finnish *mies* "man" has also been etymologically related to the names Magyar and Mansi.) It is important that Ostyak (Khant) not be confused with Ostyak Samoyed (Selkup) nor with Yenisey Ostyak (Ket, a non-Uralic, Paleo-Siberian tongue), which should also not be confused with Yenisey (Enets).

LANGUAGES OF THE FAMILY

Divisions within Finno-Ugric

The two major branches of Uralic are themselves composed of numerous subgroupings of member languages on the basis of closeness of linguistic relationship. Finno-Ugric can first be divided into the most distantly related Ugric and Finnic (sometimes called Volga-Finnic) groups, which may have separated as long ago as five millennia. Within these, three relatively closely related groups of languages are found: the Baltic-Finnic, the Permic, and the Ob-Ugric. The largest of these, the Baltic-Finnic group, is composed of Finnish, Estonian, Livonian, Votic, Ingrian, Karelian, and Veps. The Permic group consists of Komi and Udmurt; the Ob-Ugric group includes Mansi and Khant.

The Ugric group comprises the geographically most distant members of the family—the Hungarian and Ob-Ugric languages. Finnic contains the remaining languages: the Baltic-Finnic languages, Lapp, Mordvin, Mari, and the Permic tongues. There is little accord on the further subclassification of the Finnic languages, although the fairly close relationship between Baltic-Finnic and Lapp is generally recognized (and is called North Finnic); the degree of separation between the two may be compared to that between English and German. Mordvin has most frequently been linked with Mari (a putative Volga group), but comparative evidence also suggests a bond with Baltic-Finnic and Lapp (that is, West Finnic). The extinct Merya, Murom, and Meshcher groups, known only from Old Russian chronicles, are assumed to have been Finnic peoples and from their geographical location northwest of Mordvin must have belonged to West Finnic. One hypothesis for the internal relationships of the Uralic family as a whole is given in Figure 18.

The precursor of the modern Samoyedic languages is thought to have divided near the turn of the Christian Era into a northern and a southern group. North Samoyedic consists of Nenets, Enets, and Nganasan. South Samoyedic contains but a single living language, Selkup, but is known to have been represented by numerous other dialects, now extinct: Kamas, Motor, Koibal, Tofalar (Karagasy), Soyot, and Taigi.

Hungarian. Hungarian, the official language of Hungary, remains the primary language of the fertile Carpathian Basin. Bounded by the Carpathian Mountains to the north, east, and southwest, the Hungarian area is represented by almost 2,500,000 speakers outside the boundaries of Hungary—in Czechoslovakia, the Ukranian S.S.R., and

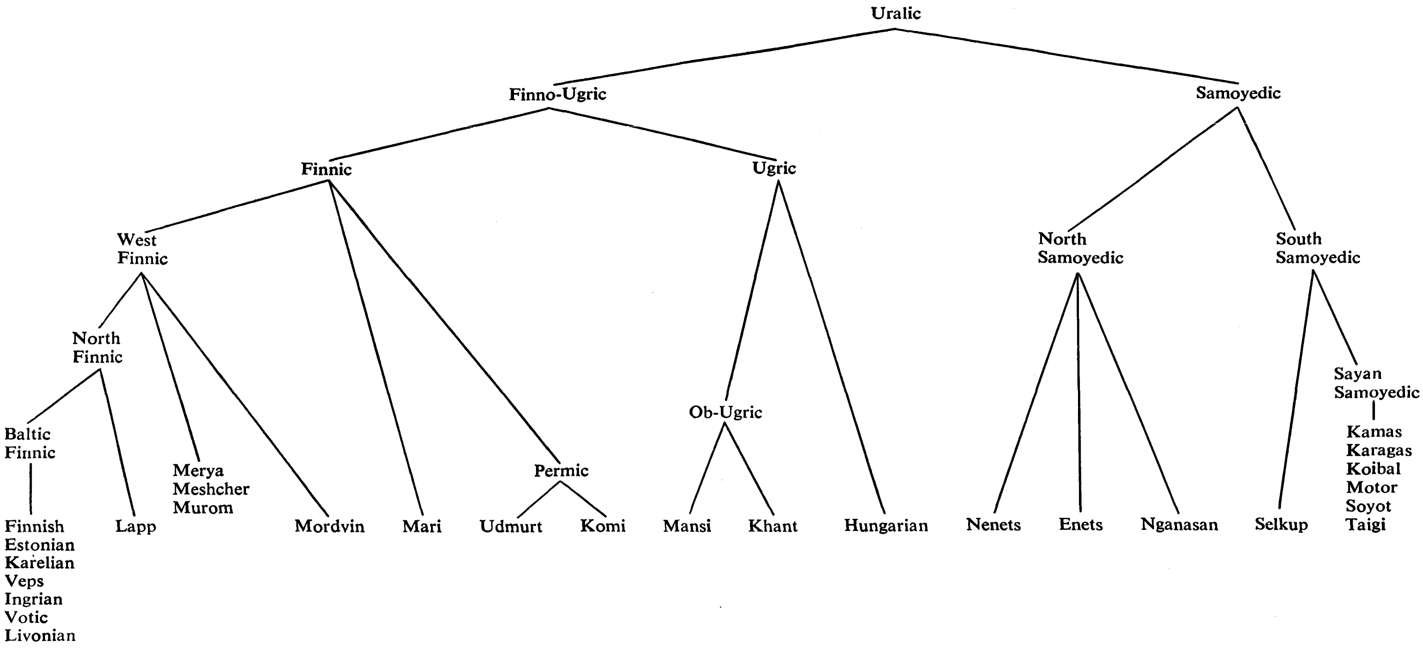


Figure 18: Family tree diagram of the Uralic languages (see text).

Romanian Transylvania (some 1,700,000). To the south, a substantial Hungarian population (over 450,000) extends into central Yugoslavia. Hungarian emigrant communities are found in many parts of the world, especially in North America and Australia.

The ancestors of the Hungarians, following their separation from the other Ugric tribes, moved south into the steppe region below the Urals. As mounted nomads, in contact with and often in alliance with Turkic tribes, they moved westward, reaching and conquering the sparsely settled Carpathian Basin in the period 895–896. The Hungarians came under the influence of Rome through their first Christian king, Stephen (István), in 1001, and the use of Latin for official purposes continued into the 19th century. Following a Hungarian defeat at the Battle of Mohács in 1526, Hungary was occupied by Turkish forces, who were replaced by German Habsburg domination in the late 17th century. Concern for a common literary medium, closely tied with Hungarian nationalism, began in the late 18th century. More recent foreign influences on the language were suppressed and replaced by native words and constructions. The literary form received a broad dialect base, facilitating its use as a national language.

Dialects of Hungarian

Modern Hungarian has eight major dialects, which permit a high degree of mutual intelligibility. Budapest, the nation's capital, is located near the junction of three dialect areas: the South, Transdanubian, and Palóc (Northwestern). As a result of unfavourable treaties following both world wars, especially the Treaty of Trianon, two dialects (Central Transylvanian and Szêkely) lie almost entirely within Romania, and the remaining six dialects radiate outward into neighbouring countries.

The Hungarians' own name for themselves is *magyar.* Other Western appellations, such as the French *hongrois,* German *Ungar,* and Russian *vengr,* all stem from the name of an early Turkic tribal confederation, the *on-ogur* (meaning "ten tribes"), which the Hungarians joined in their wanderings toward the west, and does not indicate relationship with the ancient Huns, a Turkic tribe. One of the earliest recorded references to the Hungarians, a Byzantine geographical survey of Constantine Porphyrogenitus (died 959) entitled *De administrando imperio,* lists the *megyer* as one of the Hungarian tribes, but, as was typical in early reports, the Hungarians were not distinguished from their Turkish allies.

Ob-Ugric: Khant and Mansi. The Ob-Ugric peoples, the Khant and the Mansi (Vogul), are among the smallest Finno-Ugric groups. Although their numbers have declined over recent centuries, the Khant language is still maintained by about 14,000 speakers, and approximately 52 percent of the 7,600 Mansi still claim it as their mother tongue. To a large extent they have been assimilated by their Russian and Tatar neighbours. These two peoples are widely dispersed along the Ob River and its tributaries, for the most part within the Khant-Mansi Autonomous Okrug, which has its administrative centre in Khant-Mansiysk at the junction of the Ob and the Irtysh. The Mansi are found along the western tributaries primarily north of the Irtysh and just east of the Urals; a few speakers are also found in the Arctic lands west of the Urals. The Khant live along both the Ob itself and its tributaries.

Because of the great distances between the various groups, the dialects of both languages show considerable divergence. They are usually designated by the name of the river on which they are spoken. Mansi has four main dialect groups, of which one (Tavda) is practically extinct and another (Konda) is no longer spoken by the youth of the area. The largest dialect group (Northern) is centred on the Sosva and serves as the basis for the literary language. Khant is divided into three main dialects: a northern dialect in the general area of the mouth of the Ob, an eastern dialect extending from east of the Irtysh to the Vakh and Vasyugan tributaries, and a southern dialect lying between the other two. Literary Khant has been based primarily on the northern group, but standardization remains weak, and in recent decades other dialects have also been used.

Khant and Mansi dialects

After the division of the Proto-Ugric language into separate languages, it is likely that the precursors of the Ob-Ugric tribes were still centred west of the Urals well within historic times. The Old Russian Chronicles of Nestor, which assigned them the common name *jugra,* places them in the vicinity of the Pechora River in 1092; they did not shift to the Ob waterways until several centuries later. Both of the Ob-Ugric languages first appeared in printed form in 1868 as a result of gospel translations published in London, but it was not until after the formation of their autonomous okrug in 1930 that any sort of literary form of Khant and Mansi really existed. Until 1937 numerous books were published using a modified Latin (roman) alphabet; since then Cyrillic has been used. To a certain extent, elementary education is conducted in the native languages within the autonomous okrug.

Finnish. Finnish, together with Swedish (an unrelated North Germanic language), serves as the official language of Finland. It is now spoken by more than 5,000,000 people, including 94 percent of the inhabitants of Finland plus nearly 500,000 Finns in North America, Sweden, and the Soviet Union. It is also recognized as an official language in the Karelian A.S.S.R., alongside Russian.

Modern Finnish dialects

Finnish as the common language of the Finns is not the direct descendant of one of the original Baltic-Finnic dialects; rather, it arose through the interacton of several separate groups in the territory of modern Finland. These included the Häme; the southwestern Finns (originally called Suomi), who appear to be close relatives to the Estonians, because they arrived directly from across the Gulf of Finland; and the Karelians, perhaps themselves a blend of Veps and more western Finnic groups. Early Russian chronicles refer to these as *jemj, sumj,* and *korela.* The intermixture of the three groups is still reflected in the distribution of the five main modern dialects, which form a western and an eastern area. The western area contains the southwest dialect (near Turku), Häme (south central), and a northern dialect subgroup (largely a mixture of the other two plus eastern traits). The eastern area consists of the Savo dialect (perhaps a blend of the original Karelian and Häme dialects) and a southeastern dialect, which strongly resembles Karelian. The Finnish word for their land and their language is *suomi,* the original meaning of which is uncertain. The first use of the term Finn *(fenni)* is found in the 1st century AD in Tacitus' *Germania,* but this usage is generally considered to refer to the ancestors of the Lapps, who have also been labelled Finns at various times. (The province of Norwegian Lappland is called Finnmark.)

The first book in Finnish was an ABC book from 1543 by Mikael Agricola, founder of the Finnish literary language; five years later it was followed by Agricola's translation of the New Testament. Finnish was accorded official status in 1809, when Finland entered the Russian Empire after six centuries of Swedish domination. The publication of the national folk epic, the *Kalevala,* created from folk songs collected among the eastern dialects by folklorist and philologist Elias Lönnrot (first edition in 1835; substantially expanded in 1849), gave increased impetus to the movement to develop a common national language encompassing all dialect areas.

Estonian. Estonian serves as the official language of the Estonian S.S.R., located immediately south of Finland across the Gulf of Finland. There are about 1,000,000 speakers, most of them living within the Estonian S.S.R. but also elsewhere in the Soviet Union and abroad, especially in North America and Sweden. Modern Estonian is the descendant of one or possibly two of the original Baltic-Finnic dialects. The modern language has two major dialects, a northern one, which is spoken in most of the country, and a southern one, which extends from Tartu to the south. The northernmost dialects share many features with the southwestern Finnish dialect. The Estonians' own name *eesti* came into general use only in the 19th century. The name *aestii* is first encountered in Tacitus, but it is likely that it referred to neighbouring Baltic-Finnic peoples.

The first connected texts in Estonian are religious translations from 1524; the *Wanradt-Koell Catechism,* the first book, was printed in Wittenberg in 1535. Two centres of culture developed—Tallinn (fomerly Reval) in the north and Tartu (Dorpat) in the south; in the 17th century they gave rise to two literary languages. Influenced by the Finnish *Kalevala,* the Estonian author Friedrich Reinhold Kreutzwald fashioned a national epic, the *Kalevipoeg* ("Son of Kalevi"), which appeared in 20 songs between 1857 and 1861. As with the *Kalevala,* this was instrumental in kindling renewed interest in a common national literary form in the late 19th century.

Karelian, Veps, Ingrian, Votic, Livonian

Smaller Baltic-Finnic groups. The five less numerous Baltic-Finnic groups—Karelian, Veps, Ingrian, Votic, and Livonian—lie within the territory of the Soviet Union, largely in the general vicinity of the Gulf of Finland. The Karelians, Veps, and Livonians were among the original Baltic-Finnic tribes; Votic is considered to be an offshoot of Estonian, and Ingrian, a remote branch of Karelian. None of these languages currently has a literary form, although unsuccessful initial attempts to establish one have been made for all but Votic (for Livonian as early as the 19th century, for the others during the 1930s). Since the beginning of the 20th century the numbers of these Baltic-Finnic speakers have been drastically reduced, and, with the exception of Karelian and Veps, their extinction within several generations seems certain. Ingrian, Votic, and Livonian, each with fewer than 1,000 speakers, were not reported in the 1979 census.

Karelian, the largest of these groups, with about 86,000 speakers—not counting those Karelians who emigrated into Finland following World War II—lies along a broad zone just east of the Finnish border from just north of Leningrad to the White Sea. A separate group of Karelians is found far to the south near Kalinin (formerly Tver) on the upper Volga. Karelian has two major dialects, Karelian proper and Olonets (*aunus* in Finnish), which is spoken northeast of Lake Ladoga. One of the first historical mentions of the Karelians is found in a report of the Viking Ohthere to King Alfred of England at the end of the 9th century; this indicates that they were already on the southern Kola Peninsula as neighbours of the Lapps and gives their name as *beorma.*

The language of one of the original Baltic-Finnic tribes, Veps, is spoken southeast on a line connecting lower Lake Ladoga with central Lake Onega. About one-third of the 8,000 Veps still consider the language their native tongue—a sharp decline from the 26,172 speakers reported in the mid 1800s. A small Baltic-Finnic group, the Ludic dialects, is found between Veps and Karelian and is generally considered as a blend of the two major groups, rather than a separate language; the dialects are more closely akin to Karelian. The Ingrians and the Votes live on the southern Gulf of Finland in the border area between Estonia and Soviet Russia, where they have survived because the border area has been closed to outsiders, even within the Soviet Union. Livonian, maintained in a dozen villages on the northernmost tip of Latvia, on the Courland Peninsula, has about 500 speakers, but the language is not used among the younger generation.

Dialect groups of Lapp

Lapp. The Lapps are widely distributed, from central Norway northward and eastward across northern Sweden and Finland to the Kola Peninsula. The number of Lapps has increased over the past century to more than 30,000, but the number of Lapp speakers has declined rapidly in recent decades as the language has given way to the various official national languages. Lapp is generally divided into three main dialect groups, each with various subtypes. These dialects are virtually mutually unintelligible, so that when speakers of different Lapp groups meet, they generally converse in Finnish, Swedish, or Norwegian. To speak of a single Lapp language is thus misleading. Lapp represents a group of at least four or five languages at least as diverse as the separate Baltic-Finnic languages. The largest group, North Lapp (with roughly two-thirds of all speakers), is centred in northern Norway, Sweden, and Finland. East Lapp consists of two small groups in eastern Finland—Inari and Skolt—plus Kola Lapp in the Soviet Union. South Lapp is still represented by a few speakers scattered from central Norway to north central Sweden.

North Lapp has had a literary tradition that began with the 17th-century Swedish Lapp Bible and other religious translations; in the mid-20th century elementary schools that used Lapp as the language of instruction were found in many larger North Lapp communities. Two basic variants of the literary language are in use. One, in Norway and Sweden, employs a special Lapp orthographic system devised to accommodate a wide range of dialectal variation; a second, in Finland, is based on a more narrow adaptation of Finnish orthography. Each of the two types has numerous local variants, and progress toward a common Lapp orthography has been slow.

A wide range of theories has been posited to reconcile the apparent conflict between the different racial features of the Lapps and the Baltic-Finnic peoples and their strong linguistic similarities. The Lapps have been said to represent a non-Uralic people (the so-called Proto-Lapps) or perhaps a Samoyed tribe that abandoned its own language for a late stage of Finnic, perhaps early Baltic-Finnic. Another view is that the Lapps are one of the original Finnic tribes and not linguistically more closely related to Baltic-Finnic. In the absence of any compelling evidence, no conclusive answer can be provided. On the other hand, the fact that the non-Uralic origin of the Baltic Finns is rarely considered, despite recent evidence suggesting that the Lapps best represent the racial characteristics of the Finno-Ugric peoples, leads to the possibility of a certain degree of chauvinism in the earlier hypotheses suggesting outside origins for the Lapps. In any case, it is clear that the Lapps were already present north of the Gulf of Finland prior to the arrival of the first Baltic-Finnic tribes, and from there they may have extended over much of the Scandinavian Peninsula. They have been mentioned as the northern neighbours of the north Germanic tribes in numerous historical sources of the 1st millennium of the Christian Era. The Lapps were taxed by the Norwegians in the 9th century and by the Karelians in the 13th century and since that time have continually retreated northward under pressure from their southern neighbours. The Lapps' own name for themselves, *sabme,* is etymologically related to the Finnish dialect name, *häme.*

Mordvin, Mari, and the Permic languages

Other Finnic languages. Mordvin, Mari, and the Permic languages—Udmurt and Komi—are each recognized by autonomous republics within the Russian Soviet Federated Socialist Republic, where they share official status with Russian. Mordvin, Mari, and Udmurt are centred on the middle Volga River, in roughly the area considered to have been the original home of Proto-Finno-Ugric. Because of their location, the history of these groups over the past millennium has been closely tied to that of the Turkic Bulgars, the Tatars (until 1552), and then the Russians. The Komi, having moved far to the north, eventually reaching into the Arctic tundra, did not come under Bulgar or Tatar influence. A written form of early Komi, Old Permic, was used in religious manuscripts in the 14th century, and a native Komi literary tradition stems from the 19th century. Grammars of Mari and Udmurt prepared by Russian linguists appeared in 1775, but native literary development in these languages, as well as in Mordvin, is of recent origin. Although these groups currently enjoy the status of large minorities, even in their respective autonomous areas, their numbers have increased over the past century, and they have maintained ethnic consciousness.

Mordvin. Mordvin, with more than 900,000 speakers (about 78 percent of the 1,177,000 Mordvins reported in the early 1980s), is the fourth-largest Uralic group. The Mordvins are widely scattered over an area between the Oka and Volga rivers, some 200 miles southwest of Moscow. Less than half of their number live within the Mordvinian Autonomous Soviet Socialist Republic. Mordvin has two main dialects, Moksha and Erzya, which are sometimes considered separate languages. Both have literary status. Although the Mordvins do not have a common designation for themselves beyond the two dialect names, the name Mordens appears in the 6th-century *Getica* of Jordanes and is no doubt related to the Permic word for "man," *murt/mort.*

Mari. Mari (Cheremis) is currently maintained by about 570,000 speakers (approximately 78 percent of 627,000 in the early 1980s), primarily in an area north of the Volga between Kazan and Gorki, northeast of the Mordvin area, especially within the Mari A.S.S.R. Mari's three main dialects are: the Meadow dialect, used by the largest group north of the Volga and the basic dialect of the Mari A.S.S.R.; Eastern Mari, used by a small group near Ufa, originally speakers of the Meadow dialect who emigrated in the late 18th century; and the Mountain dialect, to the west and on the south bank of the Volga. The Mountain and Meadow dialects both serve as literary languages and differ from each other only in minor details.

Udmurt and Komi

The Permic languages. Speakers of the two closely related Permic languages number close to 900,000—roughly 590,000 Udmurts and 285,000 Komi. Udmurt is concentrated largely in the vicinity of the lower Kama River just east of the Mari A.S.S.R., in the Udmurt A.S.S.R. Only very minor dialectal differences are found within Udmurt. The Komi language area extends into the Nenets National Okrug far to the north. Lesser groups of Komi are found as far west as the Kola Peninsula and east of the Urals. Three major dialects are recognized, although the differences are not great: Komi Zyryan, the largest group, which serves as the literary basis within the Komi A.S.S.R.; Komi-Permyak, the dialect of the Komi-Permyak National Okrug, where it has literary status; and Komi-Yazva, spoken by some 4,000 Komi to the east of the Komi-Permyak National Okrug and south of the Komi A.S.S.R.

The Samoyed languages. Nenets, with the largest number of speakers of all the Samoyed languages, has added a substantial increase in the number of its speakers over the past century, from 9,245 in 1897 to about 28,000 in the early 1980s. Two distinct groups of Nenets differ in dialect as well as in cultural traditions: the Forest Nenets, a smaller, more concentrated group in the wooded area north of the central Ob; and the Tundra Nenets, a group whose territory stretches roughly 1,000 miles eastward from the White Sea. These are the "Samoyadj" of Nestor's chronicles, but little is known of the history of any of the Samoyed peoples until recent centuries. Nenets alone among the Samoyed languages can claim a native literature, although both it and Selkup have been in written form since the 1930s. Evidence of the cultural prestige of certain Nenets tribes is seen in the adoption of a Samoyed language by Khant speakers on the Yamal Peninsula. Enets is spoken by a dwindling group of several hundred Samoyeds near the mouth of the Yenisey, just east of the Nenets. Nganasan, spoken by the northernmost Eurasian people, is found north and east of the Enets-speaking group, centring on the Taymir Peninsula. The number of Nganasans has remained fairly constant, and they seem to have a high degree of ethnic identity (about 75 percent of 900 Nganasans still claimed Nganasan as their mother tongue in the early 1980s).

Selkup, the last of the southern Samoyed languages, is represented by scattered groups of speakers who live on the central West Siberian Plain between the Ob and the Yenisey. Only half of the 3,500 Selkup speakers still considered the language their mother tongue in the early 1980s.

EARLY HISTORY

Hypotheses concerning early Uralic cultures and languages

Determining the geographical location, material culture, and linguistic characteristics of the earliest stages of Uralic at a period thousands of years prior to any historical records is a problem beset with enormous difficulties; consensus among Uralic scholars is limited to a handful of general hypotheses.

The original homeland of Proto-Uralic is considered to have been in the vicinity of the central Urals, possibly centred west of the mountains. Following the dissolution of Uralic, the precursors of the Samoyeds gradually moved northward and eastward into Siberia. The Finno-Ugrians moved to the south and west, to an area close to the junction of the Kama and Volga rivers.

Knowledge of the location of these early groups is based on several kinds of indirect evidence. One approach attempts to reconstruct their natural environment on the basis of shared cognate words for plants, animals, and minerals and on the distribution of these words in the modern languages. For example, cognates (related words) designating certain types of spruce are found in all the Uralic languages except Hungarian (Finnish *kuusi,* Lapp *guossâ,* Mordvin *kuz,* Komi *koz,* Khant *kol,* Nenets *xādy,* Selkup *kūt*). Because the range of this type of fir tree is restricted to more northern climates, it is generally assumed that the widespread consistent association of the name and the tree suggests a period in which Proto-Uralic was spoken within that zone. Several other terms for plants (*e.g.,* Finnish *muurain* "cloudberry" [*Rubus arcticus*]), a term for metal (Estonian *vask* "copper," Hungarian *vas* "iron," Nganasan *basa* "iron"), and a word for "reindeer" (Lapp *boaʒo*) are also consistent with a northern Ural location. Great caution is necessary in such matters, because the association of words and objects can also result from borrowing, perhaps long after the period of Uralic unity; especially such culturally mobile items as "metal" and "reindeer" cannot with certainty be traced to a Proto-Uralic community. The central Volga location of Proto-Finno-Ugric is rather strongly supported by an abundance of shared terminology dealing with beekeeping, which constitutes a rather significant part of the culture of this region.

A second approach to determining the location of Proto-Uralic is based on contacts with other, unrelated languages as evidenced by loanwords from one group to the other. Early Finno-Ugric borrowed numerous terms from very early dialects of Indo-European. Though these words are entirely lacking from the Samoyed languages, within the Finno-Ugric division they are shared by the most remotely related members and show the same phonetic relationships as the native Finno-Ugric vocabulary. Examples include agricultural and bee-culture terminology (*e.g.,* "honey": Finnish *mete,* Komi *ma,* Hungarian *méz* [compare Indo-European **medhu-*]; "pig": Finnish *porsas,* Komi *porś*); several numerals ("hundred": Finnish *sata,* Hungarian *száz*); mineral words ("salt": Finnish *suola,* Komi *sol*); and the word for orphan (Finnish *orpo,* Hungarian *árva*). The nature of these borrowings, together with the linguist's relatively richer knowledge of early Indo-European, supports a southward movement of Proto-Finno-Ugric and also provides some insight into the culture of the Finno-Ugrians.

Loanwords from unrelated languages

The central Volga location is also supported by the geographical distribution of the daughter languages. Except for Hungarian, which moved westward across the steppe, the Finno-Ugric languages form two chains distributed along major waterways, with the junction of the Kama and Volga at their centre. One chain extends northward along the Kama, across the northern tip of the Urals into the Ob watershed, then south along the Ob and its tributaries. The second extends northeast along the Volga to the Gulf of Finland. The extinct Merya, Murom, and Meshcher languages were once links in this chain. Finally, assumptions about the more distant relationships of Uralic have influenced views concerning its original location. Earlier, proponents of the Ural-Altaic hypothesis tended to place the Uralic homeland in south central Siberia, near the sources of the Ob and Yenisey, but there is no support for this view.

LINGUISTIC CHARACTERISTICS

The linguistic structure of Proto-Uralic has been partially reconstructed by a comparison of the similarities and differences among the known Uralic tongues. Not all existing similarities can be attributed to a common Uralic origin; some may also reflect universal pressures and limitations on language structure (*e.g.,* the tendency to weaken stopped consonants between vowels, the modifying of a sound to become more similar to a preceding or following sound) or the influence of neighbouring, even genetically unrelated language structures (*e.g.,* the various types of vowel harmony [see below] in Finno-Ugric probably reflect such areal pressure).

Phonological characteristics. The correspondences of sounds in cognate Uralic words are illustrated in Table 36. Thus, a *p* in the beginning of a Finnish word corresponds to *f* in Hungarian (*puu* : *fa*); a Finnish *k* is matched by Hungarian *h* before a back vowel (*a, o*), otherwise by *k;* within the word, Finnish *t* is matched by Hungarian *z* and *nt* by *d*; Finnish initial *s* sometimes corresponds to Hungarian *s* (spelled *sz*) and sometimes to no consonant at all (*syli* : *öl*). In most of these instances, Finnish has retained the consonants of the Proto-Uralic consonant system. One exception is *nt,* which was originally **mt;* the *m* has become *n,* matching the position of articulation of the adjacent *t.* (An asterisk marks a form that is not found in any document or living dialect but is reconstructed as having once existed in an earlier stage of a language.) A second Finnish innovation is the loss of the distinction between the two original *s* sounds, **s* and palatalized **ś.* (Palatalization is the modification of a sound by simultaneous raising of the tongue to or toward the hard palate.) Hungarian maintains a distinction, but the original **s* words have lost this sound. By careful examination of such systematic relationships it is possible to sketch out much of the phonological structure of early Uralic. The reconstructions of Table 36 (last column) are based on the view that the vowel system of Baltic-Finnic is relatively more conservative, whereas the consonant contrasts have been best preserved in Lapp.

Finnish and Hungarian sound correspondences

Consonants. The following consonant sounds are generally posited for the early stages of Uralic: **p*, **t*, **č*, **k*, **s*, **š*, **ð*, **l*, **r*, **m*, **n*, *ŋ, **j*, **v* (*č* is pronounced as the *ch* in "chip," *š* as the *sh* in "ship," *ð as the th* in "then," *ŋ* as the *ng* in "sing," *j* as the *y* in "yet"), and the palatalized alveolar sounds **t′, ś, ð′, l′, ń,* plus a few others less well established. Modern Finnish has a much smaller inventory of consonants, having lost the palatalized alveolar sounds and **č, *š, *ð,* and **ŋ*. Hungarian, on the other hand, has a larger number of consonants by virtue of a newly introduced distinction between sounds made with and without vibration of the vocal cords (voicing), such as voiceless *p, t, s* as opposed to voiced *b, d, z; e.g., dél* "noon" : *tél* "winter." Other Uralic languages, such as Komi, have also acquired a voicing contrast (*e.g., doj* "pain" : *toj* "louse"), but the geographical distribution of those languages in which the voicing contrast plays an active role leaves little doubt as to its areal (regional) origin under the influence of Indo-European and Turkic languages.

Vowels. Essentially nothing is known of the Proto-Uralic vowels, and there is little agreement about the nature of the Proto-Finno-Ugric vowel system. It is clear, however, that, in contrast to a relatively limited number

Proto-Uralic vowels

Table 36: Representative Cognates in Selected Uralic Languages

English translation	Finnish	Estonian	Lapp	Mari	Komi	Khant	Hungarian	Nenets	Proto-Uralic
head; end	*pää*	*pea*	—	—	*pom*	—	*fej*	*ṕä-*	** päŋe*
tree	*puu*	*puu*	—	*pu*	*pu*	—	*fa*	*ṕǟ*	** puve*
fish	*kala*	*kala*	*guolle*	*kol*	—	*kul*	*hal*	*xal′ä*	** kala*
house, hut	*kota*	*koda*	*goatte*	*kuðo*	*-ka*	*kat*	*ház*	—	** kota*
who	*ken*	*ke(s)*	*gi*	*ke*	*kin*	—	*ki*	*x́ib′ä*	** ken*
hand	*käte-*	*kät-*	*giettâ*	*kö*	*ki*	*köt*	*kéz*	—	** käte*
louse	*täi*	*täi*	*dik'ke*	*ti*	*toj*	*tögtəm*	*tetű*	—	** täjka*
know	*tunte-*	*tunde-*	*dow'dâ-*	—	*təd*	—	*tud*	*tumda-*	** tumte-*
give	*anta-*	*anda-*	*vuow'de-*	*omta-*	*ud-*	*ŏntas*	*ad*	—	** amta-*
eye	*silmä*	*silm*	*čal'bme*	*šinča*	*śin*	*sem*	*szem*	*sew*	** silmä*
heart	*sydäm-*	*südam-*	*čâððam-*	*šüm*	*śələm*	*səm*	*szív*	*sēj*	** śüðam-*
lap	*syli*	*süli*	*sâllâ*	*šəl*	*syl*	*jöl*	*öl*	—	** süle*
vein	*suoni*	*soon*	*suodnâ*	*šön*	*sən*	*jan*	*ín*	*tēn-*	** sōne*
mouse	*hiiri*	*hiir*	—	—	*šyr*	*junkər*	*egér*	—	** šiŋer*
ice	*jää*	*jää*	*jiegŋâ*	*ij*	*ji*	*jöŋk*	*jég*	—	** jäŋe*
blood	*veri*	*veri*	*vârrâ*	*vür*	—	—	*vér*	—	** vere*
water	*vete-*	*vet-*	—	*vüt*	*va*	—	*víz*	*jīd-*	** vete*
go	*men-*	*min-*	*mânnâ-*	*mija-*	*mun-*	*mən-*	*men-*	*min-*	** mene-*
one	*yhte-*	*üht-*	*ok'tâ*	*ikte*	*ət′ik*	*ĭt*	*?egy*	—	** ükte*
two	*kahte-*	*kaht-*	*guok'te*	*kok*	*kyk*	*kät*	*két*	—	** kakte*
three	*kolme*	*kolm*	*gol'bmâ*	*kum*	*kujim*	*koləm*	*három*	—	** kolm-*
four	*neljä*	*neli*	*njæl'lje*	*nyl*	*ńol*	*ńəlä*	*négy*	—	** neljä*
five	*viite-*	*viit-*	*vit'tâ*	*vič*	*vit*	*vet*	*öt*	—	** vit(t)e*
six	*kuute-*	*kuut-*	*gut'tâ*	*kut*	*kvajt*	*kut*	*hat*	—	** kut(t)e*

*Unattested, reconstructed form.

of consonants, Finno-Ugric must have had a fairly large number of vowels (nine to 11 are usually posited). One hypothesis is that the original vowel system was essentially like that of Finnish, which has eight vowel sounds: *i, ü, u, e, ö, o, ä, a* (*ü*—spelled *y* in the standard orthography—and *ö* are front rounded vowels, as in German; *ä* is a low front vowel, as *a* in "cat"). Hungarian has a similar system, although not all dialects have a separate *ä* sound, which is not distinguished from *e* in the orthogaphy. A second approach posits a Proto-Uralic vowel structure closely resembling that of Khant, with seven full vowels and three reduced vowels.

The early Finno-Ugric system of vowels most likely possessed quantitative vowel contrasts (long versus short, or full versus reduced). Such contrasts are present in Baltic-Finnic, Lapp, and Ugric and within Samoyedic; *e.g.,* Finnish *tulen* "of fire" and *tuulen* "of wind," *tuleen* "into fire," and *tuuleen* "into wind"; Hungarian *szel* "slice" and *szél* "wind," *szelet* "wind" (accusative case), and *szelét* "its wind" (accusative). The possibility of influence by neighbouring languages cannot be ruled out in the case of vowel length, because western Finno-Ugric languages have been in close contact with Slavic and Germanic languages with similar vowel contrasts, and the eastern languages form an areal group among themselves. The remaining languages lack vowel quantity and are in intimate contact with Russian, which has lost the original contrastive vowel quantity of Indo-European. The Izhma dialect of Komi, adjacent to Nenets, has superficial contrasts such as *pi* "son" versus *pī* "cloud," but this vowel length is the result of a change of an *l* at the end of the syllable to a vowel.

Stress. In numerous Uralic languages—including Finnish, Estonian, Hungarian, and Komi—stress is automatically on the first syllable of the word; it is likely that Proto-Uralic also had word-initial stress. Closely related to this initial stress is the apparent severe limitation on early Finno-Ugric noninitial vowels; the full range of contrasts was permitted only in the first syllable. In certain languages, such as Eastern Mari and the Yazva Komi dialect, stress is not bound to a given syllable, and determining the place of stress requires information concerning vowel quality as well; *e.g.,* Yazva *śībdinə* "to bind," *liććina* "to descend," *l′iśīna* "wood" (the two stressed *i*′s are phonologically tense; *ś, ć, l′* are palatalized consonants). Stress at the end of a word is also found; *e.g.,* in Eastern Mari and Udmurt. Nganasan has a mora-counting stress, falling on the third unit of vowel length from the end of the word (where short vowels count as one unit, long vowels as two).

Conditions for vowel harmony

Vowel harmony. Vowel harmony is among the more familiar traits of the modern Uralic languages. Although most Uralic scholars trace this feature back to Proto-Uralic, there is good reason to question this view. Vowel harmony is said to exist when certain vowels cannot occur with other specific vowels within some wider domain, generally within a word. For example, of the eight vowels of Finnish, within a simple word, any member of the set *ü, ö, ä* prohibits the use of any member of the set *u, o, a,* but *i* and *e* may occur with either set. That is, within a word, vowels that are either rounded, such as *ü, ö, u, o,* or low, such as *ä, a,* must agree with each other in frontness or backness. (The distinction is marked phonetically by putting two dots over the front vowels.) The unrounded front vowels, *i* and *e,* may occur with any of the other vowels. Thus, from *talo* "house" one may form *talossa* "in (the) house," but for *kynä* "pen" the comparable form is *kynässä* "in (the) pen"; similarly, *talossansa* "in his house" contrasts with *kynässänsä* "in his pen" and *talossansako* "in his house?" with *kynässänsäkö* "in his pen?", whereas *taloni* "my house" and *kynäni* "my pen" have the same ending because *i* can occur with either of the two sets of vowel classes. Hungarian has essentially the same system, differing only in certain minor details (short *e* is the front vowel counterpart of *a*); *e.g., asztal* "table," *asztalok* "tables," *asztalokban* "in the tables," but *föld* "land," *földök* "lands," *földökben* "in the lands." Similar though less general front–back vowel-harmony systems are found in given dialects of Mordvin, Mari, Mansi, Khant, and Kamas.

Frequently confused with the true harmony situations above are partial and total assimilations of vowels in adjacent syllables. These assimilations illustrate a universal tendency of vowel interaction and are of relatively recent origin; they are best held apart from the question of vowel harmony. Examples of vowel assimilations abound. In Finnish an unstressed *e* in the illative case ("place into") is totally assimilated to a preceding vowel, even across an intervening *h*: *talo* + *hen* becomes *taloon* "into the house," *talo* + *i* + *hen* yields *taloihin* "into the houses," *työ* + *hen* becomes *työhön* "into the work." The Hungarian allative case ("place to or toward which") shows an assimilation of the phonetic feature of lip rounding with front vowels in addition to the standard vowel harmony; thus, *ház-hoz* "to the house," *kéz-hez* "to the hand," *betu-höz* "to the letter." Apart from such nonharmony alternations, no support for rounding harmony is found in Uralic.

Possible influence of Turkic and other languages

Considered from an areal viewpoint, two aspects of Uralic vowel harmony must be considered. First, those languages that show productive or active vowel harmony, with the exception of Baltic-Finnic, have had recent Turkic neighbours whose languages exhibited vowel harmony. For languages such as Mansi and Khant, dialects with vowel harmony are located closer to Tatar groups. Second, the original homeland of Uralic lies in the centre of an enormous hypothetical areal grouping, labelled by some as the "Eurasian language union." The languages of this "union" are said to be characterized by two features: (1) the absence of a tonal accent (changes in pitch that change meaning, as is found in Chinese, Swedish, or Serbian) and (2) the contrast of plain and palatalized consonants (as in Russian). The phonetic basis for the consonantal contrast between nonpalatalized and palatalized is acoustically the same as the contrast of front and back vowels (*i.e.,* it involves the timbre of the second formant). Indeed, in Erzya Mordvin, vowel harmony and palatalization appear to be conditioned by essentially the same rules. Instead of seeking a genetic explanation of vowel harmony in Uralic, a somewhat more recent areal origin—in part under Turkic influence—must be considered. Of significance is the further consideration that, among the northwestern languages far from Turkic influence, it is precisely Lapp and the Baltic-Finnic Estonian and Livonian that do not have vowel harmony and that have developed special syllable-accent systems (thus, they lack both traits of the Eurasian union).

Consonant gradation. The alternation of consonants known as consonant gradation (or lenition) is sometimes thought to be of Uralic origin. In Baltic-Finnic, excluding Veps and Livonian, earlier single stops were typically replaced by voiced and fricative consonantal variants, and geminate (double) stops were weakened following a stressed vowel when the next syllable was closed; that is, *_p_ alternated with *_v_ and *_b_; *_t_, with *_ð and_ *_d;_ *_k_, with *ɣ and *_g;_ *_pp_ with *_p;_ and so on. Finnish thus shows pairs such as *mato* "worm" and *madon* "of the worm," *matto* "rug" and *maton* "of the rug," *poika* "boy" and *pojan* "of the boy," *lintu* "bird" and *linnun* "of the bird," *selkä* "back" and *selän* "of the back." Estonian shows the same type of alternations, with considerable difference in detail; *e.g., sada* "hundred" and *saja* "of a hundred," *madu* "snake" and *mao* "of the snake," *lind* "bird" and *linnu* "of the bird," and *selg* "back" and *selja* "of the back." Most of the Lapp languages exhibit similar alternations, but the process applies to all consonants and, moreover, works in reverse—single consonants are doubled in open syllables; *e.g., čuotte* "hundred" and *čuoðe* "of a hundred," *borra* "eats" and *borâm* "I eat." The change of *t* to *ð* however, is not a part of Lapp gradation but rather a general process that voices and weakens all single stops between voiced sounds (in this case, vowels).

Despite their essential differences, the Baltic-Finnic and Lapp gradations appear to be areally related. The Baltic-Finnic type, which represents a more plausible phonetic change, indicates that early Lapp may have acquired its gradation under Baltic-Finnic influence. The evidence within Baltic-Finnic points to a relatively late, post-Proto-Baltic-Finnic origin. The existence of analogous consonant weakening in various Samoyedic languages (Nganasan, Selkup) is the result of independent innovation.

Length variations in Estonian sounds

Syllable-accent structures. Closely related to the gradation phenomena is the development of syllable-accent structures in Estonian, Livonian, and Lapp. Estonian is known for its unique quantity alternations of three contrastive vowel and consonant lengths—thus, *vara* "early" versus *vaara* "of the hillock" (*aa* = long *ā*) versus *vaara* "hillock (partitive)" (here *aa* = extra-long *â*); *lina* "linen" versus *linna* "of the city" (*nn* is pronounced as two short *n*'s) versus *linna* "into the city" (here *nn* is pronounced as long *n̄* plus short *n*; the contrast with the previous *nn* is not shown in the standard orthography). The extra-quantity contrast is in fact found with all stressed syllable types containing at least one vowel or consonant following its first vowel; thus, *taevas* "sky" (with short *e*) versus *taevas* "in the sky" (with long *ē*); *osta* "buy!" (with short *s*) versus *osta* "to buy" (with long *s̄*), whereas a two-syllable form such as *osa* "part" (*o/sa*) with only a single vowel in the first syllable is incapable of such a quantity contrast. A multitude of analyses of Estonian quantity have been proposed, although not all have recognized the phenomenon as a function of whole syllables bound to stress—in other words, that it is an accent phenomenon. One orthographic dictionary (by E. Muuk), for example, utilizes this principle, placing a grave accent mark before syllables with extra quantity. Otherwise, Estonian orthography marks the three degrees of duration only for stops: *b, d, g* indicate single short (voiceless lenis) stops (*tuba* "room"); *p, t, k* are plain geminates, or double consonants (*tupe* "of the sheath"); and *pp, tt, kk* mark extra-long geminates (*tuppa* "into the room," *tuppe* "into the sheath"). Because the extra quantity is in part tied to an original open next syllable, it frequently operates together with gradation alternations; *e.g., linnu* "of the bird" versus *lindu* "bird (partitive)," with extra quantity.

The syllable quantity accent in Lapp superficially resembles that in Estonian and, like the former, occurs only under stress and is in part conditioned by the openness of the next syllable. In North Lapp (Utsjoki), alternations in paradigms involve three grades of quantity shaping: *mânâm* "I go" (*â* is a Lapp letter for a somewhat rounded *a*) versus *mânna* "he goes" versus *mân'ne* "goer"; *dieđam* "I know" versus *dietta* "he knows" versus *diet'te* "knower"; *juol'ge* "leg" versus *juolge* "of the leg." This series of contrasts shows a three-stage decrease in initial-vowel duration and a three-stage increase in the duration of the first consonant after the first vowel or vowels. The other northern and eastern Lapp languages display similar alterations, but there is considerable diversity in the phonetic details.

Proto-Uralic sentence structure

Grammatical characteristics. The grammatical structures of the various Uralic languages, despite numerous superficial differences, generally indicate a basic Proto-Uralic sentence structure of (subject) + (object) + main verb + (auxiliary verb)—the parenthesized elements are optional, and the last element is the finite (inflected) verb, which is suffixed to agree with the subject in person and number. This pattern has been best preserved in the more eastern languages, especially Samoyed and Ob-Ugric; *e.g.,* Nenets *ťiky pevśumd'o-m saravna t'eńe-vaʔ* "we well remember that evening" (literally, "that evening-[accusative] well remember-we"); Mari *joltaš-em-blak lum tol-mə-m buč-aš tüŋal-ət* "my friends begin to wait for the coming of snow" (literally, "friend-my-[plural] snow coming-[accusative] wait-to begin-they"). This order is common but optional in the languages of central Russia. Lapp, Baltic-Finnic, and Hungarian now show the typical European subject-verb-object order; *e.g.,* Finnish *isä osti talo-n* "father bought a house (-genitive)," Hungarian *János keres egy ház-at* "John seeks a house (-accusative)." Although these more western languages have relatively "free" word order, the object precedes the verb only for special emphasis; *e.g.,* Hungarian *János egy házat keres* "John is looking for a *house* (and not something else)," Estonian *ma ta-lle nuia ei anna* "I won't give *him* a *club*" (literally, "I him-to club not give"). Estonian sentence structure somewhat resembles that of German, with its tendency to place the finite verb in second position while the rest of the verb complex remains at the end of the sentence; *e.g., mehe-d ol-i-d ammu koju jõud-nud* "the men had got home long ago" (literally, "man-[plural] be-[past]-they long-ago home get-[past participle]").

In place of a verb "have," the Uralic languages use the verb "be," expressing the agent in an adverbial (locative or dative) case; *e.g.,* Finnish *isä-llä on talo* "father has a house" (literally, "father-at is house"), Hungarian *Jánosnak van egy ház-a* "John has a house" (literally, "John-to is one house-his"). In Proto-Uralic the copula verb "be" was lacking in simple predicate adjective or noun sentences, although the predicate was probably marked to agree with the subject. The following Hungarian sentences reflect this situation: *a ház fehér* "the house [is] white," *a ház-ak fehér-ek* "the houses [are] white." In Nenets and Mordvin such nonverbal predicates are conjugated for subject agreement and tense in the manner of intransitive verbs; *e.g.,* Nenets *mań xańenadmʔ* "I am a hunter," *pydaŕiˀ xańenadiʔ* "you two are hunters„" *mań xańenadamź* "I was a hunter," *pydaraʔ xańenadać* "You (plural) were hunters." Otherwise, a wide range of grammatical usage is found. In Baltic-Finnic and Lapp the use of a copula verb is obligatory, in Permic it is optional, and in Hungarian the copula is absent only in the 3rd person ("he, she") in a nonpast tense.

Use of negative sentences

Negative sentences in Proto-Uralic were indicated by means of a marker known as an auxiliary of negation, which preceded the main verb and was marked with suffixes that agreed with the subject, and perhaps tense. This is best reflected in the Finnic and Samoyedic languages; *e.g.,* Finnish *mene-n* "I go," *e-n mene* "I don't go," *mene-t* "you go," *e-t mene* "you don't go." Ugric employs undeclined negative particles (*e.g.,* Hungarian *nem*), and in Estonian only negative imperative forms are still conjugated, although colloquial Estonian has initiated a tense distinction; *e.g., ma/sa ei tule* "I/you don't come" and *ma/sa e-s tule* "I/you didn't come."

In Proto-Uralic questions were formed with interrogative pronouns, beginning with **k-* and **m-*, illustrated by Finnish *kuka* "who," *mikä* "what" and Hungarian *ki* "who, *mi* "what." Yes–no questions were formed by attaching an interrogative particle to the verb, as in Finnish *mene-n-kö* "am I going?", *e-n-kö minä mene* "am *I* not going?" (in Finnish, the verb also shifts to initial position). The use of intonation (changes in pitch) in interrogative sentences is currently widespread. In Hungarian it is the only way to form direct yes–no questions, although in indirect questions a particle *-e* is used; *e.g., a házak fehérek?* (with sharply rising intonation of the next to the last syllable, dropping again on the final syllable) "are the houses white?", *nem tudom, fehérek-e a házak* "I don't know whether the houses are white."

Conjunction, the connecting of clauses, phrases, or words, was formerly without the aid of specialized conjunctions. In the modern languages the conjunctions are largely borrowings from Germanic (Finnish *ja* "and") and Russian (Mari *da* "and; in order to," *a* "but," *ńi . . . ni* "neither . . . nor," *jesle* "if"). Both coordination and subordination in sentences are marked by a wide range of constructions, especially by means of infinitive verbs, participles, and gerunds; e.g., Mari *keče peš purgəžan poranan ulmaš* "the weather was very stormy and snowy" (literally, "weather very stormy snowy was"), *ača-ž abašt* "their father and mother" (literally, "father-his mother-their"), *nuno batə-ž-den* "he and his wife" (literally, "they wife-his with"); Finnish *kirja-n lue-ttu-a-ni* "when I had read the book, . . ." (literally, "book-[genitive] read-[past passive participle-partitive case]-my"), *luke-akse-ni kirja-n* "in order for me to read the book" (literally, "read-to-[translative case] my book-[genitive]").

Case suffixes and postpositions were and are used to show the function of words in a sentence. Prefixes and prepositions were unknown in Proto-Uralic. Adjectives, demonstrative pronouns, and numerals originally did not show agreement in case and number with the noun, as is still the case in Hungarian; *e.g., a négy nagy ház-ban* "in the four large houses." Finnish, however, has initiated a case–number agreement system much like that in neighbouring Indo-European languages; *e.g., neljä-ssä iso-ssa talo-ssa* "in the four large houses." The case system of the Proto-Uralic language contained an unmarked nominative case, an accusative, a case of separation (ablative), a locative (essive) case, and a case of direction (lative), plus possibly several others. The modern languages show a range of from three cases in Khant, six in Lapp, 14 in Finnish, up to 16 to 21 for Hungarian (the case status of several suffixes is debatable). The average number of cases is around 12. For the most part, these cases are the same for all nouns (nouns are not classified for gender; and 3rd person pronouns generally do not distinguish between "he, she"), singular and plural, and many are similar in function to English prepositions.

Cases in the modern Uralic languages

The distinction between a case and a preposition is often based on arbitrary and superficial criteria. Postpositions, preposition-like elements placed at the end of words, are generally more independent, and also function as adverbs. They often resemble inflected nouns (*e.g.*, Finnish *takana* "behind": *talo-n takana* "[at] behind the house [-genitive]," *talo-n taka-a* "from behind the house," *taka-osa* "back part").

Table 37: Case Endings in Several Uralic Languages

Finnish	Komi	Hungarian	Nenets	English translation
talo-ssa	*kerka-yn*	*ház-ban*	*xarda-xa-na*	"in (the) house"
talo-i-ssa	*kerka-jas-yn*	*ház-ak-ban*	*xarda-xa- ʔ-na*	"in (the) houses"
talo-sta	*kerka-yś*	*ház-ból*	*xarda-xa-d*	"from (the) house"
talo-i-sta	*kerka-jas-yś*	*ház-ak-ból*	*xarda-xa-t* (from *xa- ʔ-d*)	"from (the) houses"

The original case relationships of essive–lative–ablative form a three-way set of contrasts that has been extended into several parallel series of cases in the modern languages—perhaps under areal influence. For example, Finnish uses essentially the original three in relatively abstract functions (essive, a state of being, *-na*; translative, a change of state, *-ksi*; partitive, a case of separation, [-*t*]*a*), and also adds an *-s-* element to indicate internal relationship (_-ssa_ from *_s_ + _na_ "in"; _-hen,_ or a vowel + _n,_ etc. from *_s_ + _ń_ "into"; _-sta_ "out of"), and an _-l-_ element to indicate external relationship (_-lla_ from *_l_ + _na_ "on, at," _-lle_ from *_l_ + _k_ "onto, to," _-lta_ "off of, from"). Hungarian has nine cases similarly organized into three series of three, the internal set of which (*-ben* "in," *-be* "into," *-ből* "out of") has recently developed from a noun with the meaning "intestines" (*bél*). In Finnish the personal pronouns are declined throughout on a pronoun stem; *e.g.*, *minä* "I," *minu-ssa* "in me," *minu-n* "me (genitive)," and so on. In Hungarian, however, only the nominative and accusative forms are formed this way, and the remaining cases are formed by adding the possessive suffixes to a form of the case marker (sometimes expanded); *e.g.*, *te* "you (singular)," *teged-et* "you (accusative)," *benn-ed* "in you," *belé-d* "into you," *belől-ed* "out of you."

Number in Uralic nouns

The inflection of nouns for number (singular and plural) in the Uralic languages is much looser than in the Indo-European languages. Suffixes for the plural in the various Uralic languages are so diverse as to suggest that early stages of Uralic did not possess a specialized number marker; *e.g.*, Finnish *-t* and *-i-*, Mari *-blak,* Komi *-jas.* A dual-plural distinction ("two" as opposed to "more than two") is found in Lapp, Ob-Ugric, and Samoyedic, but here again the specific elements cannot be traced to a common source. If Proto-Uralic had plural and dual suffixes, they were probably used only with the personal pronouns. In the modern languages personal pronouns often take a plural marker different from that of the nouns, and in Lapp the dual formation is restricted to pronouns and personal affixes.

The category of definiteness (like English "the") is marked in numerous ways in the modern languages and originally appears to have been tied to the manner of number marking in Uralic (plural being reflected by indefiniteness). Hungarian alone has a definite article, *a*(*z*), a demonstrative in origin; Mordvin has three sets of inflectional endings: indefinite, definite singular, and definite plural (*kudo-so* "in a house," *kudo-so-ńt′* "in the house," *kudo-t′ńe-sə* "in the houses"). Nearly all the more eastern members have a definite marker that is identical with the 3rd or 2nd person possessive suffix (Komi *kerka-ys/yd* "the house" or "his/your house").

In possessive constructions the possessor noun precedes the possessed noun, or in the case of a personal pronoun possessor, possessive suffixes are used; *e.g.*, Finnish *isä-n talo* "father's house" (*-n* = genitive, *talo-ni/si* "my/your house"; Hungarian *János ház-a* "John's house" (*-a* = possessive construction marker), *ház-am/ad* "my/your house." Although in earlier stages the possessive suffixes followed the case suffixes, more recent case formations (especially from original postpositions) have led to restructuring of this order; *e.g.*, Finnish *talo-i-ssa-ni* "in my houses," but Hungarian *ház-a-i-m-ban* "in my houses" (*-i-* = plural); Komi *kerka-yd-ly* "for your house" (*-yd-* = "your"), *kerka-śyd* "from your house," where two fixed orders coexist. The Proto-Uralic comparative construction was similar to the Finnish *talo-a iso-mpi* "house-from larg-er" (= "larger than a house"); *cf.* Hungarian *egy ház-nál nagy-obb* "house-by larg-er" (in dialects also *ház-tól* "house-from"); Komi *kerka dor-yś yǰyd-ǰyk* "house by-from larg-er." Parallel "than" type conjunctions are now common in the more western languages; *e.g.*, "larger than a house" in Finnish can also be expressed as *isompi kuin talo* (*kuin* = "than"), and in Hungarian *nagyobb mint egy ház* (*mint* = "than").

Compounding and derivation in word formation

The formation of nouns in Proto-Uralic included compounding (adding two or more words together) as well as derivation by the use of suffixes (word endings). In noun + noun constructions, including titles of address, the qualifying noun came first; *cf.* Hungarian *házhely* "house site," *Szabó János úr* "Mr. John Szabó"; Finnish *taloryhmä* "group of houses," *Sirpa täti* "Aunt Sirpa." The rich system of derived words in Uralic together with the various inflectional suffixes led to relatively long words; *cf.* Finnish *talo-ttom-uude-ssa-ni-kin* "even in my houselessness" (literally, "house-less-ness-in-my-even"), Hungarian *ház-atlan-ság-om-ban* "in my houselessness."

The Proto-Uralic verb was inflected for tense-aspect (*_-pa_ indicated "nonpast," *_-ka_ indicated "perfect nonpast; imperative," *_-ja_ indicated "past") and mood (*_-ne_ indicated "conditional-potential"). The use of auxiliary verbs to indicate tenses was unknown, although Lapp, Baltic-Finnic, and Hungarian now have essentially a Germanic-type tense system, with perfect formations based on the "be" verb; *e.g.*, Finnish *mene-n* "I go," *ole-n men-nyt* "I have gone" ("be-I go-[past participle]"), *men-i-n* "I went," *ol-i-n men-nyt* "I had gone," *men-isi-n* "I would go," *ol-isi-n men-nyt* "I would have gone." Under Germanic and Slavic influence both Estonian and Hungarian have developed separable verbal prefixes with adverbial and aspectual meanings; *e.g.*, Estonian *ära söö-* "eat (perfective)" and *ta sõ-i kala ära* "he ate the fish" versus *ta sõi kala* "he was eating fish," *ta hakkas kala ära söö-ma* "he began to eat (up) the fish"; Hungarian *meg-tanul* "learn" (perfective) and *János megtanul-t magyar-ul* "John learned Hungarian" versus *János tanult magyarul* "John was learning Hungarian," *János tanult meg angolul* "*John* learned English," *János nemetül tanult meg* "John learned *German*" (with special emphasis as indicated).

Proto-Uralic did not have specialized voice markers, such as the Indo-European passive; rather, the function of voice was interwoven with topicalization (a way of indicating the main subject of a sentence), emphasis, and definiteness of the subject and object as well as with verbal aspect. An indefinite subject of an intransitive verb or an indefinite object were marked with the ablative case (*_-ta_), but a definite object took the accusative marker (*_-m_) and other subject situations were unmarked (nominative). This system is best preserved in Finnish: *vesi* (-nominative) *juoksee* "the water is running" versus *vettä juoksee* "there is water running," *juon vede-n* "I will drink the water" (_-n_ is from older *_-m_) versus *juon vettä* "I drink water." (Note that aspect as well as tense is affected by these case distinctions.)

Subjective and objective conjugations

The widespread use of separate subjective and objective conjugations among the Uralic languages (as in Mordvin, Ugric, and Samoyedic) are the result of an original system for singling out the subject or object for emphasis (focus), and not simply a device for object–verb agreement (similar to subject agreement). For example, Nenets *tymʔ xada-v* "*I* killed a deer (focus on the agent)" versus *tymʔ xada-dmʔ*

"I killed a *deer* (focus on the object)," in which *-v* signifies "I . . . it" (the objective conjugation) and *-dm*ˀ signifies "I" (the subjective conjugation). Note also the objective forms *xada-n* "I killed [them]," *xada-r* "you (singular) killed [it]," *xada-d* "you (singular) killed [them]," and so on for nine possible subjects (three persons times singular, dual, plural) times two object numbers (singular and nonsingular [not actually distinguished with 3rd-person subjects]); and the subjective forms *xad-n* "you (singular) killed" and so on, for nine subject agreements. Hungarian similarly opposes definite and indefinite conjugations: two different sets of personal endings are used—one with transitive verbs with definite objects and the other elsewhere; *e.g., olvas-om/od a level-et* "I/you read the letter" versus *olvas-ok/ol egy level-et* "I/you read a letter." Along with its subjective and objective conjugations Khant has added a so-called passive conjugation (*cf. kitta-j-m* "I am being sent," *-j-* = "passive") as an extension of the earlier focus-topicalization system. Mari and Komi have two past tense formations with related function. Again, the westernmost languages have passive constructions similar to those in Slavic and Germanic.

Verbal derivation was richly developed already in Proto-Uralic with a wide variety of verbal nouns, infinitives, and participles. Each of the three tense-aspect markers was apparently used as a participial formative (*cf.* Finnish *lähde* from **läkte-k* "source," *lähtijä* "one who leaves," *lähtevä* from **-pa* "leaving"). Several of the modern languages have made extensive use of their native derivational processes to eliminate foreign loanwords; *e.g.,* for "telephone" Finnish has *puhelin,* which is derived from *puhel-* "talk," just as *soitin* "musical instrument" comes from *soitta-* "to play." The Uralic finite verb originally may have been based on participial constructions parallel to the noun plus predicate adjective sentences (like Hungarian *a ház fehér* "the house [is] white"). Thus, one may reconstruct sentences like **ema tumte-pa* "mother [is] knowing," **ema tumte-pa-ta* "mothers [are] knowing" (with subject number expressed only in the predicate [agreement]) to explain the close similarity of participial and finite verb constructions such as Estonian *tundev ema* "knowing mother," *tundvad emad* "knowing mothers," *ema tunneb* "mother knows," *emad tunnevad* "mothers know." (R.T.H.)

ALTAIC LANGUAGES

The term Altaic includes the languages of the Turkic, Mongolian, and Manchu-Tungus language families. Named after the Altai Mountains, the languages are spoken by more than 105,000,000 people spread over an area that extends from the northeastern region of the Asian continent through the northern and northwestern provinces of China, Mongolia, Central Asia, southern Siberia, the Volga region, and Turkey, down to the Near East and the Balkan Peninsula.

Question of genetic relationship

On the basis of correspondences in vocabulary and structural similarities, several scholars have concluded that the Altaic languages are genetically related and thus use the term Altaic as the name of a language family. Others consider the correspondences and similarities to be only traces of ancient contacts or areal (regional) convergences and reject the genetic relationship of these languages. For them, the term Altaic refers to a language group that displays a certain historical and typological unity. Still others admit the possibility that the languages are related but do not consider the relation provable on the basis of present knowledge.

Attempts have been made to determine further genetic relations of the Altaic languages. The examination of structural and etymological correspondences gave rise to the hypothesis that the Uralic (including Finnish and Hungarian) and Altaic languages are related; this hypothetical group is called the Ural-Altaic language family. But the Ural-Altaic theory has an ever-decreasing number of supporters. Some adherents of the Altaic relationship are seriously considering the possibility of genetic ties to certain ancient strata of the Korean language.

As a result of the historically very active role of the Altaic peoples, their languages are found spread over a large geographical area. Altaic peoples overpowered their present territories in Asia and Europe in succeeding waves. At one time or another they have played a dominant role in the history of China, Iran, Byzantium, the Arab caliphate, and India, and their migrations had effects on the history of eastern Europe as well. Their present political formations in Inner and Central Asia and in the Near East are of great importance.

Languages of the Altaic group. The following tables give information about the literary languages and the major dialects of the three language families, their geographical division, and the number of speakers. Historical connections of the individual languages are also indicated. The names of subdivisions of the language families are of geographical or historical origin, in conformity with scientific traditions.

Because the name Turkish usually referred to a single language—Ottoman Turkish, the idiom spoken in Turkey, the Balkan countries, Cyprus, and elsewhere—the term Turkic is preferred in scientific literature for the whole language family.

On the evidence of certain archaic features in which Chuvash corresponds to the Mongolian languages, adherents of the theory of an Altaic relationship tend to look upon Chuvash as a separate unit among the Turkic and Mongolian languages.

Mutual intelligibility is possible within each of the Turkic language groups, but there are major difficulties in

Table 38: Turkic Languages*

	number of speakers
Common Turkic, or z-Turkic, languages	
Southeast (Uighur or Chagatai) group	
Uzbek (U.S.S.R. 12,925,000, Afghanistan 1,390,000, China 7,000)	14,322,000
Uighur (China 5,570,000, U.S.S.R. 187,000); the	5,757,000
literary language, is also used by the Salars (China 35,000) and Yellow Uighurs (China 4,000)	39,000‡
Southwest (Oğuz or Turkmen) group	
Turkish (Turkey 39,515,000, W. Germany 1,500,000, Balkan countries 940,000, Cyprus 115,000, U.S.S.R. 82,000, Arab countries 30,000, other countries 22,000)	42,204,000
Azerbaijani (Iran 6,646,000, U.S.S.R. 5,460,000, Afghanistan 45,000);	12,150,000
other Oğuz dialects spoken in Iran (near Azerbaijani)	970,000‡
Turkmen (U.S.S.R. 2,107,000, Iran 590,600, Afghanistan 330,000, Arab countries 243,000, Turkey 92,000)	3,363,000
Gagauz (U.S.S.R. 157,000, Bulgaria 5,000)	162,000
Northwest (Kipchak) group	
Karaim (U.S.S.R.)	6,000
Kumyk (U.S.S.R.)	226,000
Karachay-Balkar (U.S.S.R., Karachai 85,000, Balkar 44,000)	129,000
Tatar (U.S.S.R. 5,479,000, Romania 24,000, Turkey 10,000, China 10,000, Bulgaria 11,000, other countries 8,000)	5,542,000
Bashkir (U.S.S.R.)	928,000
Kazakh (U.S.S.R. 6,596,000, China 871,000, Mongolian People's Republic 73,000, Afghanistan 5,000)	7,545,000
Kara-Kalpak (U.S.S.R. 307,000, Afghanistan 4,000)	311,000
Kirgiz (U.S.S.R. 1,946,000, China 124,000, Afghanistan 46,000)	2,116,000
Nogay (U.S.S.R.)	55,000
Northeast (Siberian or Altai) group	
Khakass (U.S.S.R.); the literary language	57,000‡
used also by the Shors (U.S.S.R.)	58,000
Altai (formerly Oirot; U.S.S.R.)	45,000‡
Tuvinian (U.S.S.R. 166,000, Mongolian People's Republic 21,000)	187,000
Yakut (U.S.S.R.)	316,000
Khalaj (Iran)	25,000
r-Turkic language	
Chuvash (U.S.S.R.)	1,445,000
Total	97,958,000

*The geographical location of the language area and data specifying the distribution and number of speakers are given in parentheses. †1981 estimate. ‡Latest available figure.

Table 39: Mongolian Languages*

	number of speakers†
Western group	
Kalmyk (U.S.S.R.)	135,000
Oyrat (Mongolian People's Republic 49,000, China 110,000)	159,000
Mongol (Afghanistan)	50,000‡
Eastern group	
Mongol (Mongolian People's Republic 1,288,000, China 2,487,000)	3,775,000
Buryat (U.S.S.R. 321,000, Mongolian People's Republic 37,000)	358,000
Daghur (China)	72,000
Monguor (China)	93,000
Santa (China)	259,000
Total	4,901,000

*The geographical location of the language area and data specifying the distribution and number of speakers are given in parentheses. †1981 estimate. ‡Latest available figure.

comprehension among the groups, and understanding is impossible among the speakers of the Yakut, Khalaj, and Chuvash tongues and those of the so-called common Turkic languages.

The name Mongolian is used for the members of the language family taken collectively, as well as for certain idioms spoken in the Mongolian People's Republic (Outer Mongolia) and China (Inner Mongolia). To make an exact distinction between the family and these languages, the latter are sometimes designated by the name Mongol. The reverse use of these terms, however, is also known.

Kalmyk, Mongol, and Buryat (Buriat) are contemporary literary Mongolian languages. There are no great differences among these languages, and mutual intelligibility thus exists, though the possibility of mutual comprehension is decreasing among speakers of isolated and nonwritten dialects.

The census datum for Manchu has an ethnic rather than a linguistic value. It is generally assumed that Manchu must be considered as the language of a minor group, and it is, in fact, almost an extinct language.

There are considerable differences between the northern and southern groups of this family and among the individual languages and dialects of the southern group. Despite the immense geographical area over which they are spread, languages and dialects belonging to the northern group are less differentiated.

LINGUISTIC CHARACTERISTICS

Sound harmony

Phonology. Simplicity is the characteristic feature of the phonological (sound) systems of the Altaic languages. Richness in vowels is compensated for by the scantiness of consonants. In addition, clusters of consonants are rare. Sound harmony is one of the most characteristic features of the Altaic languages. Its essence is that in an individual word only certain combinations of vowels may occur, and, in the case of certain consonants, specific variants must be selected according to the type of vowels in the word. The best known forms of vowel harmony are palatal and labial harmony; *i.e.,* only back (velar) or front (palatal) vowels and rounded (labial) or unrounded (nonlabial) vowels may occur in a word. For example, in Turkish, back vowels include *a, ı, o, u;* front vowels include *e, i, ö, ü;* rounded vowels are represented by *ö, ü, o, u* and unrounded vowels by *a, ı, e, i.* This harmony extends to the suffixes, too, producing variants when added to words belonging to different vowel classes. The change in the vowels of the last two syllables (both suffixes) in the following Turkish words is notable: *at-lar-da* "on the horses," *it-ler-de* "on the dogs," *gör-ür-üm* "I see," *getir-ir-im* "I bring."

Table 40: Manchu-Tungus Languages*

	number of speakers†
Southern (Manchu) group	
Manchu (China)	4,145,000
Nanai (U.S.S.R. 6,000, China 1,000)	7,000
Other dialects (U.S.S.R. 12,000, China 33,000)	45,000
Northern (Tungus) group	
Evenk (U.S.S.R. 12,000, China 7,000, other countries 1,000)	20,000
Lamut, or Even (U.S.S.R.)	7,000
Total	4,224,000

*The geographical location of the language area and data specifying the distribution and number of speakers are given in parentheses. †1981 estimate.

Sound harmony in the Altaic languages is the product of a complicated historical development. The form of vowel harmony varies from one language to another. Sound harmony is less developed in the Manchu-Tungus group, in which it may occasionally reflect an ancient state or special aberration. Adherents of the Altaic relationship maintain that the ancient state of these languages has been characterized by a dynamic stress on the first syllable, while musical tone on the end syllable was the result of a later development. The ancient state seems to be reflected in the contemporary Mongolian tongues and a group of Manchu-Tungus languages; in Turkic and in another group of the Manchu-Tungus languages, stress shifted to the end syllable.

Grammar. From the point of view of grammatical features, there is a great difference between the Altaic languages and the Indo-European languages (which include English, French, and German). Altaic languages are agglutinative; that is, word formation consists of adding suffixes to the word roots. The root may take a large number of suffixes (*e.g.,* Turkish *gel-me-dik-ler-i-nden* "because of their not coming"), though the order of the combination of suffixes is governed by strict rules. While certain suffixes in the Indo-European languages may perform several functions (*e.g.,* Latin *casarum* "of the houses," in which the suffix *-arum* indicates a feminine, plural, genitive form), in the Altaic languages these functions are distributed among separate suffixes (*e.g.,* Turkish *ev-ler-in* "of the houses," which consists of a stem plus a plural suffix plus a genitive suffix).

Altaic languages lack definite articles (such as English "the") and grammatical gender (*e.g.,* German masculine, feminine, and neuter words), and use the singular form of nouns with numerals (*e.g.,* "two girl"). In the Manchu-Tungus languages, a distinction is made in the 1st person plural of the personal pronoun between exclusive ("we"—without you) and inclusive ("we"—with you). The same peculiarity is found in Mongolian and, from the Mongolian influence, in the Turkic languages of Siberia. There are no prepositions in the Altaic languages, but postpositions are used very frequently. For example, an Altaic speaker does not say the equivalent of "until night" or "for John" but rather "night until" and "John for." In the Tungus languages, nominal (noun) inflection has approximately 20 cases, while the Manchu languages show a simpler state, with about five to eight cases. Turko-Mongolian languages also show a simple case system: they have about six to ten cases. Because the plural is expressed by independent suffixes, the case suffixes are identical in singular and plural.

Parts of speech in the Altaic languages

Parts of speech in the Altaic languages are less differentiated than they are in the Indo-European languages; essentially, only nominals (words used as nouns), verbs, and particles (words that are neither nouns nor verbs) may be considered as separate parts of speech. In contrast with the Tungus languages, the Turkic and Mongolian languages do not possess any morphological means, such as specific suffixes, to distinguish between nouns and adjectives. Altaic languages are rich in verbal nominals (nouns, adjectives, and adverbs derived from verbs), as well as in nouns of action, participles, and gerunds; these perform important grammatical functions in Altaic sentences. A conspicuous ancient Altaic peculiarity is that verbal predicative expressions (finite verbs) are of verbal-nominal origin; that is, these verbal formations are in most cases verbal nouns in predicate positions and express their personal relations with added personal or possessive pronouns: *e.g.,* Old Turkic *bil-t-im* "I knew" is literally "having known my," and Old Turkic *bil-ir-biz* "we know" is literally "knowing we." In Altaic languages, there is no word that corresponds to the English "to have." To express this concept, a periphrastic construction (circumlocution) is used; *e.g.,* an Altaic speaker might say "The book is with me" rather than "I *have* the book."

Despite major differences in syntax, there are also significant common Altaic features. The modifier (attribute, appositive), for example, always precedes the modified word (*e.g.,* a noun), and the noun in the genitive case (the possessive) precedes the possessor (*e.g.,* "captain of the ship" is "of the ship, captain"). In the Tungus languages the attribute agrees in number and case with the modified noun. This feature is completely lacking in the Turkic languages, though agreement in number existed in the early periods of Mongolian.

In Altaic sentence structure, subordinate clauses that begin with conjunctions or relative pronouns ("The man *who was bald*") are alien; wherever they appear, they indicate foreign influence. Instead of subordination, coordination by verbal nominals (nouns) serves as a basis for sentence construction. The following Turkish sentences illustrate this feature. *Ahmedin bize gelmesini isterim,* translated as "I want Ahmed to come to us," is literally "I want Ahmed's coming to us"; *Şehre giden otobüs gecikti* "The bus, which goes to the town, was late" is literally "The bus going to the town was late"; *Eve varınca hemen telefon etti* "As soon as he arrived home he telephoned" is literally "Arriving home, he immediately telephoned." As a result of a more varied system of morphology, the Manchu-Tungus languages are characterized by a higher degree of syntactic freedom, while the Turkic and Mongolian languages have a higher degree of syntactic constraints (requirements for the positions of the elements of a sentence) and a simpler system of morphology.

Altaic languages have changed at a slow rate. The archaic character of the contemporary language is particularly conspicuous in the Manchu-Tungus and Mongolian languages. Because of this difference in historical character, one language may provide useful information for the history of another in regard to the development of earlier, supposedly common features.

Vocabulary. During an immense span of space and time, the vocabulary of the Altaic languages has shown a high degree of receptivity. Opponents of the genetic Altaic relationship look upon the ancient common elements of the vocabulary of these languages as loanwords resulting from once-frequent contacts between Turkic and Mongolian and between Turko-Mongolian and Manchu-Tungus. These correspondences in vocabulary show closer ties between the Turkic and Mongolian languages than between the Turko-Mongolian and Manchu-Tungus languages. The oldest loanwords of the vocabulary are connected with the Semitic, Indo-European, and Uralic languages. In the eastern part of the Altaic language area, many loanwords came from the Chinese to the Old Turkic and Middle Mongolian languages, as well as to contemporary Turkic and Mongolian, and also to the Manchu-Tungus languages, particularly to Manchu.

Oldest loanwords from Semitic, Indo-European, Uralic

Through the medium of religious texts translated from Sanskrit, Tocharian, and Sogdian, many loanwords found their way to Uighur and from there to Mongolian. Mongolian also had contacts with the Tibetan language through the Buddhist literature and through certain dialects. A strong linguistic interaction may be observed among the local Turkic and Mongolian languages in Mongolia and southern Siberia and among the Turkic and Iranian languages in Central Asia.

A number of Mongolian words were adopted into the Yakut language, which in turn had contact with the Tungus languages. The spread of the Islāmic religion paved the way for Arabic and Persian loanwords to enter the Turkic languages. Because of relations among the Ottoman Empire, Turkey, and the Western European civilizations, a considerable number of French, Italian, German, and English words were borrowed into Turkish. Russian provided the major part of the vocabulary of modern civilization for the Turkic languages of the U.S.S.R., although there was also considerable influence from other spheres. The Balkan, Anatolian, Caucasian, and certain Uralic languages had only minor influence on the Turkic languages.

In certain Turkic languages, the influence of linguistic contacts goes far beyond the simple borrowing of words. Although in the East the development of the Buddhist translation literature was followed by a comparatively modest linguistic influence by the source languages, the Arabic and Persian influence that followed the adoption of Islāmic religion and culture proved very strong. An avalanche of Arabic and Persian words flooded the vocabulary of the literary languages of the settled Islāmic-Turkic peoples and also filtered into the dialects. This layer of loanwords also led to changes in the structure (*i.e.,* the phonological system and sound harmony) of certain languages. Iranian influence gave rise to similar profound changes in Uzbek. Currently, Russian influence has been greatly intensified by the continuous contact taking place within an identical political and cultural framework and also by the increasingly widespread bilingualism of the Turkic peoples. This process is reflected in the growing number of calques (loan translation words), as well as in syntactic (sentence) structure.

In the course of linguistic contacts, the Altaic languages naturally also exerted a considerable influence on other languages; *e.g.,* Persian, Russian, Finno-Ugric, the Balkan languages.

Earliest Turkic records

Writing systems. The Altaic peoples used a variety of writing systems. The first-known writing of the Turkic peoples is the runic script (named after the Germanic runes), which is most probably of Semitic origin, particularly Aramaic. The first dated texts originate from Inner Asia in the 8th century; they are called the Orkhon inscriptions. Archaeological investigations indicate that this early script was spread over an extremely large area. The Turkic peoples (mainly Uighurs) who settled in the Tarim Basin used Brāhmī and Manichaean scripts in the 8th to 10th century, as well as the Uighur writing, as it developed from the Sogdian cursive writing, from the 8th to 17th century. Nestorian (Christian) Turkic peoples in Central Asia used the Estrangelo script of Syriac origin in the 13th and 14th centuries.

After the adoption of Islām in the 11th century, the Arabic script became the most widespread writing system among the Turkic peoples. In Anatolia, certain linguistic groups or minorities used the Greek alphabet (16th–20th centuries) and the Armenian script (17th–20th centuries). The latter was also employed by the Turkified Armenians in the Ukraine in the 16th and 17th centuries. Between the 16th and 19th centuries, Karaites used the Hebrew script; Latin and Cyrillic writing was introduced to them in the 19th century. In 1929 the Arabic script was replaced in Turkey by Latin writing. Among the Turkic peoples of the U.S.S.R., the Latin script was introduced in the 1920s; in the 1930s it was replaced by the Cyrillic script. Turks living in China, Iran, and the Arab countries still use the Arabic script.

Mongolian script developed from Uighur writing in the 12th century. After some Latin antecedents, it was replaced by the Cyrillic script among the Buryats in 1938 and among the Mongols in 1944. Mongols living in China still use the Mongolian script. The 'Phags-pa or Pa-sse-pa alphabet, introduced on the state's initiative in 1269–72, was based on the Tibetan script and did not take root. It became extinct in a few decades. From the 17th century, the Oyrats used a reformed version of the Mongolian writing, called Oyrat script. After 1917 this writing system was replaced among the Kalmyks by the Cyrillic alphabet; the Latin alphabet was introduced in 1931, and in 1937 the Cyrillic script was reintroduced. Oyrats living in China still use the Oyrat script.

The Manchu script, a reformed or adapted version of Mongolian writing, came into existence in the 17th century; it is probably still used in private life in China. For Manchu-Tungus people living in the U.S.S.R., the Cyrillic script was introduced in 1937–38 after some experiment with the Latin alphabet.

The writings of the Mongolian-speaking Khitans, who ruled in China from the 10th to the 12th century, and those of the Juchens, who ruled in China in the 12th and 13th centuries and who spoke a Manchu-Tungus language, have not yet been definitely deciphered.

HISTORY

Original habitat of the Altaic peoples

The original habitat of the Altaic peoples is supposed to have been on the steppe area bordering on the Altai and Ch'ing-hai mountains between Tibet and China; in the north, it probably stretched to the Siberian taiga region adjacent to the steppe. It is assumed that Tungus peoples lived in the northern and northeastern parts of this area and the Mongols in the central and southeastern parts. Turkic peoples probably inhabited the northwestern and western parts, while the southern and southwestern region was occupied by Hunnic groups. Although the Hun language is not known, Huns are supposed to have spoken an Altaic idiom. In historical times, the Uralic, Indo-Iranian, Sino-Tibetan, and Korean languages have bordered on this Altaic language area.

Advocates of a genetic relationship between the Altaic languages believe that it was probably in the area delineated above that the Proto-Altaic language (*i.e.,* the parent language of the modern Altaic tongues) began to divide into the currently known main groups. Those who oppose the Altaic theory maintain that this area was the scene of contacts among the individual language groups. In either case, from this original habitat, the Altaic peoples moved to their present territories. In the steppe migrations, which were natural consequences of a nomadic way of life and which dramatically changed the ethnic and linguistic map of central Eurasia, the Turkic and Mongolian peoples played the leading role. The Manchu-Tungus peoples' separate migration toward the northeastern Asian taiga and tundra might have taken place in an early period. The south- and southwest-bound migrations of the southern Manchu-Tungus peoples were restricted to the East Asian regions.

The Turkic languages. *Division into two linguistic groups.* It was probably at a very early date that the Turkic languages split into the *z* and *š* group (commonly called *z*-Turkic) and *r* and *l* group (commonly called *r*-Turkic); these groups are so known because of certain regular phonetic contrasts involving those sounds, *e.g., z*-Turkic *yüz,* "hundred," contrasts with *r*-Turkic *śĕr* "hundred"; *š*-Turkic *yaš,* "age," contrasts with *l*-Turkic *śul* "age." The first traces of the *r* and *l* division appeared on the south Russian steppes. It is assumed that the Huns also were speakers of an *r*- and *l*-type Turkic language and that their migration was responsible for the appearance of this language in the West. The *r*- and *l*-type language is now documented only by Chuvash, a language considered as a descendant of a Volga-Bulgarian language. The rest of the Turkic languages are of the *z*- and *š*-type.

The split of the *z*- and *š*-type Turkic languages, which appeared first in Inner Asia and later developed in the presently known language groups and languages, evolved over several centuries.

Literary languages and linguistic records. As was stated above, the earliest known records of the Turkic languages are runic inscriptions left by the Turks in Mongolia. The Uighurs overthrew the Turkish Empire there, and, after the disintegration of their empire in 840, they settled down in the Tarim Basin, where exposure to the Manichaean and Buddhist religions gave rise to a sizable literature, written in various scripts, that flourished from the 9th to the 14th century. In the 11th century, during the rule of the Qarakhanids (the first Turkic representatives of Islām), a new literature came into existence in eastern Turkistan. These literary efforts clearly show the intensification of Arabic and Persian influence, which culminated in the Khwārezmian literature (13th–14th centuries) and the Chagatai literature (15th–16th centuries), as well as in the postclassical products of Chagatai (17th–19th centuries). (Khwārezmian and Chagatai are Middle Turkic languages that possess significant literatures.) Despite the influence exerted by other Turkic language groups, these Inner and Central Asian literary languages exhibit an organic continuity; they might be considered as the antecedents of the contemporary Uzbek and New Uighur literary languages.

Arabic and Persian influence on Turkic literature

The main movement of the Oğuz branch of Turkic peoples from western Turkistan was toward the Caucasus, Iran, Anatolia, and the Balkans. The first continuous records of this language group come from 13th-century Anatolia. Its linguistic material is linked with the contemporary literary language of Turkey (*i.e.,* Turkish). The Azerbaijani and Turkmen languages also have long literary traditions. The first records of the Azerbaijani language date from the 14th century; the Turkmen language came into use only gradually, because the Chagatai language was commonly used instead by writers.

The migration of the Kipchak Turks was toward the south Russian steppes. A 14th-century textbook known as the *Codex Cumanicus,* used by missionaries, constitutes the first record of the languages of these peoples. Linguistic records (*e.g.,* dictionaries, grammars, literary works) of the Kipchaks, who as mercenaries went to Egypt and Syria where they also came into power, date from the 13th to 17th century. Armenian Turks of the Ukraine left remarkable written material from the 16th and 17th centuries. All these Kipchak linguistic records, although they represent languages that have disappeared, are still connected indirectly with the contemporary Kipchak literary languages. Because of the cultural conditions of these peoples, the first records of most of these languages originated in the 20th century.

The remainder of the contemporary Turkic literary languages also developed in the 20th century.

The diffusion of the Turkic peoples over the centuries and the changes in their political and cultural centres were accompanied by the emergence of a number of literary languages. These literary languages are not necessarily the direct antecedents of the contemporary—and recently developed—literary languages. In most cases, only their membership in the same language group (Uighur, Kipchak, Oğuz) may be established, thus indicating an indirect historical connection.

Stages of the languages. On the basis of linguistic records, Turkic languages that were spoken from the 8th to the 10th century are termed Old Turkic, those of the 11th to 15th century are called Middle Turkic, and those of the 16th to 20th century are known as New Turkic languages. To specify the form of the Turkic languages that resulted from language reforms that were instituted in the mid-20th century, it is customary to use the term newest, or modern. The division into periods is based on external facts of these languages rather than the internal history and thus merely gives a practical chronological overview.

The Mongolian languages. Although the Mongolian script probably evolved as early as the 12th century, its earliest known record, the inscription of Yesunke, dates from 1225. The *Secret History of the Mongols,* which was written in all probability in 1240, has been handed down in Chinese transcription. As regards the development of spoken Mongolian, the history of Mongolian is divided into three periods: Old Mongolian (to the 12th century), Middle Mongolian (13th–16th centuries), and New Mongolian (17th–20th centuries). Written Mongolian, which was originally based on a Middle Mongolian dialect, was characterized for centuries by an increasing conservativism in regard to its relation to the spoken language. Its history is divided into such periods as preclassical (13th–16th centuries), classical (17th–20th centuries), and modern (20th century). Classical Mongolian is continued by the contemporary literary Mongol, which is based on the Khalkha Mongol dialect, and by the written language of Mongols living in China.

Stages of Mongolian

The Oyrats created a separate script and literary language in the 17th century. Although this separate language gradually decreased in importance among the eastern Oyrats from the second half of the 18th century, it survived among the western Oyrats (the Kalmyks), and its course of development flowed into the contemporary literary language.

It was not until 1931 that the Buryats began to develop their own literary language. Until then, literate groups had used literary Mongolian.

The Manchu-Tungus languages. Within the Manchu-Tungus language family, only the history of the Manchu language is documented by linguistic records. As a result of the Manchu rule in China (1644–1911/12), Manchu rose to the rank of official language in the country. The earliest Manchu records date from the mid-17th century. Manchu acted as the lingua franca in China's communications with other countries. The downfall of the Manchu rule in China and the gradual Sinicization of the Manchu population ended the influence of the Manchu language. Other literary languages of the Manchu-Tungus language family came to be in the 1930s.

As has been shown, the Altaic languages, unlike the Indo-European languages, have relatively very recent linguistic records—Turkic dating from the 8th century, Mongolian from the 13th century, and Manchu-Tungus from the 17th century. This situation puts limits on what can be learned from comparative historical investigations, making it extremely difficult to settle the problem of the Altaic relationship.

Altaic languages in the 20th century. The development of literary languages became a central problem in the independent political entities of the Altaic peoples that arose after World War I. The languages of the Altaic peoples in the U.S.S.R. underwent a veritable revolution in the 1920s and 1930s. Writing reforms were soon followed by the strengthening of literary languages on the basis of the individual national languages. For several illiterate peoples, new literary languages had to be created for the first time. In Turkey, written and spoken language reforms included, above all, a renewal of the vocabulary; this formed a significant objective of the Turkish president Kemal Atatürk's social program. In the ongoing process of modernization, a new literary language has come into existence. Considerable reform of writing and language also took place in Mongolia.

Creation of new literary languages

The linguistic policy pursued in the mid-20th century had an immense impact on the history of the Altaic languages. These efforts not only saved minor or nonwritten languages from extinction but strengthened the positions of major Altaic languages among the languages of modern civilizations. The linguistic consolidation has considerably contributed to enhancing the political and cultural significance of these peoples.

For information on the literature of the peoples who speak (or have spoken) Altaic languages, see CENTRAL ASIAN ARTS; ISLĀMIC ARTS.

(G.Ha.)

DRAVIDIAN LANGUAGES

The Dravidian language family, as known to date, consists of 23 languages spoken by more than 165,000,000 people in South Asia. In terms of population figures the major languages of the family may be listed in the following order: Telugu, 52,986,000; Tamil, 44,400,000; Kannada (Kannaḍa), also called Kanarese, 27,900,000; Malayalam (Malayālam), 27,500,000; Gondi, 2,460,100; Tulu (Tuḷu), 1,427,000; and Kurukh (Kuruk̲h̲), 1,358,000. The Dravidian languages are spoken in the Republic of India (mainly in its southern, eastern, and central parts), in Sri Lanka (Ceylon), and by settlers in areas of Southeastern Asia, southern and eastern Africa, and elsewhere. Brahui (Brāhuī), with 750,000 speakers in Pakistan, is isolated from all of the other members of the family. The four major languages—Telugu, Tamil, Kannada, and Malayalam—possess independent scripts and literary histories dating from the pre-Christian Era. Now recognized by the constitution of India, they form the basis of the linguistic states of Andhra Pradesh (established as the first Indian linguistic state in 1953), Tamil Nadu, Karnataka (formerly Mysore), and Kerala.

Official use of Dravidian languages

Of the Dravidian languages, Tamil has the greatest geographical extension and the richest and most ancient literature, which is paralleled in India only by that of Sanskrit. Its phonological and grammatical systems correspond in many points to the ancestral parent language, called Proto-Dravidian.

Nothing definite is known about the origin of the Dravidian family. There are vague indigenous traditions about an ancient migration from the south, from a submerged continent in what is now the Indian Ocean. According to some scholars, Dravidian languages are indigenous to India. In recent years, a hypothesis has been gaining ground that posits a movement of Dravidian speakers from the northwest to the south and east of the Indian Peninsula, a movement originating possibly from as far away as Central Asia. Another theory connects the Dravidian speakers with the peoples of the Indus Valley civilization. The Dravidian languages have remained an isolated family to the present day and have defied all of the attempts to show a connection with the Indo-European tongues, Mitanni, Basque, Sumerian, or Korean. The most promising and plausible hypothesis is that of a linguistic relationship with the Uralic (Hungarian, Finnish) and Altaic (Turkish, Mongol) language groups.

As an independent family, the Dravidian languages were first recognized in 1816 by Francis W. Ellis, a British civil servant. The actual term Dravidian was first employed by Robert A. Caldwell, who introduced the Sanskrit word *drāvida* (which, in a 7th-century text, obviously meant Tamil) into his epoch-making *A Comparative Grammar of the Dravidian or South Indian Family of Languages* (1856).

Languages of the family. Tamil is spoken by 39,400,000 people (1981 est.) in the Indian state of Tamil Nadu, by another 2,697,000 in Sri Lanka (Ceylon), by smaller numbers of people in Burma, Malaysia, Indonesia, and Vietnam (about 1,400,000), in East and South Africa (almost 250,000), and by still smaller numbers in Guyana and on the islands of Fiji, Mauritius, Réunion, Madagascar, Trinidad, and Martinique. The earliest literary monuments of the language belong roughly to the 3rd and 2nd centuries BC. There exist a number of local dialects, the major dialect regions being the northern and eastern areas combined, the western area, the southern area (split into at least four major dialects of Madurai, Tirunelveli, Nanjiland, and Ramnad), and Sri Lanka (Ceylon). Correlated with the social position of the speaker are a number of speech forms; a major division occurs between the Brahmin and the non-Brahmin varieties. In addition, there is a sharp dichotomy between the formal language and informal speech.

Dialects and forms of Tamil

Malayalam, which is closely related to Tamil, is spoken in the Indian state of Kerala by some 21,700,000 people. Possessing an independent written script, it also has a rich modern literature. There are at least three main regional dialects (North, Central, South) of Malayalam and a number of communal dialects.

In the Nīlgiris and adjacent regions, several minor tribes speak the following languages: Kota (1,400), Toda (1,145), Badaga (128,500), Irula (Iruḷa) (6,176). The less well-known languages of a number of other tribes may yet be established as independent members of the Dravidian family (*e.g.,* Kurumba, Paṇiya).

Kodagu (Koḍagu), a non-literary language of a mountainous region called Coorg, has 119,000 speakers.

Kannada (Kanarese), which is spoken by 25,700,000 people in the Indian state of Karnataka, exhibits a dichotomy between educated speech and colloquial Kannada; in the latter at least three social dialects are recognizable that may be characterized as Brahmin, non-Brahmin, and Harijan ("untouchable"). A number of regional dialects (among them are Dharwar, Bangalore, and Mangalore) also exist. Kannada has an orthography of its own and an important ancient and modern literature.

To the south of the Kannada territory, more than 1,400,000 people speak Tulu (Tuḷu), a South Dravidian language having no developed written literature.

Telugu (spoken by 52,986,000 people), the official language of the state of Andhra Pradesh, exhibits a dichotomy between the written and the spoken styles, in addition to a number of sharply distinct local and regional dialects (including Telangana, coastal area, Rayalaseema, and a "transitional" zone) and divisions between Brahmin, non-Brahmin, and Harijan speech. The language has its own

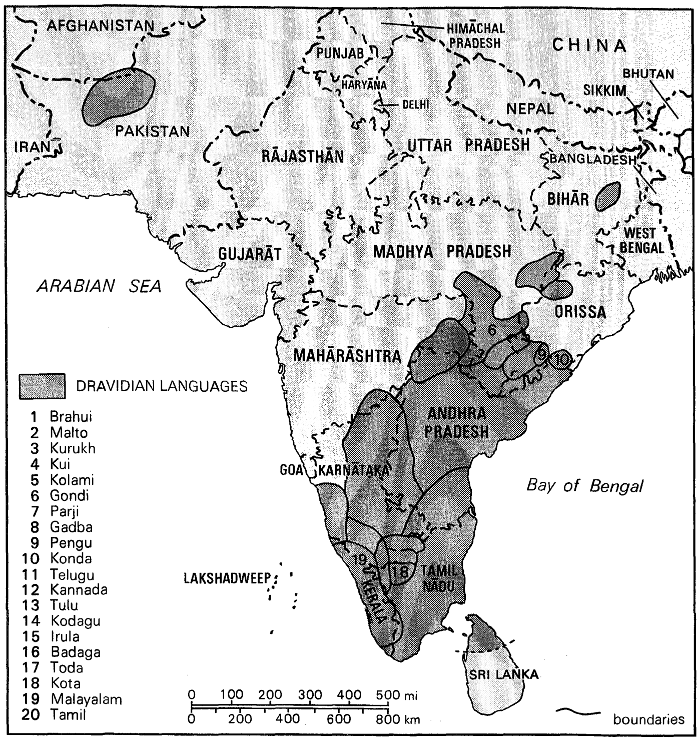


Figure 19: Distribution of the Dravidian languages.

Adapted from Ramanujan and Masica, "Toward a Phonological Typology of the Indian Linguistic Area," *Current Trends in Linguistics*, vol. 5 (1969); Mouton & Co., Publishers, The Hague

script, closely akin to that of Kannada, and an important literary tradition.

In the northern parts of Andhra Pradesh, two tribes speak Kolami (95,000 persons) with its dialect Naikri (Naikṛi), and Naiki (1,800), whereas Parji (170,000) is spoken in Bastar, Madhya Pradesh. In Orissa, Konda (Koṇḍa) is spoken by about 19,000 Konda Doras, and about 3,000 Gadbas speak the three closely related dialects of Ollari, Pottangi, and Poya; Pengu (1,900) and Manda (Manḍa) were discovered only recently, and Naiki of Chānda is also a newly investigated language (since 1961). The Khond tribes of Orissa (890,000) speak two closely related languages, Kui and Kuvi.

In Madhya Pradesh, many groups of Gonds (including about 2,460,000 persons) speak a number of Gondi dialects. To the north, in Bihār, Orissa, and Madhya Pradesh, the Oraon tribe speaks Kurukh (1,358,000), and near the borders of Bihār and West Bengal, 138,000 tribals speak Malto.

The only Dravidian language that is spoken entirely outside India is Brahui, with about 750,000 speakers who live in the Kalāt, Khairpur, and Hyderābād districts of Pakistan.

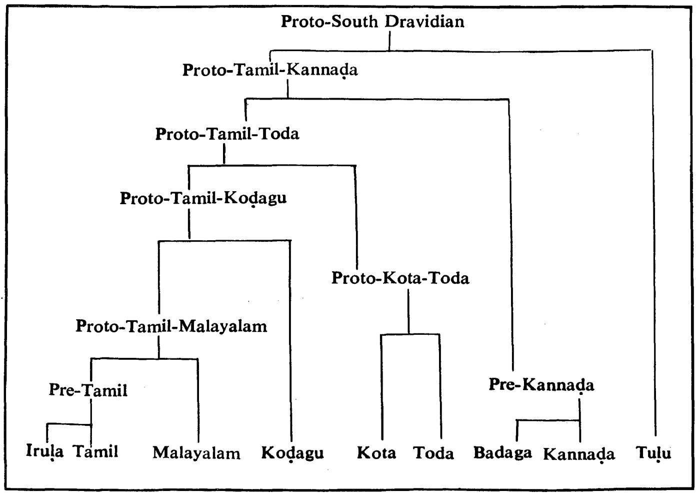


Figure 20: South Dravidian subfamily.

Historical survey of the Dravidian languages. Although in modern times speakers of the various Dravidian languages have mainly occupied the southern portion of India, while those of the Indo-Aryan (Indic) tongues have predominated in north India, nothing definite is known about the ancient domain of the Dravidian parent speech. It is, however, a well-established and well-supported hypothesis that Dravidian speakers must have been widespread throughout India, including the northwest region. This is clear because a number of features of the Dravidian languages appear in the Rigveda, the earliest known Indo-Aryan literary work, thus showing that the Dravidian languages must have been present in the area of the Indo-Aryan ones. The Indo-Aryan languages were not, however, originally native to India; they were introduced by Aryan invaders from the north. Several scholars have demonstrated that pre-Indo-Aryan and pre-Dravidian bilingualism in India provided conditions for the far-reaching influence of Dravidian on the Indo-Aryan tongues in the spheres of phonology (*e.g.*, the retroflex consonants, made with the tongue curled upward toward the palate), syntax (*e.g.*, the frequent use of gerunds, which are nonfinite verb forms of nominal character, as in "by the falling of the rain"), and vocabulary (a number of Dravidian loanwords apparently appearing in the Rigveda itself).

Dravidian features in the Rigveda

Thus a form of Proto-Dravidian, or perhaps Proto-North Dravidian, must have been extensive in north India before the advent of the Aryans. Apart from the survival of some islands of Dravidian speech, however, the process of replacement of the Dravidian languages by the Aryan tongues was entirely completed before the beginning of the Christian Era, after a period of bilingualism that must have lasted many centuries. Finally, the almost universal adoption of Indo-Aryan in the north and of Dravidian in the south has covered up the original linguistic diversity of India.

The circumstances of the advent of Dravidian speakers in India are shrouded in mystery. There are vague linguistic and cultural ties with the Urals, with the Mediterranean area, and with Iran. It is possible that a Dravidian-speaking people that can be described as dolichocephalic (longheaded from front to back) Mediterraneans mixed with brachycephalic (short-headed from front to back) Armenoids and established themselves in northwest India during the 4th millennium BC. Along their route, these immigrants may have possibly come into an intimate, prolonged contact with the Ural-Altaic speakers, thus explaining the striking affinities between the Dravidian and Ural-Altaic language groups. Between 2000 and 1500 BC, there was a fairly constant movement of Dravidian speakers from the northwest to the southeast of India, and in about 1500 BC three distinct dialect groups probably existed: Proto-North Dravidian, Proto-Central Dravidian, and Proto-South Dravidian. The beginnings of the splits in the parent speech, however, are obviously earlier. It is possible that Proto-Brahui was the first language to split off from Proto-Dravidian, probably during the immigration movement into India sometime in the 4th millennium BC, and that the next subgroup to split off was Proto-Kurukh-Malto, sometime in the 3rd millennium BC (see the family tree diagrams, Figures 20–22).

Possible Dravidian and Ural-Altaic contact

Compared to the work done on other language families, the progress in comparative Dravidian studies has been slow and firm results are still meagre. Considerable knowledge has been acquired in comparative phonology (sound systems), but correspondences have been worked out only for the sounds in the roots of words. Very little comparative work has been done on grammatical processes, and complete historical grammars of the literary languages are still lacking. Hence the reconstruction of any feature of the Dravidian protolanguage, with the possible exception of some parts of the phonology, must necessarily be considered very tentative.

The vowel system of Proto-Dravidian consisted of five vowels—**i, *u, *e, *o, *a* (an asterisk denotes an unattested, reconstructed, hypothetical form)—each having

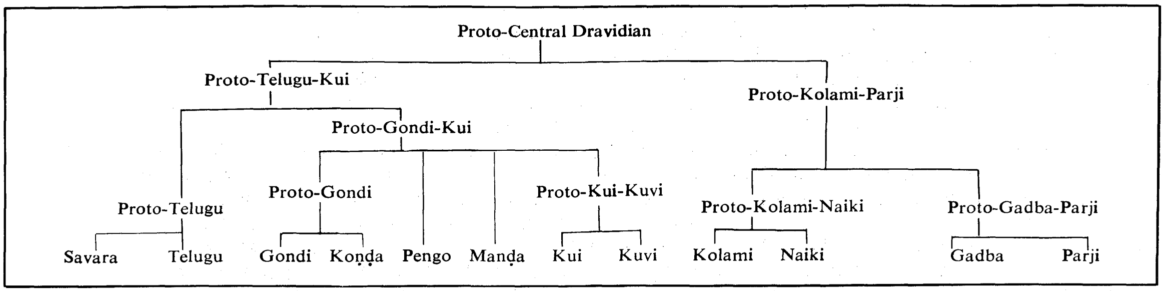


Figure 21: Central Dravidian subfamily.

two quantities, short and long. Relative stability of root vowels seems to have been the rule. The Proto-Dravidian consonant system consisted of obstruants (stops) *p, *t, *ṯ, *ṭ, *c, *k; nasals *m, *n, *ṇ, *ñ; laterals *l, *ḷ; the flap *r; the voiced retroflex continuant *ṛ; and the semivowels *y and *v. The most characteristic feature of the consonantal system was the six positions of articulation for obstruants: labial (with the lips), dental (tongue touching the back of the upper teeth), alveolar (tongue touching the upper gum ridge), retroflex (tip of tongue curled upward toward the palate and back), palatal (body of tongue touching the palate, or roof of the mouth), and velar (back of tongue touching the velum, or soft palate). The retroflex series was very distinctive and important and comprised an obstruant *ṭ, a nasal *ṇ, a lateral *ḷ, and a continuant *ṛ. No consonant of the alveolar or retroflex series began a word. In the final position all of the consonants occurred, but all of the obstruants were followed by an automatic release sound, the vowel *-u. Initial consonant clusters did not occur. There was only one series of obstruant phonemes (distinctive sounds); these sounds were voiceless (produced without vibration of the vocal cords) initially and voiced (with vocal cord vibration) between vowels. All Proto-Dravidian roots were monosyllables.

Proto-Dravidian used only suffixes, never prefixes or infixes, in the construction of inflected forms. Hence, the roots of words always occurred at the beginning. Nouns, verbs, and indeclinable words constituted the original word classes.

Process of Dravidian acculturation in India

During the 1st millennium BC, while Aryanization steadily progressed in north India, the Dravidian-speaking newcomers began to mix with the Negritos and Proto-Australoids in the south; this process of acculturation continued during the period from approximately 1200 to 600 BC. A movement of the Aryans into the south of India began sometime about 1000 BC. Before the 5th century BC, Proto-South Dravidian was probably still one language, but with two strongly marked dialects. Within Proto-Central Dravidian, a similarly deep two-way division also occurred, and as discussed above, North Dravidian must by that time have already been split into the Kurukh-Malto and Brahui subgroups (see the family tree diagrams, Figures 20–22).

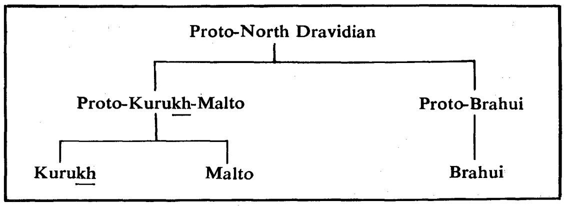


Figure 22: North Dravidian subfamily.

Apart from a possible Dravidian word in the Hebrew text of the Bible (*tukkhiyīm* "peacocks"; *cf.* Tamil *tōkai* "tail of a peacock"), the Dravidian languages enter history in Sanskrit and Greco-Roman texts. The *Cēras,* a south Indian dynasty, are possibly mentioned in the early Sanskrit text *Aitareya Āraṇyaka.* Kātyāyana, a grammarian of the 4th century BC, mentions the countries of Pāṇḍya (Tamil *pāṇṭiya*), Cōla (Tamil *cōla*), and Kerala, or Cēra (Tamil *cēra*); these lands were well known to Kauṭilya (4th century BC), the author of the earliest treatise on statecraft, and mentions of them also appear in the edicts of the great Buddhist leader Aśoka (3rd century BC). The term *drāviḍa* itself is almost certainly a Sanskritization (with an inserted "hypercorrect" *r*) of the earlier Pāli and Prākrit terms *dāmiḷo, damiḷa, dāviḍa,* which must have been derived from the Tamil name of the language, *tamil.* A number of South Dravidian words, almost all of them geographic and dynastic names, occur in such Greco-Roman sources as the *Periplus maris Erythraei* ("Circumnavigation of the Erythraean Sea") of about AD 89 and in the writing of Ptolemaeus of Naukratis of the 2nd century AD; it is also very probable that Western-language terms for rice (compare Italian *riso,* Latin *oryza,* Greek *oryza*) and ginger (compare Italian *zenzero,* German *Ingwer,* Greek *zingiberis*) are cultural loans from Old Tamil, in which they are *arici* and *iñcivēr,* respectively.

Sometime during the reign of Aśoka (3rd century BC), the two South Dravidian languages, Tamil and Kannada, developed into distinct idioms and the two cultures emerged as separate entities; a third major Dravidian linguistic and cultural unit, Telugu, appeared in the Andhra country. In the period from 300 to 100 BC, one of the pre-Tamil dialects (probably that of Madurai) gained prestige and became the standard literary language (*centamil*), the written form of early Old Tamil, which became established in poetic texts and in its earliest grammar, *Tolkāppiyam.* During the same period, about 250 BC, the Aśokan Southern Brāhmī script was adapted for Tamil and was used in short cave inscriptions by Jain monks over a period of several centuries, dating approximately from the 2nd century BC to the 5th century AD.

Early inscriptions and writings

The earliest inscriptions in Kannada may be dated at AD 450; Kannada literature begins with Nṛpatuṅga's *Kavirājamārga,* about AD 850. The oldest Telugu inscription is from AD 633, and the literature begins with the grammarian Nannaya's 11th-century translation of the Sanskrit classic the *Mahābhārata.* In Malayalam, the earliest writings are from the close of the 9th century, and the first literary text is probably the *Bhāṣākauṭalīyam,* AD 1125–1250.

Since these attested beginnings, the four languages—Tamil, Malayalam, Kannada, and Telugu—have been used continuously in administration and literature up to the present day. In addition to possessing an immense wealth of epigraphic and literary texts, they all developed pronounced features of diglossia, a dichotomy between the standardized, formal language and the informal, colloquial speech, which is divided into regional as well as social dialects. In modern times, all of the four cultivated languages have adapted quickly to new conditions resulting from economic, social, and political changes. All of these languages are used in teaching basic courses in science and the arts; and new technological terminology is coined, sometimes based either on English or Sanskrit models, but often on exclusively indigenous linguistic material (in Tamil).

To date, nothing is known about the history of the nonliterary Dravidian languages before their "discovery," which began at the end of the 18th century. The Gonds, however, are mentioned (as Gondaloi) by Ptolemy of Naukratis, writing in the 2nd century AD.

A tendency toward structural and systemic balance and stability is characteristic of the Dravidian group. Nevertheless, there is no doubt about the influence of the other languages of India. Dravidian languages show extensive lexical (vocabulary) borrowing, but only a few traits of structural (either phonological or grammatical) borrowing, from the Indo-Aryan tongues. On the other hand, Indo-Aryan shows rather large-scale structural borrowing from Dravidian, but relatively few loanwords. There is indeed a possibility of Dravidian and Indo-Aryan drawing even closer together in the future; but it is highly doubtful that a new family of languages will develop in such a way that the bases of the contributing groups (*i.e.*, Dravidian and Indo-Aryan) will be completely eliminated through the phenomena of borrowing.

Agglutinating character of Dravidian

Characteristics of the Dravidian languages. Dravidian languages would probably be called agglutinative in the categorization of the 19th-century philologists. An agglutinative language incorporates separate formal units of distinct meaning into a single word. There are some elements of "internal flexion" (*e.g.*, the alternation of short–long root vowels in derived words), however, as well as regular alternations in vowel and consonant quantities within the root. Relatively low receptivity to change results in a slower rate of change than is found in the Indo-European language family.

The degree of phonetic divergence among the Dravidian languages is not very great; hence, etymologies are not too difficult to discover. The territory occupied by Dravidian speakers in India may be characterized as a large dialect area resembling the area of the Romance languages, with numerous boundaries marked by bundles of isoglosses (an isogloss is a boundary line that separates the areas of two differing features of language usage), but also with many isoglosses enclosing more than one language. In any study of Dravidian, therefore, both evolution and diffusion must be taken into account.

Sounds of Dravidian. Compared to the reconstructed system of Proto-Dravidian phonemes (distinctive sounds), the most striking developments in vowels are the gradual elimination of the contrast between *e* and *ē* (long *e*) and *o* and *ō* (long *o*) in Brahui, as a result of the influence of Indo- Aryan languages or Iranian or both; the raising of Proto-Dravidian **e* and **o* to *i* and *u* and the lowering of these protolanguage sounds in Brahui; and the merger of Proto-Dravidian **i* and **u* with **e* and **o* in the South Dravidian languages before a consonant plus the vowel *a*. Also noteworthy are the emergence of retroflex vowels (*i.e.*, centralized vowels "coloured" by neighbouring retroflex consonants) in Kodagu and Irula; the nasalization of vowels, as in colloquial Tamil; the loss of vowels in unaccented noninitial syllables in Toda, Kota, some dialects of Kannada, and Tamil, and the resulting consonant clusters (*e.g.*, Kota *anjrčgčgvdk,* "because of the fact that [someone] will cause [someone] to terrify [someone]"). Metathesis (the transposition of sounds, as in "aks" from "ask") and vowel contraction resulted in initial consonant clusters in Telugu and other Central Dravidian languages—*e.g.*, Tamil *koḻu,* but Kui *krōga,* both meaning "fat."

Consonant changes

Among the most important consonantal developments are the loss of **c*-, a typical South Dravidian phenomenon that seems to be still in progress (*e.g.*, Proto-Dravidian **caṛ*-, but Tamil *alal* "to burn," and *talal* "to glow"); the velarization of **c*- to *k*- in North Dravidian when the sound is followed by ŭ̄ (*e.g.*, Tamil *cuṭu* "be hot," but Malto *kut*- "burn"); the palatalization of Proto-Dravidian **k*- to *c*- before front vowels in Tamil, Malayalam, and Telugu (*e.g.*, **ke*- "red," but Tamil *ce*-); and the replacement of **k*- in North Dravidian by *x* before ă̄, ŏ̄, and ŭ̄ (*e.g.*, Tamil *kal,* but Brahui *xal,* "stone"). The retroflex voiced continuant **ṛ* has been preserved only in the old stages of the cultivated languages and partly in modern Tamil and Malayalam; elsewhere, it merged with *ḷ, ḍ,* and other sounds. Some languages, notably Kannada, developed a secondary *h*-, not inherited from the parent speech (*e.g.*, Tamil *peyar,* Old Kannada *pesar,* but Modern Kannada *hersru,* "name"). According to the Dravidian scholar Bhadriraju Krishnamurti, a laryngeal (or *h*-type of sound) should be reconstructed for some items in Proto-Dravidian.

Problems of accent and intonation still remain to be worked out. Word stress is predictable, always occurring on the radical (initial) syllable and therefore being nondistinctive. The rules of sandhi (change of a sound or sounds as a result of adjacent sounds) are as complicated and delicate as in Sanskrit.

Grammatical features of Dravidian. In grammar, the absolutely prevailing process is suffixation, the addition of suffixes. Grammatical functions are, however, also expressed by composition (the compounding of word elements) and by word order. There are no prefixes or infixes. Suffixes agglutinate (are attached to one another); *e.g.*, Tamil *coṉṉatilēyiruntu* "from what was said" is composed of *col* "say" + *n* "past" + *atu* "3rd person singular neuter" + *il* "locative" + *ē* "emphatic" + *y* (an automatic insertion resulting from a sound rule) + *iruntu* "ablative" (*iruntu* comes from *iru* "be" + *nt/u* "past").

The major word classes are nouns (substantives, numerals, pronouns), adjectives, verbs, and indeclinables (particles, enclitics, adverbs, interjections, onomatopoetic words, echo words). There are two numbers and four different gender systems, the "original" probably having "male: non-male" in the singular and "person:non-person" in the plural. The pronoun has a category "inclusive:exclusive" in the 1st person plural. A characteristic derivation is that of "pronominalized" or "personal" nouns and adjectives; *e.g.*, Old Tamil *iḷai* "youth," *iḷai-y-am* "young-we," *iḷai-y-ar* "young-they."

Negative verbs

Finite forms of the verb (forms showing person and number) are, ultimately, "pronominalized" verb stems; *e.g.*, Tamil *aṭi*-(*y*)-*ēn* ("slave"—1st person singular) "I am a slave"; *nal-(l)-ēn* ("good"—1st person singular) "I am good"; *pō-v-ēn* ("go"—future—1st person singular) "I shall go." The most characteristic feature of the Dravidian verb is a full-fledged negative system: all of the positive verb forms have their corresponding negative counterparts. Verbs are intransitive, transitive, and causative; there are also active and passive forms. The main (and probably original) dichotomy in tense is past:non-past. Present tense developed later and independently in each language or subgroup.

In a sentence, however complex, only one finite verb occurs, normally at the end, preceded if necessary by a number of gerunds. Gerunds and participles, as well as verb-nouns, play an important role. The determining member always precedes the determined; *e.g.*, Tamil *poṉ* "gold" + *nakaram* "city" becomes *poṉṉakaram* "city of gold, golden city." Word order follows certain basic rules but is relatively free.

Vocabulary. In vocabulary, different Dravidian languages were receptive to loanwords in differing degrees. Among the cultivated languages, Tamil has the relatively lowest number of Indo-Aryan loanwords (18–25 percent, according to the style), whereas in Malayalam and Telugu the percentage of loanwords is substantially higher. The most important sources of loanwords have been Sanskrit, Pāli, and Prākrit (with varying degrees of importance in different periods); in modern times Urdu, Portuguese, and English have made significant contributions as well. There was only very limited lexical borrowing from one Dravidian language into another in historical times. Among all of the Dravidian languages, Brahui, in Pakistan, is inevitably the one most influenced by Indo-Aryan and Iranian; in contrast, Toda is probably the one language least influenced by any other idiom. In Tamil, there is currently a very notable and active purifying movement; it aims at removing as many borrowed "Sanskritic" (but not English) vocabulary items as possible. Such purism has not yet occurred in any other of the cultivated Dravidian languages.

Writing. Writing was first developed in Tamil Nadu, sometime about 250 BC, when the Aśokan Southern Brāhmī script was adapted for Tamil. The earliest inscriptions in Tamil script proper are the Pallava copperplates of about AD 550. The Kannada–Telugu script is based on Cālukya (6th century) inscriptions; the Grantha script, used in Tamil Nadu for Sanskrit since the 6th century, was accommodated for Malayalam and Tulu. Apart from these, Tamil has an old cursive script called Vaṭṭeḻuttu, "round script," and Malayalam possesses its own modern cursive form, Koleḻuttu, "rod-script."

(K.V.Z.)

AUSTRO-ASIATIC LANGUAGES

Austro-Asiatic languages are spoken by about 65,600,000 people scattered throughout Southeast Asia and eastern India. The family comprises about 150 languages, most of them having numerous dialects. Khmer, Mon, and Vietnamese are culturally the most important and have the longest recorded history. The rest are languages of nonurban minority groups written, if at all, only recently. The family is of great importance as a linguistic substratum for all Southeast Asian languages.

Main subfamilies

Superficially, there seems to be little in common between a monosyllabic tone language such as Vietnamese and a polysyllabic toneless Munda language such as Mundari of India; every recent study, however, confirms the underlying unity of the family. The date of separation of the three main Austro-Asiatic subfamilies—Munda, Nicobarese, and Mon-Khmer—has never been estimated and must be placed well back in prehistory. Within the Mon-Khmer subfamily itself, 12 main branches are distinguished; glottochronological estimates of the time during which specific languages have evolved separately from a common source indicate that these 12 branches all separated about 3,000 to 4,000 years ago.

Relationships with other language families have been proposed, but, because of the long durations involved and the scarcity of reliable data, it is very difficult to present a solid demonstration of their validity. In 1906 Wilhelm Schmidt, a German priest and anthropologist, classified Austro-Asiatic together with the Austronesian family (formerly called Malayo-Polynesian) to form a larger family called Austric. Paul K. Benedict, a U.S. scholar, extended the Austric theory, to include the Tai-Kadai family of Indochina and Burma and the Miao-Yao family of China, together forming an "Austro-Tai" superfamily.

Regarding subclassification within Austro-Asiatic, there have been several controversies. Schmidt, who first attempted a systematic comparison, included in Austro-Asiatic a "mixed group" of languages containing "Malay" borrowings and did not consider Vietnamese to be a member of the family. On the other hand, some of his critics contested the membership of the Munda group of eastern India. The "mixed group," recently called Chamic, is now considered to be Austronesian. It includes Cham, Jarai, Rhade, Chru, Roglai, and Hroy, and represents an ancient migration of Indonesian peoples into southern Indochina. As for Munda and Vietnamese, the recent works of the German linguist Heinz-Jürgen Pinnow on Kharia and of the French linguist André Haudricourt on Vietnamese tones have shown that both language groups are Austro-Asiatic.

Classification of the Austro-Asiatic languages. The work of classifying and comparing the Austro-Asiatic languages is still in the initial stages. In the past, classification has been done mainly according to geographic location. For instance, Khmer, Pear, and Stieng, all spoken on Cambodian territory, were all lumped together, although they actually belong to three different branches of the Mon-Khmer subfamily.

Khmer and Vietnamese: national languages

Numerically the most important, the Khmer and Vietnamese languages are also the only national languages of the Austro-Asiatic group and are regularly taught in schools and used in the mass media and on official occasions. Speakers of most other Austro-Asiatic languages are under strong social and political pressure to become bilingual in the official languages of the national unit in which they live. Most groups are too small or too scattered to win recognition, and, for many, the only chance of cultural survival lies in retreating to a mountain or jungle fastness, an old Austro-Asiatic tradition.

Phonological characteristics. In order to make general statements about the phonology (sound systems) of Austro-Asiatic languages, it is necessary to exclude Munda and Vietnamese, which, having been under the influence of Indian and Chinese languages, respectively, have acquired very divergent characteristics. The usual Austro-Asiatic word structure consists of a major syllable sometimes preceded by one or more minor syllables. A minor syllable has one consonant and one minor vowel. Most languages have only one type of minor vowel or, sometimes, three (*e.g.*, *a*, *i*, or *u*); others may also have vocalic nasal sounds, which are produced by releasing the breathstream through the nose, or liquids (*l* and *r* sounds) as minor vowels. Major syllables are composed of one or two consonants, followed by one major vowel and usually one final consonant.

Consonants. A typical consonant system for an Austro-Asiatic language would be the following (the symbols used are from the International Phonetic Alphabet):

p	t	c	k	ʔ	
b	d	j	g		
ɓ	ɗ				
m	n	ɲ	ŋ	ı	
w	r	l	s	y	h

Some languages (*e.g.*, Pearic, Semelaic) have an aspirated series of consonants, p^h, t^h, c^h, k^h, in which the sounds have an accompanying audible small puff of air, and many have no voiced stops at all (Monic, Khmer, Pearic). (Stops are consonants made with complete stoppage of the breath stream at some point in the vocal tract; voiced stops such as *b, d,* and *g* are produced with vocal cord vibration, as opposed to voiceless sounds produced without vocal cord vibration.) The imploded ɓ and ɗ are sounds pronounced by briefly drawing the air inward, causing suction, and are not truly voiced; they have sometimes been called pre-glottalized sounds or "semi-voiceless" sounds. These imploded stops are found only in a few branches of Mon-Khmer (*e.g.*, Mon, Khmer, Bahnaric), but it is possible that they existed in the ancestral language, called Proto-Mon-Khmer. Pre-glottalized nasals and liquids (*i.e.*, nasal and liquid sounds preceded by a glottal stop) are also found, sometimes as single distinctive sounds (unit phonemes) and sometimes as consonant clusters. In final position, all consonants except voiced stops can be found, but in several languages (*e.g.*, Mon, Sedang, Palaung) the number of possibilities is more reduced. Final stops are pronounced without release, nasals are often decomposed (*e.g.*, a final *m* becomes pronounced as b^m), and *s* sounds usually tend toward *h* sounds. Palatal consonants (*č* and *ñ*), produced with the blade of the tongue touching the hard palate, are commonly found at the end of words, a feature that sets Austro-Asiatic languages apart from the other languages of South Asia.

Vowels. Also characteristic of the Austro-Asiatic languages is an extraordinary variety of major vowels: systems of 30 to 35 different vowels are not uncommon (Bru has 41 distinctive vowel sounds [phonemes]). Four degrees of height are often distinguished in front and back vowels as well as in the central area. Diphthongs are not rare. Vowel length is usually distinctive: a normal vowel may contrast with an extra-short vowel of the same quality. Nasal vowels are found in several branches of the family, but, in any one language, they do not occur very frequently. A few dialects of Palaung, Wa, and North Bahnaric have a simple tone system, usually high versus low tone, but this is not typical of the Austro-Asiatic family. The Viet-Muong branch is the only one to have developed complex tone systems, probably influenced by non-Austro-Asiatic languages.

Registers, or voice qualities

Much more typical of the Austro-Asiatic family is a contrast between two series of vowels pronounced with different voice qualities, which are called registers. The voice may have a "creaky" register, a "breathy" register, or a normal one. Mon, Khmer, Jeh, Sedang, and some Palaung dialects have a two-way distinction of this sort. There is a controversy regarding the historical origin of the registers. Some believe that they were found in the original Proto-Mon-Khmer language; others, who seem to hold the more likely theory, propose that they are independent innovations in each branch, representing a transitional state from toneless to tonal languages.

Grammatical characteristics. *Morphology.* In morphology (word formation), Munda and Vietnamese again show the greatest deviations from the norm. Munda languages have an extremely complex system of prefixes, infixes (elements inserted within the body of a word), and suffixes. Verbs, for instance, are inflected for person, number, tense, negation, mood (intensive, durative, repetitive), definiteness, location, and agreement with the object. Furthermore, derivational processes indicate intransitive, causative, reciprocal, and reflexive forms. On the other hand, Vietnamese has practically no morphology.

Between these two extremes, the other Austro-Asiatic languages have many common features. (1) Except in Nicobarese, there are no suffixes. A few languages have enclitics, certain elements attached to the end of noun phrases (possessives in Semai, demonstratives in Mnong), but these do not constitute word suffixes. (2) Infixes and prefixes are common, so that only the final vowel and consonant of a word root remain untouched. It is rare to find more than one or two affixes (*i.e.,* prefixes or infixes) attached to one root; thus, because roots are mostly monosyllabic, the number of syllables per word remains very small. (3) The same prefix (or infix) may have a wide number of functions, depending on the noun or verb class to which it is added. For instance, the same nasal infix may turn verbs into nouns and mass nouns into count nouns. Sometimes, these different functions have similar meanings: for instance, reduplication, the repetition of a word or word element, may indicate plurality in nouns and repetition in verbs. This phenomenon may be widespread enough to make the distinction between basic word classes very unclear and questionable. (4) Many affixes are found only in a few fossilized forms and have often lost their meaning. (5) Expressive language and wordplay are embodied in a special word class called "expressives." These are sentence adverbials that describe noises, colours, light patterns, shapes, movements, sensations, emotions, aesthetic feelings, and so on. Some sort of symbolism, perhaps based on synaesthesia, is often observable in these words and serves as a guide for individual coinage of new words. The forms of the expressives are thus quite unstable, and the additional effect of wordplay can create subtle

Use of "expressives"

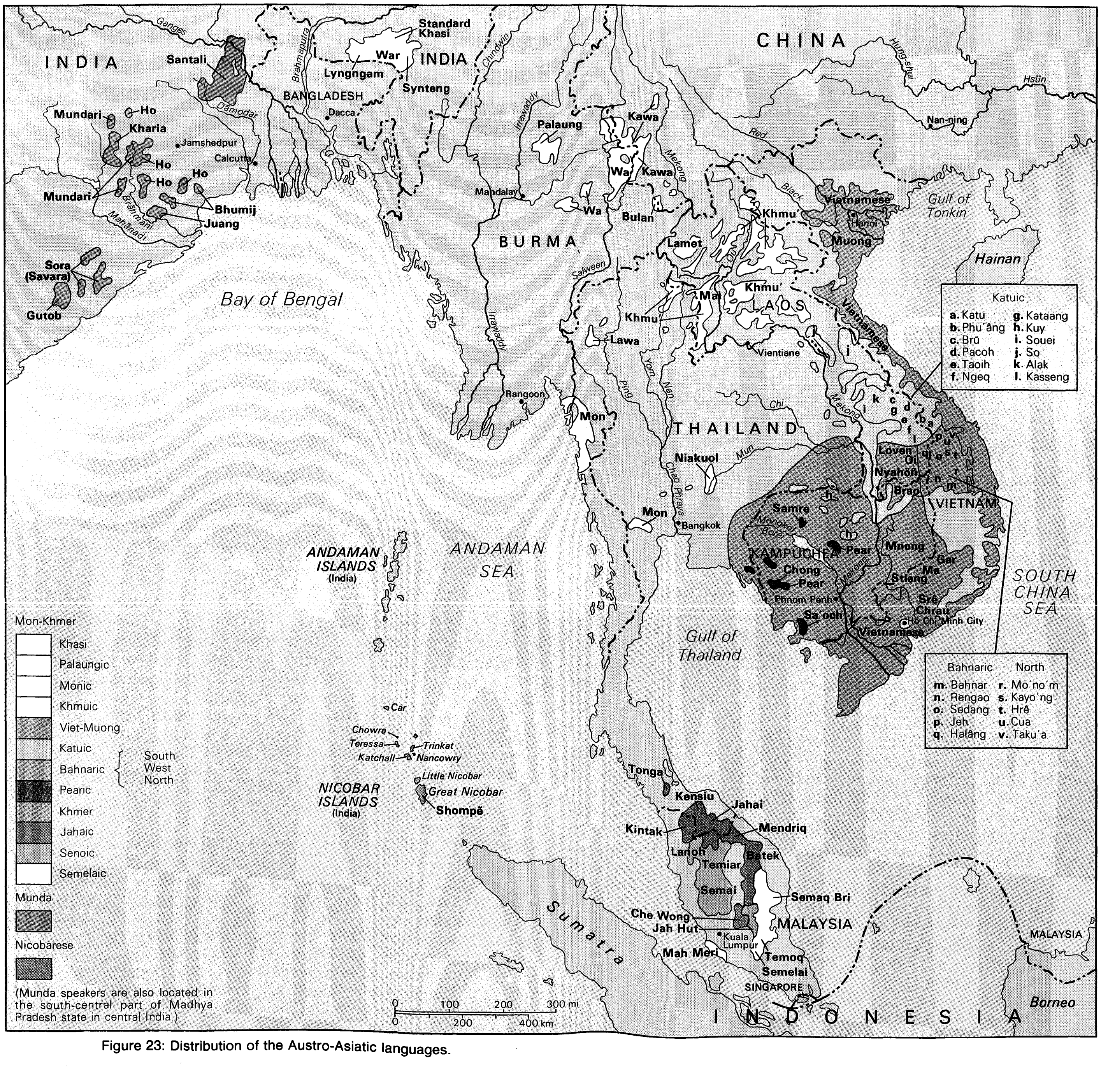


Figure 23: Distribution of the Austro-Asiatic languages.

and endless, sometimes apparently empty, structural variations. For example, in Bahnar one can say /pha:m lɛč həmrɔ:ŋ həmra:ŋ "blood flows həmrɔ:ŋ həmra:ŋ (like a torrent, irregularly)." (The slash marks indicate that the symbols enclosed are phonetic, standing for speech sounds rather than letters of the alphabet. Many Austro-Asiatic languages do not have their own writing systems and are thus recorded here in phonetic transcription.) In Semai /slɯ:č, səslɯ:č, səralɯ:č, səralĩp̄č, srlĩp̄č, shu:č, səsrhu:č, səralĩ:m/ and many other forms describe "a massage on oily skin, a snake's creeps, shiny fur, noodles in Chinese soup," and so on. The Semai forms /ɲəɲ pəlayə:ɲ, lɛɲ kəlayɯ:ɲ, pəɲ pəlayã:ɲ, puɲ pəlayõ:ɲ/ all describe "oversize hat, opening of parachute, flying disc, ridiculously large ears" and are based on wordplay with the borrowed Malay noun /payuŋ/, meaning "umbrella." Spontaneous expression rather than rational communication seems to be the dominant function of these expressives.

Syntax. In syntax, possessive and demonstrative forms and relative clauses follow the head noun; if particles are found, they will be prepositions, not postpositions (elements placed after the word to which they are primarily related), and the normal word order is subject–verb–object. There is usually no copula equivalent to the English verb "be." Thus, an equational sentence will consist of two nouns or noun phrases, separated by a pause; *e.g.*, in Rengao, /klan, bəs kən/, literally "python, snake large," means "a python (is) a large snake" or "phythons (are) large snakes." Predicates corresponding to the English "be + adjective" usually consist of a single intransitive (stative) verb; *e.g.*, in Khmer, /srəy nuh, lˀɔ:/, literally "girl that, pretty," means "that girl (is) pretty." Ergative constructions (in which the agent of the action is expressed not as the subject but as the instrumental complement of the verb) are quite common; *e.g.*, in Semai, the ergative sentence /tlɛy ˀadɛh ˀɲ-caaˀya ˀɛɲ/, literally, "banana this I-ate by me," means "I ate this banana" as does the active sentence: /ˀɛɲ ˀɲ-caaˀ tlɛy ˀadɛh/ "I I-ate banana this." Also noteworthy are sentence final particles that indicate the opinion, the expectations, the degree of respect or familiarity, and the intentions of the speaker. Munda syntax, here again, is radically different, having a basic subject–object–verb word order, like the Dravidian languages of India. It is quite conceivable that the complexity of Munda verb morphology is a result of the historical change from an older subject–verb–object to the present subject–object–verb basic structure.

Word taboos

Vocabulary. The composition of the vocabulary of the Austro-Asiatic languages reflects their history. Vietnamese, Mon, and Khmer, the best known languages of the family, came within the orbit of larger civilizations and borrowed without restraint—Vietnamese from Chinese, Mon and Khmer from Sanskrit and Pāli. At the same time, they have lost a large amount of their original Austro-Asiatic vocabulary. It is among isolated mountain and jungle groups that this vocabulary is best preserved. But there, other disruptive forces are at work. For instance, animal names are subject to numerous taboos; and the normal name is avoided in certain circumstances (*e.g.*, hunting, cooking, eating, and so on). A nickname is then invented, often by using a kinship term ("uncle," "Grandfather") followed by a pun or an expressive adverb describing the animal. In the course of time, the kinship term is abbreviated (thus many animal names begin with the same letter), the normal name is forgotten, and the nickname becomes standard. As such, it is then in turn avoided, and the process is repeated. There are also taboos on proper names; *e.g.*, after a person's death, his name and all words that resemble it are avoided and replaced by metaphors or circumlocutions. These replacements may explain why, for instance, the Nicobarese languages, which seem closely related, have few vocabulary items in common. In general, new words and fine shades of meanings can always be introduced by wordplay and from the open-ended set of expressive forms. Borrowings from the nearest majority languages are also common.

Writing systems and texts. Two Austro-Asiatic languages have developed their own orthographic systems and use them to this day. For both scripts, the letter shapes and principles of writing were borrowed from Indian alphabets (perhaps those of the Pallava kingdom in South India) that were in use in Southeast Asia at the time. Both Austro-Asiatic groups modified these alphabets in their own way, to suit the complex phonology of their languages. The most ancient inscriptions extant are in Old Mon (6th century AD), soon followed by Old Khmer in the early 7th century. The monuments of Burma, Thailand, and Kampuchea (Cambodia) have preserved a large number of official inscriptions in these two languages. Both alphabets were in turn used as models by other peoples for writing their own languages, the Thais using Khmer letters and the Burmese using Mon letters. The religious literature in Old and Middle Mon played a very important role in the spreading of Theravāda Buddhism to the rest of Southeast Asia.

Because Vietnam was a Chinese province for a thousand years, the Chinese language was used and written there for official purposes. In the course of time (perhaps as early as 8th century AD), a system called Chunom (popular writing) was developed for writing the Vietnamese language with partly modified Chinese characters. About 1650, a French missionary, Alexandre de Rhodes, devised a systematic spelling for Vietnamese, based on its distinctive sounds (phonemes). It uses the Latin (Roman) alphabet with some additional signs and several accents to mark tones. In this alphabet, tone was, for the first time in the world, recognized as a functional element and was systematically noted. At first, and for a long time, the use of this script was limited to Christian contexts, but it spread gradually, and in 1910 the French colonial administration made its use official. Now called Quoc-ngu (national language), it is learned and used by all Vietnamese.

Most other Austro-Asiatic languages have been written for less than a century; the literacy rate remains very low with a few exceptions (*e.g.*, Khasi). Dictionaries and grammars have been written only for the most prominent languages, with traditional and often insufficient methods. Many languages (*e.g.*, Wa, Kuy, Stieng, Pacoh, Katu, Muong) have only been described briefly in a few articles, and many more (Semelai, Puman, Sa'och, Riang, Lawa, Mrabri) are to Western scholars little more than names on the map. (G.Di.)

SINO-TIBETAN LANGUAGES

In the narrowest sense, the Sino-Tibetan languages include the Chinese and Tibeto-Burman languages. In terms of numbers of speakers, they comprise the world's second largest language family (after Indo-European), including over 300 languages and major dialects. In a wider sense, Sino-Tibetan has been defined as also including the Tai (or Daic), Karen, and Miao-Yao languages and even the Yenisey-Ostyak (or Ket) language in Northern Siberia (the latter affiliation seems rather untenable). Some linguists connect the Austro-Asiatic or Austronesian (Malayo-Polynesian) families, or both, with Sino-Tibetan; a suggested term for this most inclusive group, which seems to be based on premature speculations, is Sino-Austric. Other scholars see a relationship of Sino-Tibetan with the Athapascan and other languages of North America, but proof of this is beyond reach at the present state of knowledge.

Sino-Tibetan languages were known for a long time by the name of Indochinese, which is now restricted to the languages of Indochina. They were also called Tibeto-Chinese until the now universally accepted designation Sino-Tibetan was adopted. The term Sinitic also has been used in the same sense, as well as for the Chinese subfamily exclusively. (In the following discussion of language groups, the ending *-ic,* as in Sinitic, indicates a relatively large group of languages, and *-ish* denotes a smaller grouping.)

Areas in which the languages are spoken

Sinitic languages, commonly known as the Chinese dialects, are spoken in China and on Taiwan and by important minorities in all the countries of Southeast Asia (by a majority only in Singapore). In addition, Sinitic languages are spoken by Chinese immigrants in many parts of the world, notably in Oceania and in North and South America; altogether there are some 960,000,000 speakers of the Chinese dialects. Sinitic is divided into a number of language groups, by far the most important of which is Mandarin (or Northern Chinese). Mandarin, which includes Modern Standard Chinese (which is based on the Peking dialect), is not only the most important language of the Sino-Tibetan family but also has the most ancient writing tradition still in use of any modern language. The remaining Sinitic language groups include Wu (including Shanghai dialect), Hsiang (Hunanese), Kan (or Kan-Hakka), Yüeh, or Cantonese (including Canton and Hong Kong dialects), and Min (including Fuchow, Amoy, and Taiwanese).

Tibeto-Burman languages are spoken in Tibet and Burma; in the Himalayas, including Nepal, Sikkim, and Bhutan; in Assam and in Bangladesh; as well as by hill tribes all over mainland Southeast Asia and West China (the provinces of Kansu, Tsinghai, Szechwan, and Yunnan). The total number of speakers is approximately 50,000,000. Tibetic (*i.e.,* Tibetan in the widest sense of the word) comprises a number of dialects and languages spoken in Tibet and the Himalayas. Burmic (Burmese in its widest application) includes Lolo, Kachin, Kuki-Chin, the obsolete Hsi-hsia, or Tangut, and other languages. The Tibetan writing system (which dates from the 7th century) and the Burmese (dating from the 11th century) are derived from the Indo-Aryan (Indic) tradition; the Hsi-hsia system (developed in the 11th–13th century in Northwest China) was based on the Chinese model. Pictographic writing systems, which show some influence from Chinese, were developed within the last 500 years by Lolo and Moso tribes in West China. In modern times many Tibeto-Burman languages have acquired writing systems in Roman (Latin) script or in the script of the host country (Thai, Burmese, Indic, and others).

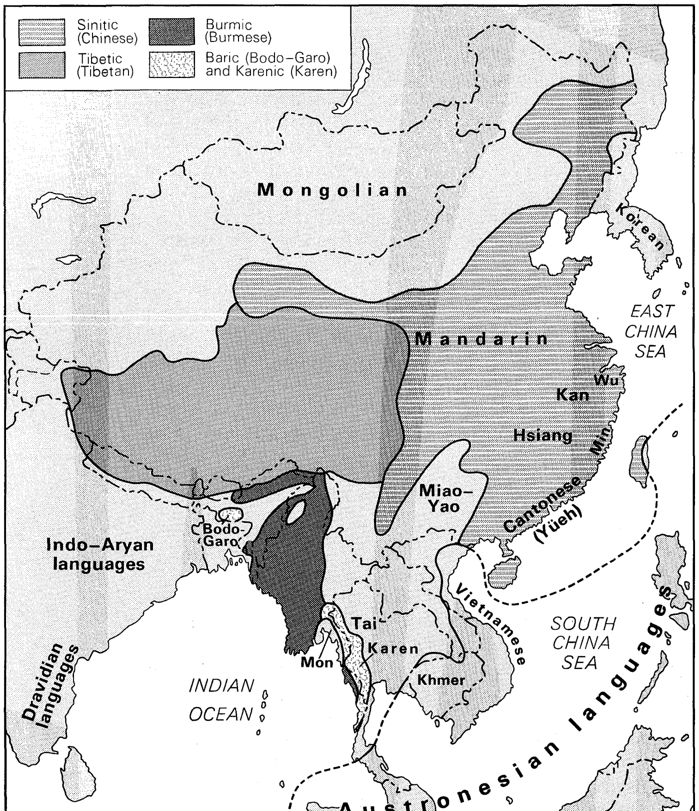


Figure 24: Distribution of the Sino-Tibetan languages.

Table 41: Tibetic (Bodic) Languages*

	areas where spoken	number of speakers†
Bodish-Himalayish		
Bodish languages		
Tibetan (with branches and dialects)	Tibet, Nepal, sporadically in India, in Bhutan, Kashmir, the Chinese provinces of Kansu, Tsinghai, Szechwan, and Yunnan	4,825,000
Central group: Lhasa, Khams, Kagate, Jad, Nyamkat (Mnyamskad)		
Southern group: Spiti, Sharpa, Sikkim, Lhoke		
Northern group: Ambo (Ngambo), Chone		
Western group: Balti, Purik (Burig), Ladākhi (Ladwags)	Kashmir	
Derge	China	
Gurung	central Nepal	
Gyarung (Rgyarung)	Tibet, Szechwan	
Himalayish languages		118,000
Kanauri branch: Thebor, Bunan, Kanashi, Chitkhuli, Manchati, Rangloi, Chamba Lahuli		
Almora branch: Rangkas, and others		
Kirantich (Bahing-Vayu)		440,000
Eastern (Bahing) branch: Bahing, Sunwar, Dumi; Khambu, Rodong, Waiing, Lambichong, Lohorung, Limbu, Yakha	eastern Nepal	
Western (Vayu) branch: Vayu, Chepang; Magari (perhaps)	central Nepal	
Mirish (Mishingish)	Assam, Tibet	
Miri (Mishing)		
Abor		
Dafla (Nyising)		
Other Tibetic languages		
Newari	central Nepal	550,000
Hruso (Hurso, Aka)	northern Assam	
Digaro (Taying)	Assam, Tibet	
Miju	Assam, Tibet	
Dhimal	Darjeeling area	

*Represents approximately 6,000,000 speakers. †Approximate.

CLASSIFICATION

The old literary languages, Chinese, Tibetan, and Burmese, are generally considered as representatives of three major divisions within Sino-Tibetan (Sinitic, Tibetic, and Burmic, respectively). A fourth literary language, Thai, or Siamese (written from the 13th century), represents what was accepted for a long time as a Tai or Daic division of Sino-Tibetan or as a division of a Sino-Tai family (see below *Tai languages*). This relationship is now more commonly considered nongenetic in that most of the shared vocabulary is more likely attributable to a history of cultural borrowing than to derivation from a common ancestral language.

Problems in classifying Karenic

Sinitic stands apart from Tibetic and Burmic on many grounds, including vocabulary, morphology, syntax, and phonology. Most scholars agree on combining Tibetic and Burmic into a Tibeto-Burman subfamily, which also includes Bodo-Garo or Baric but not Karenic. If Karenic is to be considered Sino-Tibetan, it must be set up as an independent member of a Tibeto-Karen group that includes Tibeto-Burman. The special affinities between Sinitic and Karenic (especially in syntax) are secondary. The two closely related language groups, Miao and Yao (also known as Mnong and Man), may be very remotely related to Sino-Tibetan; they are spoken in West China and northern mainland Southeast Asia and may well be of Austro-Tai stock.

In attempting to determine the exact interrelationship of the Tai (Daic) languages, Karenic, Sino-Tibetan, and several other marginal tongues, scholars must keep in mind that a discernible layer of Sino-Tibetan features in a given language may have been superimposed upon an older, non-Sino-Tibetan foundation (called the substratum language). Attributing a language to Sino-Tibetan or to another family may depend entirely on the ability of scholars to identify the substratum. Thus, if Tai is not considered as a division of Sino-Tibetan, it is because the substratum has been recognized as Austronesian; if Karen is still included among Sino-Tibetan languages on some level, it is perhaps because identification of a substratum is still lacking. Among the languages classified as Sino-Tibetan, a great many are known only from word lists or have not yet been described in a way that makes valid comparisons possible.

A number of Sino-Tibetan languages are enumerated below together with their most likely affiliation. Some scholars believe the Tibetic and Burmic divisions to be premature and that for the present their subdivisions (such as Bodish, Himalayish, Kirantish, Burmish, Kachinish, Kukish) should be considered as the classificatory peaks around which the Sino-Tibetan languages group themselves as members or more or less distant relatives. Certainly the stage has not yet been reached in which definite boundaries can be laid down and ancestral Proto-, or Common, Tibetic and Proto-, or Common, Burmic can be undisputedly reconstructed.

Tibetic languages. The Tibetic (also called the Bodic, from Bod, the Tibetan name for Tibet) division comprises the Bodish-Himalayish, Kirantish, and Mirish language groups.

Burmic languages. The Burmic division comprises Burmish, Kachinish, and Kukish.

A number of Tibeto-Burman languages that are difficult to classify have marginal affiliations with Burmic. TheLuish languages (Andro, Sengmai, Kadu, Sak, and perhaps also Chairel) in Manipur and adjacent Burma resemble Kachin; Nung (including Rawang and Trung) in the Kachin State of Burma and in Yunnan has similarities with Kachin; and Mru, with 18,000 speakers in Arakan and Chittagong Hills, Meithei (Mhithlei) in Manipur, and Mikir in Assam seem close to Kukish.

Table 42: Burmic Languages

	areas where spoken	number of speakers
Burmish (Burmese-Lolo)		
Burma branch	west China, Burma, Indochina	
Burmese and its dialects (Rangoon, Mergui, Intha, Danu, Yaw, Taungyo, Tavoyan, Arakanese) Maru (Lawng) Lashi (Letsi) Atsi (Tsaiwa)		22,170,000
Lolo branch		
Northern group: Lisu, Nyi (I), Ahi, Lolopho Southern group: Akha, Lahu Nakhi (Moso)* Ch'iang*		
Kachinish		
Kachin (Chingpaw)†	Kachin State in Burma, adjoining Assam, Shan State, Yunnan	600,000
Kukish (Kuki-Chin)		
Kuki branch	border region of Burma–South Assam–Bangladesh	447,000
Central Kuki: Lushai, Lakher, Lai (Haka) Northern Kuki: Thado, Kamhau (Tiddim), Siyin (Sizang) Southern Kuki: Sho, Yawdwin, Chinbok, Khami (with Khimi)		
Old Kuki: Rangkhol, Bete (Biate); Anai, Lamgang; Aimol, Purum	spoken by dispersed tribes driven out from their original home on the Burma–India border	
Western Kuki: Empeo, Maram, Kwoireng, Kabui, Khoirao‡ Lahupa languages§ (Tangkhul, Maring, Khoibu)		
Nāgā branch	Nāgāland, Burma	776,000
Northern: Ao Eastern group: Rengma, Sema (Simi), Angami		
Lepchā (Rong) ‖	Sikkim, east Nepal, west Bhutan	38,000

*These stand in a not clearly defined relationship to Lolo. †For cultural, non-linguistic reasons, Maru, Lashi, and Atsi are often grouped with Kachin. ‡In Manipur and Cāchār. §These languages are similar to Kuki. ‖ Usually compared to the northern Nāgā branch of Kukish, but has Baric and Himalayish affinities.

Table 43: Baric (Bodo-Garo) Languages

	areas where spoken	number of speakers*
Bodo branch	the plains of Assam	
Bodo		712,000
Dimasa		44,000
Garo branch	the hills of Meghalaya	549,000+
Achik, Abeng, Dacca Atong, Rabha, Ruga, Koch		

*Approximate.

Baric languages. The Baric, or Bodo-Garo, division consists of a number of languages spoken in Assam and falls into a Bodo branch (not to be confused with Bodic-Tibetic, and Bodish, a subdivision of Tibetic) and a Garo branch.

A group of Sino-Tibetan languages in Nāgāland (Nagish, not to be confused with the Nāga branch of Kukish; including Mo Shang, Namsang, and Banpara) has affinities to Baric.

Karenic languages. The Karenic languages of the Karen State of Burma and adjacent areas in Burma and Thailand include the two major languages of the Pho (Pwo) and Sgaw, which have about 2,530,000 speakers. Taungthu (Pa-o) is close to Pho, and Palaychi to Sgaw. There are several minor groups.

Chinese, or Sinitic, languages. Chinese as the name of a language is a misnomer. It has been applied to numerous dialects, styles, and languages from the middle of the 2nd millennium BC. Sinitic is a more satisfactory designation for covering all these entities and setting them off from the Tibeto-Karen group of Sino-Tibetan languages. *Han* is a Chinese term for Chinese as opposed to non-Chinese languages spoken in China. The Chinese terms for Modern Standard Chinese are *kuo-yü* "national language" and *p'u-t'ung-hua* "common language." Reconstructed prehistoric Chinese is known as Proto-Sinitic (or Proto-Chinese); the oldest historic language of China is called Archaic or Old Chinese (8th–3rd century BC), and that of the next period up to and including the T'ang dynasty (618–907) is known as Ancient or Middle Chinese. Languages of later periods include Old, Middle, and Modern Mandarin (the name Mandarin is a translation of *kuan-hua*, "civil servant language"). Through history the Sinitic language area has constantly expanded from the "Middle Kingdom" around the eastern Huang Ho to its present size. The persistence of a common, non-phonetic writing system for centuries explains why the word dialect rather than language has had widespread usage for referring to the modern speech forms. The present-day spoken languages are not mutually intelligible (some are further apart than Portuguese and Italian), and neither are the major subdivisions within each group. The variation is slightest in the western and southwestern provinces and the greatest along the Huang Ho and in the coastal areas. Table 44 gives the percentage of Chinese people speaking each of the various Chinese languages.

Absence of intelligibility between Chinese spoken languages

A vernacular written tradition exists mainly in Peking Mandarin and in Cantonese. An unwritten storytelling tradition has survived in most languages. The school and radio language is Modern Standard Chinese in the People's Republic of China as well as in Taiwan and Singapore. In Hong Kong, Cantonese prevails as the language of education and in the communication media. The same orthographic system is employed, with a few variations, by all speakers of Chinese.

Non-Chinese Sino-Tibetan languages of China include some Lolo-type languages (Burmish)—I (Yi, or Nyi), with over 4,800,000 speakers in Yunnan, Szechwan, Kweichow, and Kwangsi; Hani (or Akha) with about 960,000 speakers in Yunnan; Lisu, with approximately 470,000 speakers in Yunnan; Lahu, with about 270,000 speakers in Yunnan; Moso or Nasi, with 230,000 speakers in Yunnan and Szechwan; and Achang by the Burmese border. Other Sino-Tibetan languages in Yunnan and Szechwan

Table 44: Distribution of the Chinese Languages

	areas where spoken	ratio of speakers to total population (percentage)*
Mandarin	China, north of the Yangtze River; a narrow belt south of the Yangtze River in Kiangsi, Anhwei, and Kiangsu; Szechwan, Yunnan, and Kweichow; small parts of Hunan	70†
Wu	Kiangsu and Chekiang	8+
Hsiang	Hunan	5
Kan‡	Kiangsi and a corner of Hupei	2+
Hakka‡	east and north Kwangtung; parts of Fukien, Kiangsi, Kwangsi, Taiwan, Hunan, Szechwan; sporadically in other provinces	4
Yüeh (Cantonese)	Kwangtung, south Kwangsi, Hong Kong, Macau	5
Min§	Fukien, west Kwangtung, Hainan, Taiwan, south Chekiang	4+

*An estimate from the early 1980s gives the following figures of number of speakers: Mandarin 612,750,000; Wu 83,500,000; Min 77,000,000; Kan-Hakka 66,760,000; Yüeh 53,500,000; Hsiang 49,000,000. †Figure applies to those people speaking Mandarin as the mother tongue as opposed to speaking it as a second language. ‡Kan and Hakka are often combined as Kan-Hakka, or simply Kan. §Min is subdivided into Northern Min, in North Fukien with little over 1 percent of the national population, and Southern Min, in the remainder of the Min-speaking area with 3 percent of the population.

(with fewer than 100,000 speakers each) are Kachin and the closely related Tsaiwa; Ch'iang, Gyarung, Hsifan; and Pai (or Minchia, perhaps an Austro-Asiatic language).

COMMON FEATURES

At the end of the 18th and during the first half of the 19th century, a great number of languages were discovered by Western scholars in the Himalayas, in India, and in China, and word lists and grammatical sketches began to appear. By the late 19th century a foundation had been laid for Sino-Tibetan comparative studies.

The comparative method for determining relationships

The comparative method for determining genetic relationship among languages was worked out in detail for Indo-European during the latter part of the 19th century. It rests on the assumption that sound correspondences in related words and morphological units, as well as structural similarities on all levels (phonology, morphology, syntax), can be explained in terms of a reconstructed common language, or protolanguage. The morphology and syntax of the Sino-Tibetan languages are for the most part rather simple and nonspecific, and the length of time involved in the separation of subfamilies and divisions is such that comparative phonological statements are often difficult to reduce to the form of concise correspondences and laws.

A number of features have been delineated as common for the Sino-Tibetan languages. Many of them can be shown to be of a typological nature, the result of diffusion and underlying unrelated language strata.

Typological similarities. *Monosyllabicity.* The vast majority of all words in all Sino-Tibetan languages are of one syllable, and the exceptions appear to be secondary (*i.e.,* words that were introduced at a later date than Common, or Proto-, Sino-Tibetan). Some suffixes in Tibeto-Burman are syllabic, thus adding a syllable to a word, but they have a highly reduced set of vowels and tones ("minor syllables"). These features are, however, shared by contiguous languages not clearly attributable to Sino-Tibetan on the basis of shared basic vocabulary items (namely, Austro-Asiatic and Miao-Yao).

Tonality. Most Sino-Tibetan languages possess phonemic tones, which indicate a difference in meaning in otherwise similar words. There are no tones in Purik, a Western Tibetan language; Ambo, a Northern Tibetan tongue; and Newari of Nepal. Balti, another Western Tibetan language, has pitch differences in polysyllabic nouns. The tones of the remaining Tibetan dialects can be accounted for by positing an original and older system of voiced and voiceless initial sounds that eventually resulted in tones. In several Himalayish languages, tones are linked with articulatory features connected with the end of the syllable or are linked with stress features, as also in Kukish Lepcha.

Most Baric languages lack tones altogether; and Burmic, Karenic, and Sinitic tonal systems can be reduced to two basic tones ultimately probably accounted for by different syllabic endings. What can be reconstructed for Proto-Sino-Tibetan, the language from which all the modern Sino-Tibetan languages developed, are a set of conditioning factors (as, for example, certain syllabic endings) that resulted in tones; the tones themselves cannot be reconstructed. Again the features that encouraged the development of tones are not uniquely Sino-Tibetan; similar conditions have produced similar effects in Tai and Miao-Yao and—within the Austro-Asiatic languages—in Vietnamese and in the embryonic form of two registers (pitches or musical range) also in Cambodian.

Prefixes and suffixes

Affixation. Most Sino-Tibetan languages possess or can be shown to have at one time possessed derivational and morphological affixes—*i.e.,* word elements attached before or after or within the main stem of a word that change or modify the meaning in some way. Many prefixes can be reconstructed for Proto-Sino-Tibetan: *s-* (causative), *m-* (intransitive), *b-*, *d-*, *g-*, and *r-*, and many more for certain language divisions and units. Among the suffixes, *-s* (used with several types of verbs and nouns), *-t,* and *-n* are inherited from the protolanguage. The problem of whether Proto-Sino-Tibetan made use of *-r-* and *-l-* infixes (besides perhaps semivocalic infixes) has not been solved. Whether clusters containing these sounds were the result of prefixation to roots beginning in *r* and *l* (and *y*) or came about through infixation is not clear.

Initial consonant alternation. Voiced and voiceless initial stops alternate in the same root in many Sino-Tibetan languages, including Chinese, Burmese, and Tibetan (voiced in intransitive, voiceless in transitive verbs). The German Oriental scholar August Conrady linked this morphological system to the causative *s-* prefix, which was supposed to have caused devoicing of voiced stops. (Voicing is the vibration of the vocal cords, as occurs, for example, in the sounds *b, d, g, z,* and so on. Devoicing, or voicelessness, is the pronunciation of sounds without vibration of the vocal cords, as in *p, t, k, s*). Such alternating of the initial consonant cannot itself be reconstructed for the protolanguage.

Vowel alternation. The morphological use of vowel gradation (called ablaut) is well-known from Indo-European languages (*e.g.,* the vowel change in English "sing, sang, sung") and is found in several Sino-Tibetan languages, including Chinese and Tibetan. In Tibetan the various forms of verbs are differentiated in part by vowel alternation; in Sinitic some related words (known as word families) are kept apart by vowel alternation. A conditioning factor outside the vowel (perhaps stress or sandhi, the modification of a sound according to the surrounding sounds) may have been responsible for the Sino-Tibetan ablaut systems.

Indistinct word classes. Especially in the older stages of Sino-Tibetan, the distinction of verbs and nouns appears blurred; both overlap extensively in the Old Chinese writing system. Philological tradition as well as Sinitic reconstruction show, however, that frequently when the verb and the noun were written alike, they were pronounced differently, the difference manifesting itself later in the tonal system. Verbs and nouns also used different sets of particles. In this respect, Chinese resembles Tai and Austro-Asiatic, whereas Tibetan is more similar to the Altaic languages (for example, Turkish, Mongol).

Classifiers for nouns

Use of noun classifiers. The Sino-Tibetan noun is typically a collective term, designating all members of its class—*e.g.,* "man" meaning "all human beings." In a number of modern Sino-Tibetan languages, such a noun can be counted or modified by a demonstrative pronoun only indirectly through a smaller number of non-collective nouns, called "classifiers," in constructions such as "one *person* man," "one *animal* dog," and so on, much like parallel cases in Indo-European (in English, "one head of cattle"; in German, *ein Kopf Salat* "one head of lettuce"). The phenomenon is absent in Tibetan and appears late in Burmese and Chinese. Furthermore, classifiers are not exclusively Sino-Tibetan; they exist also in Miao-Yao, Tai, Austric, and Japanese. In Classical Chinese, Tai, and Burmese, the classifier construction follows the noun; in modern Chinese, as in Miao-Yao, it precedes it. Classifiers are of later origin and do not belong to Proto-Sino-Tibetan.

Word order. Although the word order of subject–object–verb and modified–modifier prevails in Tibeto-Burman, the order subject–verb–object and modifier–modified occurs in Karenic. In this respect Chinese is like Karen, although Old Chinese shows remnants of the Tibeto-Burman word order. Tai employs still another order: subject–verb–object, and modified–modifier, like Austric but unlike Miao-Yao, which follows the Karen and Chinese model. Word order, even more than any of the other distinguishing features, points to diffusion from several centres, or to unrelated substrata.

Phonological correspondences. The hypothesis that the Sino-Tibetan languages are all related and derive from a common source depends on phonological correspondences in shared vocabulary more than on any other argument. It is ironic that the clearest and most convincing results should have been obtained from studies of the Sinitic-Tai similarities, which probably do not indicate a true case of genetic relationship. In 1942 it was shown that most of the words in this grouping were cultural loans (then thought of as Chinese loanwords in Tai, now believed to a very large extent to be borrowings in the opposite direction).

A comparison of Old Chinese and Old Tibetan made by Walter Simon in 1929, although limited in some ways, pointed to enough sound resemblances in important items of basic vocabulary to eliminate the possibility of coincidental similarities between unrelated languages. A few examples of similar words in Old Tibetan and Old Chinese, respectively, follow: "bent," *khyog* and *khyuk;* "eye," *mig* and *myok;* "friend," *grogs* and *gyug;* "kill," *gsod* and *sat;* "die" *shi* and *syər;* "sun," *nyi* and *ńye;* "life," *srog* and *serŋ*. The U.S. linguist Paul Benedict brought in material from other Sino-Tibetan languages and laid down the rule that the comparative linguist should accept perfect phonetic correspondences with inexact though close semantic equivalences in preference to perfect semantic equivalences with questionable phonetic correspondences. New material and competent descriptions later made it possible to reconstruct important features of common ancestral languages within major divisions of Sino-Tibetan (notably Lolo, Baric, Tibetic, Kachin, Kukish, Karenic, Sinitic).

Chain of interrelated languages

Interrelationship of the language groups. The position of Proto-Sino-Tibetan can be defined in terms of a chain of interrelated languages and language groups: Sinitic is connected with Tibetic through a body of shared vocabulary and typological features, similarly Tibetic with Baric, Baric with Burmic, and Burmic with Karenic. The chain continues at both ends, connecting Sinitic to Tai and Tai to Austronesian (Malayo-Polynesian) and also connecting Karenic with Austro-Asiatic. Considerations of basic vocabulary versus cultural loans and diffusion versus inheritance have led scholars to believe that only the members of the chain from Sinitic to Karenic share a common ancestral language; especially Sinitic and Karenic are under suspicion for containing only superstrata of Sino-Tibetan origin.

Proto-Tibeto-Burman was monosyllabic. Some grammatical units may have had the form of minor syllables before the major syllable (**ma-*, **ba-*) or after the major syllable (*-*ma*, *-*ba*). (An asterisk [*] indicates that the form it precedes is unattested and has been reconstructed as a possible ancestral form.) The consonants were three voiceless stops (*p, t, k*), which were aspirated in absolute initial position, three voiced stops (*b, d, g*), and three nasal sounds (*m, n, ŋ* [as the *-ng* in "sing"]). There were five continuant sounds (*s, z, r, l,* and *h*) and two semivowels (*w, y*). In final position there was only one set of stops, but there were a number of initial and final clusters mainly resulting from the addition of prefixes and suffixes. Three degrees of vowel opening existed with two members in each: *i* and *u, e* and *o, a* and *aa* (short and long *a*). Length may have been relevant also with the *i* and *u* and *e* and *o* vowels. The conditioning factors that led to the development of tones can be shown to have been voiced–voiceless contrast in initial and final consonants and consonant clusters. Because the conditioning factors were involved with morphological process (affixation and consonant alternation), tonal systems could also acquire certain grammatical or structural functions. An independent morphological system was vowel alternation.

The sound system of Proto-Karenic appears closely related to that of Proto-Tibeto-Burman. The tonal classes can be reduced to two, which connect Karen to Burmic, Sinitic, Tai, and Miao-Yao.

Greater dissimilarity is encountered with respect to Proto-Sinitic. The contrast of aspirated and unaspirated voiceless stops in initial position is most likely the result of lost initial cluster elements as in Proto-Tibeto-Burman. The voiced stops also possess the aspirated–unaspirated distinction. Unlike Tibeto-Burman, two series of stops in syllable final position are posited for Old Chinese, but whether the contrast involved aspiration or voicing is not clear. One series is in general without an exact correspondence in Tibeto-Burman languages, but Burmish Maru and Kuki-oid Mru both have final stops in a number of these words. Similar isolated cases are found in Tibetan and in Tai.

Old Chinese has two more relevant points of articulation, or sound-producing positions of the mouth, than Proto-Tibeto-Burman: palatal (in which the tongue surface touches the palate) and retroflex (in which the tip of the tongue is curled upward toward the palate). But these two types of sounds may be explained as the result of influence from lost Proto-Sinitic medical sounds (a palatal *-y-* and a retroflex *-r-*). The relationship between these specific medial sounds and similar elements in Tibeto-Karen is, however, not certain. Dental affricate sounds in Old Chinese, which begin as stops with complete stoppage of the breath stream and conclude as fricatives with incomplete air stoppage and audible friction, can at least be explained partly as metathesized (transposed) forms of prefix *s-* plus a dental sound in Proto-Sinitic (*e.g., st* changes to *ts*). Old Chinese possessed initial consonant clusters containing *-l-* as a second element, so Proto-Sinitic can reasonably be supposed to have had the same three medial elements as Proto-Tibeto-Burman: *-y-*, *-l-*, and *-r-*. There are few, if any, traces in Old Chinese of the more complicated clusters and the minor syllables of Tibeto-Burman.

Vowel system of Old Chinese

The vowel system of Old Chinese as reconstructed (1940) by the linguist Bernhard Karlgren to account especially for the language of the "Classic of Poetry" (a collection of poetry, entitled the *Shih Ching,* from around 800–600 BC) seems surprisingly complicated as compared to that of later Sinitic and to that of Proto-Tibeto-Burman. Probably some of the vowels were in reality diphthongs or combinations of vowels plus consonants.

As in Karen and Burmese-Loloish, the tones of Sinitic can be reduced to two (the modern Sinitic languages have from two to as many as eight or nine). Monosyllabicity of roots and morphological affixation were characteristic features of Proto-Sino-Tibetan as of Proto-Tibeto-Karen.

CHARACTERISTICS OF THE MODERN CHINESE LANGUAGES

All modern Sinitic languages—*i.e.,* the "Chinese dialects"—share a number of important typological features. They have a single syllabic structure of the type consonant–semivowel–vowel–semivowel–consonant. Some languages lack one set of semivowels, and, in some, gemination (doubling) or clustering of vowels occurs. The languages also employ a system of tones (pitch and contour) and sometimes glottal features and occasionally stress. For the most part, tones are lexical (*i.e.,* they distinguish otherwise similar words); in some languages tones also carry grammatical meaning. Non-tonal grammatical units (*i.e.,* affixes) may be smaller than syllables, but usually meaningful units consist of one or more syllables. Words can consist of one syllable, of two or more syllables each carrying an element of meaning, or of two or more syllables that individually carry no meaning. For example, Modern Standard Chinese *t'en* "sky, heaven, day" is a one-syllable word; *jih-t'ou* "sun" is composed of *jih* "sun, day," a word element that cannot occur alone as a word, and the noun suffix *t'ou;* and *hu-t'ier* "butterfly" consists of two syllables, each having no meaning in itself (this is a rare type of word formation). The Southern languages have more monosyllabic words and word elements than the Northern ones.

Sinitic grammatical characteristics

The Sinitic languages distinguish nouns and verbs with some overlapping, as do Sino-Tibetan languages in general. There are noun suffixes that form different kinds of nouns (concrete nouns, diminutives, abstract nouns, and so on), particles placed after nouns indicating relationships in time and space, and verb particles for modes and aspects. Adjectives act as one of several kinds of verbs. Verbs can occur in a series (concatenation) with irreversible order (*e.g.*, the verbs "take" and "come" placed next to one another denote the concept "bring"). Nouns are collective in nature, and only classifiers (see above) can be counted and referred to singly. Specific particles are used to indicate the relationship of nominals (*e.g.*, nouns and noun phrases) to verbs, such as transitive verb–object, agent–passive verb; in some of the languages this system forms a sentence construction called ergative, in which all nominals are marked for their function and the verb stays unchanged. Final sentence particles convey a variety of meanings, such as "question, command, surprise, new situation." The general word order of subject–verb–object and complement and modifier–modified is the same in all the languages, but the use of the preposed particles and verbs in a series varies considerably. Grammatical elements of equal or closely related values in various languages are very often not related in sounds.

The Sinitic languages fall into a Northern and a Southern group. The Northern languages (or Mandarin dialects) are closer to each other than the Southern ones (Wu, Hisiang, Kan, Yüeh, Min).

Modern Standard Chinese (Mandarin). Modern Standard Chinese is based on the Peking dialect, which is of the Northern, or Mandarin, type. It employs about 1,300 different syllables. There are 22 initial consonants, including stops (made with momentary, complete closure in the vocal tract), affricates (beginning as stops but ending with incomplete closure), aspirated consonants, nasals, fricatives, liquid sounds (*l, r*), and a glottal stop. The medial semivowels are *y* (*i*), ɥ (*ü*) and *w* (*u*). In final position, the following occur: nasal consonants, *ṛ* (retroflex r), the semivowels *y* and *w*, and the combinations ŋ*r* and *wr*. There are nine vowel sounds, including three varieties of *i* (retroflex, apical, and palatal). Several vowels combine into clusters.

The four tones of Mandarin

There are four tones: (1) high level; (2) high rising crescendo; (3) low falling diminuendo with glottal friction (with an extra rise from low to high when final); (4) falling diminuendo. Unstressed syllables have a neutral tone, which depends on its surroundings for pitch. Tones in sequences of syllables that belong together lexically and syntactically ("sandhi groups") may undergo changes known as tonal sandhi, the most important of which causes a third tone before another third tone to be pronounced as a second tone. The tones influence some vowels (notably *e* and *o*), which are pronounced more open in third and fourth tones than in first and second tones.

A surprisingly low number of the possible combinations of all the consonantal, vocalic, and tonal sounds are utilized. The vowels *i* and *ü* and the semivowels *y* and ɥ never occur after velar sounds (*e.g., k*) and occur only after the palatalized affricate and sibilant sounds (*e.g., tŝ*), which in turn occur with no other vowels and semivowels.

There are many alternative interpretations of the distinctive sounds of Chinese; the interaction of consonants, vowels, semivowels, and tones sets Modern Standard Chinese apart from many other Sinitic languages and dialects and gives it a unique character among the major languages of the world. The two most widely used transcription systems (romanizations) are Wade-Giles (first propounded by Sir Thomas Francis Wade in 1859 and later modified by Herbert A. Giles) and the official Chinese transcription system today, known as the *pinyin zimu* ("Pinyin phonetic spelling") or simply Pinyin (adopted in 1958). Table 45 presents the romanizations of both the Wade-Giles and the Pinyin systems. In Wade-Giles, aspiration is marked by ' (*p' t'*, and so on). The semivowels are *y, yü,* and *w* in initial position; *i, ü,* and *u* in medial; and *i* and *u* (but *o* after *a*) in final position. Final retroflex *r* is written *rh.* The tones are indicated by raised figures after the syllables (^{1}, ^{2}, ^{3}, ^{4}).

The Pinyin system indicates unaspirated stops and affricates by means of traditionally voiced consonants (*e.g., b, d*) and aspirated consonants by voiceless sounds (*e.g., p, t*). The semivowels are *y, yu,* and *w* initially; *i, iu,* and *u* medially; and *i* and *u* (*o* after *a*) finally. Final retroflex *r* is written *r.* The tones are indicated by accent markers, 1 = -, 2 = ´, 3 = ˘, 4 = `, (*e.g., mā, má, mă, mà* = Wade-Giles *ma*^{1}, *ma*^{2}, *ma*^{3}, *ma*^{4}).

Wade-Giles is used in the following discussion of Modern Standard Chinese grammar.

Grammatical features of modern standard Chinese

The most common suffixes that indicate nouns are *-erh* (as in *ch'ang-erh* "song"; *cf. ch'ang* "sing"), *-tzu* (as in *fang-tzu* "house"), and *-t'ou* (as in *mu-t'ou* "wood"). A set of postposed noun particles express space and time relationships (*-li* "*inside,*" *-hou* "after"). An example of a verbal affix is *-chien* in *k'an-chien* "see" and *t'ing-chien* "hear." Important verb particles are *-le* (completed action), *-kuo* (past action), *chih* or *-che* (action in progress). The directional verbal particles *-lai* "toward speaker" and *-ch'ü* "away from speaker" and some verbal suffixes can be combined with the potential particles *te* "can" and *pu* "cannot"—*e.g., na-ch'ü-lai* "take out," *na-pu-ch'ü-lai* "cannot take out"; *t'ing-chien* "hear," *t'ing-te-chien* "can hear." The particle *te* indicates subordination and also gives nominal value to forms for other parts of speech (*e.g., wo* "I," *wo-te* "mine," *wo-te shu* "my book," *lai* "to come," *lai-te* "coming," *lai-te jen* "a person who comes"). The most important sentence particle is *le,* indicating "new situation" (*e.g., hsia-yü-le* "now it is raining," *pu-lai-le* "now there is no longer any chance that he will be coming"). *Ko* is the most common noun classifier (*i* "one," *i-ko-jen* "one person"); others are *so* (*i-so-fang-tzu* "one house") and *pen* (*liang-pen-shu* "two books").

Adjectives can be defined as qualitative verbs (*hao* "to be good") or stative verbs (*ping* "to be sick"). There are equational sentences with the word order subject–predicate—*e.g., wo-ihih pei-ching-jen* "I am a Peking-person (*i.e.,* a native of Peking)"—and narrative sentences with the word order subject (or topic)–verb–object (or complement)—*e.g., wo ch'ih-fan* "I eat rice," *wo chu tsai pei-ching* "I live in Peking." The proposed object takes the particle *ba* (*wo da ta* "I beat him," *wo ba ta da le yi dun* "I gave him a beating"), and the agent of a passive construction takes *bei* (*wo bei ta da le yi dun* "I was given a beating by him").

Standard Cantonese. The most important representative of the Yüeh languages is Standard Cantonese of Canton, Hong Kong, and Macau. It has fewer initial consonants than Modern Standard Chinese (*p, t, ts, k* and the corresponding aspirated sounds *ph, th, tsh, kh*; *m, n,* ŋ; *f, s, h*; *l, y*), only one medial semivowel (*w*), more vowels than Modern Standard Chinese, six final consonants (*p, t, k, m, n,* ŋ), and two final semivowels (*y* and *w*). The nasals *m* and ŋ occur as syllables without a vowel.

Cantonese tonal system

There are three tones (high, mid, low) in syllables ending in *-p, -t,* and *-k*; six tones occur in other types of syllables (mid level, low level, high falling, low falling, high rising, low rising). Two tones are used to modify the meaning of words (high level °, and low-to-high rising *), as in *yin*° "tobacco," but *yin* "smoke" and *nöy** "daughter," but *nöy* "woman." Some special grammatical words also have the tone °. There is no neutral tone and little tonal sandhi.

There are more than 2,200 different syllables in Standard Cantonese, or almost twice as many as in Modern Standard Chinese. The word classes are the same as in Modern Standard Chinese. The grammatical words, although phonetically unrelated, generally have the same semantic value (*e.g.*, the subordinating and nominalizing particle *kε*, Modern Standard Chinese *te*; *m* "not," Modern Standard Chinese *pu;* the verbal particle for "completed action" and the sentence particle for "new situation," both *le* in Modern Standard Chinese, are Standard Cantonese *tsɔ* and *lɔ,* respectively).

Min languages. The most important Min language is Amoy from the Southern branch of Min. The initial consonants are the same as in Standard Chinese. There are

Table 45: Romanization of Chinese and Chinese Numerals

consonants				approximate pronunciation
Chinese phonetic symbols		equivalent		
printed	name	EB preferred (Wade-Giles)	alternative (Pinyin)	
ㄅ	po	p	b	*bab*y
ㄆ	p'o	p'	p	*pepp*er
ㄇ	mo	m	m	*maim*
ㄈ	fo	f	f	*fif*ty
ㄉ	te	t	d	*did*
ㄊ	t'e	t'	t	*t*ie
ㄋ	ne	n	n	*no*
ㄌ	le	l	l	*lil*y
ㄍ	ko	k	g	*go*
ㄎ	k'o	k'	k	*k*in
ㄏ	ho	h	h	Ger. Bu*ch**
ㄐ	chi	ch	j	*j*eer†
ㄑ	ch'i	ch'	q	*ch*eap†
ㄒ	hsi	hs	x	Ger. Bü*ch*er
ㄗ	tzu	ts‡	z	bi*ds*
ㄘ	tz'u	ts'§	c	bi*ts*
ㄙ	ssu	s‖	s	*s*and
ㄓ	chih	ch	zh	*j*ug†
ㄔ	ch'ih	ch'	ch	*ch*ug†
ㄕ	shih	sh	sh	*sh*y
ㄖ	jih	j	r	¶
一	i	y	y	*y*ard
ㄨ	wu	w	w	*w*e

vowels				
ㄚ	ia	a	a	c*o*t♀
ㄜ	e	e♁	e	c*o*mmand
一	i	i	i	be*a*t
ㄓ ㄔ ㄕ ㄖ	chih, chi'ih, shih, jih	ih□	i	maj*or*♀
ㄗ ㄘ ㄙ	tzu, tz'u, ssu (szu)	*u*◇	i	b*ir*d♀
ㄛ	o	o	o	w*oo*d
ㄨ	wu	u	u	f*oo*d
ㄩ	yü	ü	ü▲	Ger. f*ü*hlen
ㄦ	erh	erh	er	j*our*ney+
ㄞ	ai	ai	ai	s*i*te
ㄠ	ao	ao	ao	n*ow*
ㄢ	an	an	an	n*on*

consonants				approximate pronunciation
Chinese phonetic symbols		equivalent		
printed	name	EB preferred (Wade-Giles)	alternative (Pinyin)	
ㄤ	ang	ang	ang	Fr. r*ang*
ㄟ	ei	ei	ei	f*a*de
ㄣ	en	en	en	*un*do
ㄥ	eng	eng	eng	h*ung*
ㄡ	yu	ou⊕	ou	kn*ow*
ㄨㄥ	ung	ung	ong	Austr. b*oong*
一ㄚ	ia	ia	ia	*ya*rd
一ㄝ	ieh	ieh	ie	*yi*p, id*ea*
一ㄠ	iao	iao	iao	*yow*l
一ㄡ	iu	iu	iu	*yo*ke
一ㄢ	ien	ien	ian	*yen*
一ㄣ	in	in	in	k*in*ky
一ㄤ	iang	iang	iang	*Yon*kers♀
ㄨㄚ	ua	ua	ua	*gua*va
ㄨㄛ	o	o**	uo	*wooe*r
ㄨㄞ	uai	uai	uai	*wi*de
ㄨㄟ	ui	ui††	ui	*way*
ㄨㄢ	uan	uan	uan	Fr. d*ouame*
ㄨㄣ	un	un	un	Fr. *jeune*, Ger. t*un*
ㄨㄤ	uang	uang	uang	*Wong*♀ (Chinese surname)
ㄩㄝ	üeh	üeh	üe	*new wi*cket
ㄩㄢ	üan	üan	üan	*new wen*
ㄩㄣ	ün	ün	ün	*new win*dow

numerals

Chinese/ Japanese	Arabic	Chinese/ Japanese	Arabic	Chinese/ Japanese	Arabic
〇	0	十一	11	二十二	22
一	1	十二	12	二十三	23
二	2	十三	13	二十四	24
三	3	十四	14	二十五	25
四	4	十五	15	二十六	26
五	5	十六	16	二十七	27
六	6	十七	17	二十八	28
七	7	十八	18	二十九	29
八	8	十九	19	三十	30
九	9	二十	20	一百	100
十	10	二十一	21	一千	1,000

*Velar, heavily aspirated *h*. †*j* and *ch* in *j*eer and *ch*eap are followed by front vowels; *j* and *ch* in *j*ug and *ch*ug are followed by back vowels. ‡Written *tz* before *ŭ*. §Written *tz'* before *ŭ*. ‖Written *ss* or *sz* before *ŭ*. ¶Retroflex *r*, which is pronounced with tongue at hard palate, slightly curled back. ♀In U.S., not British, pronunciation. ♁Written *o* after *k, kh, h*. □Occurs only after *ch, ch', sh, j*. ◇Occurs only after *tz, tz', ss (sz);* standard Wade-Giles uses *ŭ*. ▲Written without umlaut after *j, a, x*. +Retroflex *r*, with tongue at hard palate, slightly curled back. ⊕Written *u* after *y*. **Written *uo* after *k, k', h, sh*. ††Written *uei* after *k, k'*.

two semivowels (*y, w*), six vowels and several vowel clusters, plus the syllabic nasal sounds *m* and *ŋ* functioning as vowels, the same semivowels as in Standard Cantonese, and, in addition, a glottal stop (ʔ) and a meaning-bearing feature of nasalization, as well as a combination of the last two features. There are two tones in syllables ending in a stop, five in other syllables. Tonal sandhi operates in many combinations.

Fuchow is the most important language of the Northern branch of Min. The very extensive sandhi affects not only tones but also consonants and vowels, so that the phonetic manifestation of a syllable depends entirely on interaction with the surroundings. There are three initial labial sounds (*p, ph, m*), five dental sounds (*t, th, s, l, n*), three palatal sounds (*tś, tśh, ń*), and five velars (*k, kh, h,* ʔ, and ŋ). Syllables can end in -*k*, -*ŋ*, ʔ (glottal stop), a semivowel, or a vowel. The tones fall into two classes: a comparatively high class comprising high, mid, high falling, and high rising (only in sandhi forms) and a rather low one, comprising low rising and low rising-falling (circumflex). Certain vowels and diphthongs occur only with the high class, others occur only with the low class, and the vowel *a* occurs with both classes. Sandhi rules can cause tone to change from low class to high class, in which case the special vowels also change.

Tones of Fuchow

Other Sinitic languages. *Hakka.* Of the Kan (Kan-Hakka) languages, Hakka of Mei-hsien in Kwangtung is best known. It has the same initial consonants, final consonants, and syllabic nasals as Standard Cantonese; the vowels are close to Modern Standard Chinese. Medial and final semivowels are *y* and *w*. There are two tones in syllables with final stops, four in the other syllabic types.

Suchow. Suchow is usually quoted as representative of the Wu languages. It is rich in initial consonants, with a contrast of voiced and voiceless stops as well as palatalized and non-palatalized dental affricates, making 26 consonants in all. (Palatalized sounds are nonpalatal sounds formed with simultaneous movement of the tongue toward the hard palate. Dental affricates are sounds produced with the tongue tip touching the teeth and then drawing slightly away to allow air to pass through, producing a hissing sound.) Medial semivowels are as in Modern Standard Chinese. In addition, there are also ten vowels and four syllabic consonants (*l, m, n, ŋ*); *-n* and *ŋ* occur in final position, as do the glottal stop and nasalization.

Shanghai dialect. The Shanghai dialect belongs to Wu. The prevalent tendency is for there to be only two tones or registers (high and low), which are related in an automatic way to the initial consonant type (voiceless and voiced).

Hsiang languages. The Hsiang languages, spoken only in Hunan, are divided into New Hsiang, which is under heavy influence from Mandarin and includes the language of the capital Ch'ang-sha, and Old Hsiang, closer to the Wu languages, as spoken for instance in Shuang-feng. Old Hsiang has 28 initial consonants, the highest number for any major Sinitic language, and 11 vowels, plus the syllabic consonants *m* and *n*. It also uses five tones, final *n* and *ŋ*, and nasalization, but no final stops.

HISTORICAL SURVEY OF CHINESE

Vocabulary. Old Chinese vocabulary already contained many words not generally occurring in the other Sino-Tibetan languages. The words for "honey" and "lion," probably also "horse," "dog," and "goose," are connected with Indo-European and were acquired through trade and early contacts. (The nearest known Indo-European languages were Tocharian and Sogdian.) A number of words have Austro-Asiatic cognates and point to early contacts with the ancestral language of Muong-Vietnamese and Mon-Khmer; *e.g.,* the name of the Yangtze River, **klawŋ,* Cantonese *kɔŋ,* Modern Standard Chinese *chiang,* is still the word for "river," pronounced *kroŋ* and *kloŋ,* in some modern Mon-Khmer languages. Words for "tiger," "ivory," and "crossbow" are also Austro-Asiatic. The names of the key terms of the Chinese calendar ("the branches") have this same non-Chinese origin. It has been suggested that a great many cultural words that are shared by Chinese and Tai are Chinese loanwords from Tai. Clearly, the Chinese received many aspects of culture and many concepts from the Austro-Asiatic and Austro-Tai peoples whom they gradually conquered and absorbed or expelled.

Chinese practices in word borrowing and word creation

From the 1st century AD, China's contacts with India, especially through the adoption of Buddhism, led to Chinese borrowing from Indo-Aryan (Indic) languages, but, very early, native Chinese equivalents were invented. Sinitic languages have been remarkably resistant to direct borrowing of foreign words. In modern times this has led to an enormous increase in Chinese vocabulary without a corresponding increase in basic meaningful syllables. For instance, *t'ieh-lu* "railroad" is based on the same concept expressed in the French *chemin de fer,* using *t'ieh* "iron" and *lu* "road"; likewise, *tien-hua* "telephone" is a compound of *tien* "lightning, electricity" and *hua* "speech." A number of such words were coined first in Japanese by means of Chinese elements and then borrowed back into Chinese. The reason that China has avoided the incorporation of foreign words is first and foremost a phonetic one; such words fit very badly into the Chinese pattern of pronunciation. A contributing factor has been the Chinese script, which is ill-adapted to the process of phonetic loans. In creating new words for new ideas, the characters have sometimes been determined first and forms have arisen that cannot be spoken without ambiguity ("sulfur" and "lutecium" coalesced as *liu,* "nitrogen" and "tantalum" as *tan*). It is characteristic of Modern Standard Chinese that the language from which it most freely borrows is one from its own past: Classical Sinitic. In recent years it has borrowed from Southern Sinitic languages under the influence of statesmen and revolutionaries (Chiang Kai-shek was originally a Wu speaker and Mao Tse-tung a Hsiang speaker). Influence from English and Russian (in word formation and syntax) has been increasingly felt.

Historical periods: pre-Classical Chinese. The history of the Chinese language can be divided into three periods, pre-Classical (*c.* 1500 BC–*c.* AD 200), Classical (*c.* 200–*c.* 1920), and post-Classical Chinese (with important forerunners as far back as the T'ang dynasty).

The pre-Classical Chinese is further divided into Oracular Chinese (Shang dynasty [18th–12th century BC]), Archaic Chinese (Chou and Ch'in dynasties [1111–206 BC]), and Han Chinese (Han dynasty [206 BC–AD 220]).

Archaic Chinese—Early, Middle, and Late

Oracular Chinese is known only from rather brief oracle inscriptions on bones and tortoise shells. Archaic Chinese falls into Early, Middle (*c.* 800–*c.* 400 BC), and Late Archaic. Early Archaic is represented by bronze inscriptions, parts of the "Classic of History" (*Shu Ching*), and parts of the "Classic of Poetry" (*Shih Ching*). From this period on, many important features of the pronunciation of the Chinese characters have been reconstructed (see below). The grammar depended to a certain extent on unwritten affixes. The writing system kept apart forms with or without infixes. Early Archaic Chinese possessed a 3rd person personal pronoun in three cases (nominative *kyəg,* or *ghyəg,* accusative *cyəg,* genitive *kywat*). No other kind of written Chinese until the post-Classical period possessed a nominative of the 3rd person pronoun, but the old form survived in Cantonese (*khöy*) and is also found in Tai (Modern Thai *khăw*).

Middle Archaic Chinese comprises the earliest writings of the Confucian school. Important linguistic changes had taken place, which became still more pronounced in Late Archaic, the language of the two major Confucian and Taoist writers, Meng-tzu (Mencius) and Chuang-tzu, as well as of other important philosophers. The grammar by then had become more explicit in the writing system, with a number of well-defined grammatical particles, and it can also be assumed that the use of grammatical affixes had similarly declined. The process used in verb formation and verb inflection that later appeared as tonal differences may at this stage have been manifested in other suprasegmental features, such as different types of laryngeal phonation. The word classes included nouns, verbs, and pronouns, all with several subclasses and particles. The use of a consistent system of grammatical particles to form noun modifiers, verb modifiers, and several types of embedded sentences (*i.e.,* sentences that are made to become parts of another independent sentence) disappeared in Han Chinese and was gone from written Chinese until the emersion of post-Classical Chinese. In Modern Standard Chinese the subordinating particle *te* combines the functions of several Late Archaic Chinese particles, and the verb particle *le* and the sentence particle *le* have taken over for other Late Archaic forms.

Han and Classical Chinese. Han Chinese developed more polysyllabic words and more specific verbal and nominal (noun) categories of words. Most traces of verb formation and verb conjugation began to disappear. An independent Southern tradition (on the Yangtze River), simultaneous with Late Archaic Chinese, developed a special style, used in the poetry *Ch'u Tz'u* ("Elegies of Ch'u"), which was the main source for the refined *fu* (prose poetry). Late Han Chinese developed into Classical Chinese, which as a written idiom underwent few changes during the long span of time it was used. It was an artificial construct, which for different styles and occasions borrowed freely and heavily from any period of pre-Classical Chinese but in numerous cases without real understanding for the meaning and function of the words borrowed.

At the same time the spoken language changed continually, as did the conventions for pronouncing the written characters. Soon Classical Chinese made little sense when read aloud. It depended heavily on fixed word order and on rhythmical and parallel passages. It has sometimes been denied the status of a real language, but it was certainly one of the most successful means of communication in the history of mankind. It was the medium in which the poets T'ao Yüan-ming (365–427), Li Po (701–762), and Tu Fu (712–770) and the prose writer Han Yü (768–824) created some of the greatest masterpieces of all times and was the language of the Neo-Confucianist philosophy (especially of Chu Hsi [1130–1200]), which was to influence the West deeply. Classical Chinese was also the language in which the Italian missionary Matteo Ricci (1552–1610) wrote in his attempt to convert the Chinese Empire to Catholic Christianity.

Post-Classical Chinese. Post-Classical Chinese, based on dialects very close to the language now spoken in Northern China, probably owes its origin to the Buddhist storytelling tradition; the tales appeared in translations from Sanskrit during the T'ang dynasty (618–907). During the Sung dynasty (960–1206) this vernacular type of language was used by Buddhists and Confucianists alike for polemic writings; it also appeared in indigenous Chinese novels based on popular storytelling. From the Yüan dynasty (1206–1368) the vernacular was used also in the theatre.

Adoption of the National Language

Modern Standard Chinese has a threefold origin: the written post-Classical language, the spoken standard of Imperial times (Mandarin), and the vernacular language of Peking. These idioms were clearly related originally, and combining them for the purpose of creating a practical national language was a task that largely solved itself once the signal had been given. The term National Language (*kuo-yü*) had been borrowed from Japanese at the beginning of the 20th century, and, from 1915, various committees considered the practical implications of promoting it. The deciding event was the action of the May Fourth Movement of 1919; at the instigation of the liberal savant Hu Shih, it rejected Classical Chinese (also known as *wen-yen*) as the standard written language. (Hu Shih also led the vernacular literature movement of 1917; his program for literary reform appeared on January 1, 1917.) The new written idiom has gained ground faster in literature than in science, but there can be no doubt that the days of Classical Chinese as a living medium are numbered. After the establishment of the People's Republic of China some government regulation was applied successfully, and the tremendous task of making Modern Standard Chinese understood all over China was effectively undertaken. In what must have been the largest scale linguistic plan in history, untold millions of Chinese, whose mother tongues were divergent Mandarin or non-Mandarin languages or non-Chinese languages, learned to speak and understand the National Language; with this effort literacy was imparted to great numbers of people in all age groups.

The writing system. The Chinese writing system is nonalphabetic. It applies a specific character to write each meaningful syllable or each nonmeaningful syllabic that is part of a polysyllabic word.

Logographic form of writing

When the Chinese script first appeared, as used for writing Oracular Chinese (from *c.* 1500 BC), it must already have had a considerable development behind it. Although many of the characters can be recognized as originally depicting some object, many are no longer recognizable. The characters did not indicate the object in a primitive nonlinguistic way but only represented a specific word of the Chinese language (*e.g.,* a picture of the phallic altar to the earth is used only to write the word earth). It is therefore misleading to characterize the Chinese script as pictographic or ideographic; nor is it truly syllabic, for syllables that sound alike but have different meanings are written differently. Logographic (*i.e.,* a system using symbols representing entire words) is the term that best describes the nature of the Chinese writing system.

Verbs and nouns are written by what are or were formerly pictures, often consisting of several elements (*e.g.,* the character for "to love" depicts a woman and a child; the character for "beautiful" is a picture of a man with a huge head-dress with ram's horns on top). The exact meaning of the word is rarely deducible from even a clearly recognizable picture, because the connotations are either too broad or too narrow for the word's precise meaning. For example, the picture "relationship of mother to child" includes more facets than "love," a concept that, of course, is not restricted to the mother–child relation, and a man adorned with ram's horns undoubtedly had other functions than that of being handsome to look at, whereas the concept "beautiful" is applicable also to men in other situations, as well as to women. Abstract nouns are indicated by means of concrete associations. The character for "peace, tranquillity" consists of a somewhat stylized form of the elements "roof," "heart," and "(wine) cup." Abstract symbols have been used to indicate numbers and local relationships.

Related words with similar pronunciations were usually written by one and the same character (the character for "to love, to consider someone good" is a derivative of a similarly written word "to be good"). This gave rise to the most important invention in the development of the Chinese script—that of writing a word by means of another one with the same or similar pronunciation. A picture of a carpenter's square was primarily used for writing "work, craftsman; to work" and was pronounced *kuŋ*; secondarily it was used to write *kuŋ-* "to present," *ghuŋ* "red," *kuŋ* "rainbow," *klawŋ* "river," and others. During the Archaic period this practice was developed to such a degree that too many words came to be written as one character. In imitation of the characters that already consisted of several components an element was added for each meaning of a character to distinguish words from each other. Thus "red" was no longer written with a single component but acquired an additional component that added the element "silk" on the left; "river" acquired an additional component of "water," and so on. The original part of the character is now referred to as its phonetic and the added element as its radical.

Ch'in dynasty standardization of characters

During the Ch'in dynasty (221–206 BC) the first government standardization of the characters took place, carried out by the statesman Li Ssu. A new, somewhat formalized style known as seals was introduced—a form that generally has survived until now, with only such minor modifications as were necessitated by the introduction of the writing brush around the beginning of the Christian Era and printing around AD 600. As times progressed, other styles of writing appeared, such as the regular handwritten form *k'ai* (as opposed to the formal or scribe style *li*), the running hand *hsing,* and the cursive hand *ts'ao,* all of which in their various degrees of blurredness are explicable only in terms of the seal characters.

The Ch'in dynasty standardization comprised somewhat over 3,000 characters. In addition to archaeological finds, the most important source for the early history of Chinese characters is the huge dictionary *Shuo-wen,* compiled by Hsü Shu in *c.* AD 100. This work contains 9,353 characters, a number that certainly exceeds that which it was or ever became necessary to know offhand. Still, a great proliferation of characters took place at special times and for special purposes. The *Kuang-yün* dictionary of 1007 had 26,194 characters (representing 3,877 different syllables in pronunciation). The *K'ang-hsi tzu-tien,* a dictionary of 1716, contains 40,545 characters, of which, however, fewer than one-fourth were in actual use at the time. The number of absolutely necessary characters has probably never been much over 4,000–5,000 and is today estimated at fewer than that.

Attempts at phonetic writing system

By the 20th century the feeling had become very strong that the script was too cumbersome and an impediment to progress. The desire to obtain a new writing system necessarily worked hand in hand with the growing wish to develop a written language that in grammar and vocabulary approached modern spoken Chinese. If a phonetic writing were to be introduced, the classical language could not be used at all because it deviates so markedly from the modern language. None of the earlier attempts gained any following, but in 1919 a system of phonetic letters (inspired by Japanese kana) was devised for writing Mandarin. (In 1937 it received formal backing from the government, but World War II stopped further progress.) In 1929 a National Romanization, worked out by the author and language scholar Lin Yü-t'ang, the linguist Yuen Ren Chao, and others, was adopted. This attempt also was halted by war and revolution. A rival Communist effort known as *Latinxua,* or Latinization of 1930, fared no better. An attempt to simplify the language by reducing the number of characters to little over 1,000 failed because it did not solve the problems of creating a corresponding "basic Chinese" that could profitably be written by the reduced number of symbols.

The government of the People's Republic of China has taken several important steps toward solving the problems of the Chinese writing system. The first and basic step of making one language, Modern Standard Chinese, known all over the country has been described above. In 1956 a simplification of the characters was introduced that made them easier to learn and faster to write. Most of the abridged characters were well-known unofficial variants, used in handwriting but previously not in printing; some were innovations. In 1958 a romanization known as the phonetic system, or *pinyin zimu* (see Table 45), was introduced. This system is widely taught in the schools and is used for many transcription purposes and for teaching Modern Standard Chinese to non-Chinese peoples in China and to foreigners. The Pinyin is conceived as a script that will gradually replace characters.

Reconstruction of protolanguages. For reconstructing the pronunciation of older stages of Sinitic, the Chinese writing system offers much less help than the alphabetic systems of such languages as Latin, Greek, and Sanskrit within Indo-European or Tibetan and Burmese within Sino-Tibetan. Therefore, the starting point must be a comparison of the modern Sinitic languages, with the view of recovering for each major language group the original common form, such as Proto-Mandarin for the Northern languages, Proto-Wu and others for the languages south of Yangtze. Because data are still lacking from a great many places, the approach has until recently been to compare major representatives of each group for the purpose of reconstructing the language of the important dictionary *Ch'ieh-yün* of AD 601 (Sui dynasty), which mainly represents a Southern language type. One difficulty is that the language in a given area represents a mixture of at least two layers: an older one of the original local type, antedating the language of the *Ch'ieh-yün,* and a younger one that is descended from the *Ch'ieh-yün* language or a slightly younger but closely related tongue—the so-called T'ang-koine or the standard spoken language of the T'ang dynasty (618–907). The relationship of the protolanguages is further complicated by the fact that different substrata of non-Chinese stock underlie most if not all of the major languages.

Influence of the T'ang layer on Sinitic

The degree to which the Sinitic languages have been influenced by the T'ang (or Middle Chinese) layer varies. In the North the Old Chinese layer still dominates in phonology; in Min the two layers are kept clearly apart from each other and the Middle Chinese layer is most important in the reading pronunciation of the characters; in Yüeh both Chinese layers are of the Southern type and are typologically close to a Tai substratum.

The Old Chinese layer is characterized by early decay of final consonants, late development of tones from sounds or suprasegmental features located toward the end of the syllable, change of final articulation type because of similar initial type (as in syllables with voiced glottal activity both at the opening and at the end that lose the final voicedness; a phenomenon later manifested as a tonal change), and influence of sounds and tones in a syllable on those of surrounding ones (sandhi).

The New Southern stratum in Sinitic languages is characterized by early change of final articulation types into tones, extensive development of registers according to type of initial consonant, and late or no loss of final stops. The Old layer cannot be the direct ancestor of the New layer. The division into Northern and Southern dialects must be very old. It might be better to speak of a T'ang and a pre-T'ang layer or a T'ang or Han layer (the Han dynasty, 206 BC–AD 220, was one of extensive settling in most parts of what is now China proper).

Tones of the *Ch'ieh-yün* dictionary

For a long time it was assumed that the *Ch'ieh-yün* dictionary represented the language of the capital of the Sui dynasty, Ch'ang-an in the present province of Shensi, but recent research has demonstrated that its major component was the language of the present-day Nanking area with a certain attempt at compromise with Northern speech habits. As its first criterion for classifying syllables, the *Ch'ieh-yün* dictionary takes the tones, of which it has four: *p'ing, shang* (transcribed as :—*e.g., pa*:), *ch'ü* (transcribed as -—*e.g., pa-*), and *ju,* or even, rising, falling, and entering ("checked") tones. The entering tone comprised those syllables that ended in a stop (*-p, -t, -k*). The rising and falling tones may have retained traces of the phonetic conditioning factor of their origin, voiced and voiceless glottal or laryngeal features, respectively. The even tone probably was negatively defined as possessing no final stop and no tonal contour.

Next, the dictionary is divided according to rhymes, of which there are 61, and, finally, according to initial consonants. Inside each rhyme an interlocking spelling system known as *fan-ch'ieh* was used to subdivide the rhymes. There were 32 initial consonants and 136 finals. The number of vowels is not certain, perhaps six plus *i* and *u,* which served also as medial semivowels. There were probably more vowels than in either Archaic Chinese or in Modern Standard Chinese, another indication that the development of the Northern Chinese phonology has not passed the stage represented by *Ch'ieh-yün.*

There are additional sources for reconstructing the *Ch'ieh-yün* language: Chinese loanwords in Korean and Japanese (Japan has two different traditions—Go-on, slightly older than *Ch'ieh-yün* but representing a Southern language type like *Ch'ieh-yün,* and Kan-on, contemporary with *Ch'ieh-yün* but closer to the Northern tradition) and Chinese renderings of Indo-Aryan (Indic) words. Voiced stops are recovered through Wu, Hsiang, and Go-on (*e.g.,* Modern Standard Chinese *t'ien* "field," Wu and Hsiang *di,* Go-on *den, Ch'ieh-yün dhien*), final stops especially through Yüeh and Japanese (*e.g.,* Modern Standard Chinese *mu* "wood," Yüeh *muk,* Go-on *mok*[*u*], *Ch'ieh-yün muk*), and retroflex initial sounds from Northern Chinese (*e.g.,* Modern Standard Chinese *sheng* "to live," *Ch'ieh-yün* ṣʌŋ [the ṣ is a retroflex]).

Early Archaic Chinese is the old stage for which the most information is known about the pronunciation of characters. The very system of borrowing characters to write phonetically related words gives important clues, and the rhymes and alliteration of the "Classic of Poetry" (*Shih Ching*) furnish a wealth of details. Even though scholars cannot always be sure that prefixes and infixes are correctly recovered, and though the order in which recoverable features were pronounced in the syllable is not always certain (*rk-* or *kr-*, *-wk* or *-kw*, and so on), enough details can be obtained to determine the typology of Old Chinese and to undertake comparative work with the Tibeto-Burman and Karenic languages. The method employed in this part of the reconstruction of Chinese has been predominantly internal reconstruction, the use of variation of word forms within a language to construct an older form. As knowledge of the old layer of modern languages and dialects increases, however, the comparative method, which draws on similarities in several related tongues, gains importance. Through further internal reconstruction, features of the Proto-Sinitic stage, antedating Archaic Chinese, can then be restored.

TIBETO-BURMAN LANGUAGES—HISTORY AND CHARACTERISTICS

Influence of geography and social factors on Tibeto-Burman languages

The Tibeto-Burman languages have evolved from the ancestral language, Proto-Tibeto-Burman, in vastly different ways and at their own pace, in accordance with the geographical and social factors that have determined the fate of Central and South Asian peoples. Some tribes have been stationary; others have swept over huge areas. As a result, conservative or archaic features do not occur in only one contiguous part of the language area and innovations in another. The nearest genetic relations are often not identical with the closest typological ones.

Tibetan. Of the modern Tibetan languages and dialects, the Western ones have preserved initial consonant clusters and final stops most faithfully and have had the least compensational development of tones. Most Central languages and dialects, including Lhasa, have lost all consonant clusters and final stops and in the process have acquired a larger inventory of single consonants and a system of tones. These changes and reductions are linked to a similar reshaping of certain grammatical processes of word formation that now operate only through suprasegmental and syllabic elements. To a surprising degree, however, Modern Central Tibetan possesses grammatical categories identical with or very similar in content, though not in form, to those of Classical Tibetan (a similar relationship as that of Modern Standard Chinese to Old Chinese). The relationship of nouns to the main verb is indicated through postposed particles, the agent of a transitive verb indicated as the one by whom the action is performed, and the subject of an intransitive verb expressed as the object or goal of the action. Nominal modifiers precede nouns, and verbal modifiers follow them. The main verb, always placed after all nouns, is followed by particles expressing aspect and tense.

Old Tibetan pronunciation can be reconstructed by comparison of modern dialects and through the very conservative alphabetic script of Indian origin that goes back to the 7th century AD and found its present form in the 9th century. The orthography is far removed from present-day Standard Tibetan pronunciation.

Old Tibetan is one of the most archaic of the Tibeto-Burman languages. It retained Tibeto-Burman final stops and final *-r, -l, -s* and also the initial voiced consonants. Many Old Tibetan consonant clusters may be referred to Proto-Tibeto-Burman. The case particles and complicated verbal conjugation perhaps represent an elaboration on somewhat simpler tendencies in the protolanguage.

Himalayish languages

Some Himalayish languages are in certain respects as archaic as Tibetan, although most initial clusters are gone. An old feature is the connection of voiced–voiceless initial consonants with intransitive–transitive verbs. Because they have developed the feature of incorporating agent and object pronouns in verbs (and of possessive pronouns in nouns), these have been known as "pronominalizing" languages. An influence from contiguous Indo-European languages seems possible, but not certain.

Some Kirantish languages have retained consonant clusters and voiced initial consonants; others have given up both. Bahing has maintained the connection between voicedness and transitivity. Within Mirish, which has kept voiced initial sounds, Abor retains final stops and Dafla has some initial consonant clusters. Kachinish is quite conservative; prefixes are well retained as are voiced inital consonants, although some reshuffling has taken place in this respect.

Burmese. Within Burmish, Modern Standard Burmese has undergone a set of radical changes. Initial *ts-* and *tsh-* have become *s-* and *sh-*; *s* has become θ (*th* as in "thin"); *y-* and *r-* have coalesced as *y-*, and *ky-* and *kr-* as a palatal *c* (*ty*). Furthermore, all final consonants except nasals have coalesced as glottal stops, and all nasals have resulted in nasalization of the preceding vowel. In addition, the quality of vowels has been greatly altered. As was the case in Tibetan, in spite of great phonetic changes, grammatical categories are close to those scholars envisage for Proto-Tibeto-Burman. Cases of nouns and aspects of verbs are expressed through postposed particles.

Among the Burmese dialects, Arakanese is especially conservative, and the closely related language Maru is one of the most archaic within Tibeto-Burman in respect to final consonants. The Lolo languages lost most consonant clusters, as did all Burmish languages, but tend otherwise to be conservative in their treatment of initial sounds and radical in the loss and change of final consonants. Nung has retained final stops lost in Burmish (*-r, -l, -s*) and possesses a set of prefixes like Kachin.

The Burmese writing system

Study of the Burmese writing system, in combination with comparative linguistic work, makes possible the reconstruction of Old Burmese. The language of the Myazedi inscription of 1113 is close in its sound system to written Burmese in its present form, which dates back to at least the 15th century. The writing system was taken over from the Mon people, who had developed their writing from Pyu, a Sino-Tibetan language known in Burma from *c.* AD 500. It is alphabetic of an Indian type but represents a separate Southern line of development.

Old Burmese is phonetically further from Proto-Tibeto-Burman than is Tibetan. Initial clusters have mostly disappeared but are felt in the development of initial consonants. Some clusters with *-w-* and liquid sounds have been retained. The tonal system of Burmese (unlike that of Tibetan) developed to compensate for the loss of final features.

The Baric languages maintain a few older prefixes and developed some of their own. They tend to retain or merge *r* and *l.* The relationship of voicedness and transitivity is retained, but a great reshuffling of initial consonants took place, as it did also in Meithei and Kukish. The Kukish languages are found in all stages of development, but many of them are among the most archaic and most important for the reconstruction of Tibeto-Burman. Prefixes are best preserved in Old Kuki and in the Naga branch, whereas vowels seem very archaic in Lushai of the Central branch (the Lushai vowels have a difference of length that must in some way be explained in terms of Proto-Tibeto-Burman). Some Kukish languages incorporate pronouns in verbs. Mru, distantly related to Kukish, is noted for a number of archaic features, including final consonants lost elsewhere in Tibeto-Burman.

(S.C.E.)

TAI LANGUAGES

The name Tai denotes a family of closely related languages, of which the Thai, or Siamese, language of Thailand is the most important member. Because the word Thai has been designated as the official name of the language of Thailand, it would be confusing to use it for the various other languages of the family as well. Tai is therefore used to refer to the entire group.

Geographical distribution

Spoken in Thailand, Laos, Burma, Assam in northeast India, northern Vietnam, and the southwestern part of China, the Tai languages together form an important group of languages in Southeast Asia. In some countries they are known by different tribal names or by designations used by other peoples; *e.g.,* there is Shan in Burma; Pai-i in Yunnan, China; Chuang-chia in Kwangsi, China; Chung-chia, Dioi, Jui, Yoi, Yay, or Pu-yi in Kweichow, China; Tho, Nung, White Tai, Black Tai, Red Tai, etc. in northern Vietnam; and Khün, Lü, etc. in Thailand and Laos. Ahom, an extinct language once spoken in Assam, has a considerable amount of literature. The Tai languages are divided into three linguistic groups—the Southwestern, the Central, and the Northern. Thai and Lao, the official languages of Thailand and Laos, respectively, are the best known of the languages.

The number of Tai speakers is estimated at 64,000,000. Of these, 38,565,000 are in Thailand, 17,800,000 in China, and about 7,900,000 in Laos, northern Vietnam, and Burma. There are tremendous variations between several estimates, and these figures may serve as only rough indications of the Tai populations.

Relationship to other languages. The Tai languages are traditionally assumed to be related to the Sino-Tibetan family of languages with which they share similar phonological structures, tone systems, and some lexical items; the relationship, however, has never been definitely established. The Tai tongues are related to the Kam-Sui languages of Kweichow, China, and to the Bê language in Hainan; but because of the great divergences in both vocabulary and phonological development, the name Tai is used only to designate a more restricted group that excludes such languages as Kam, Sui, and Bê. It has also been suggested, chiefly on the basis of some similar vocabulary items, that the Tai languages are related to the Kadai languages of southwestern China and Hainan Island and ultimately to the Austronesian (Malayo-Polynesian) languages, rather than to the Sino-Tibetan tongues.

Classification. Classifications have been made according to the geographical location of the Tai speakers, social, political, and cultural criteria, and literacy versus nonliteracy. The classification used for this article is based on linguistic relationships proposed in 1959–60; the criteria for it are lexical (involving similarities in vocabulary) and phonological (involving similarities in sounds and systems

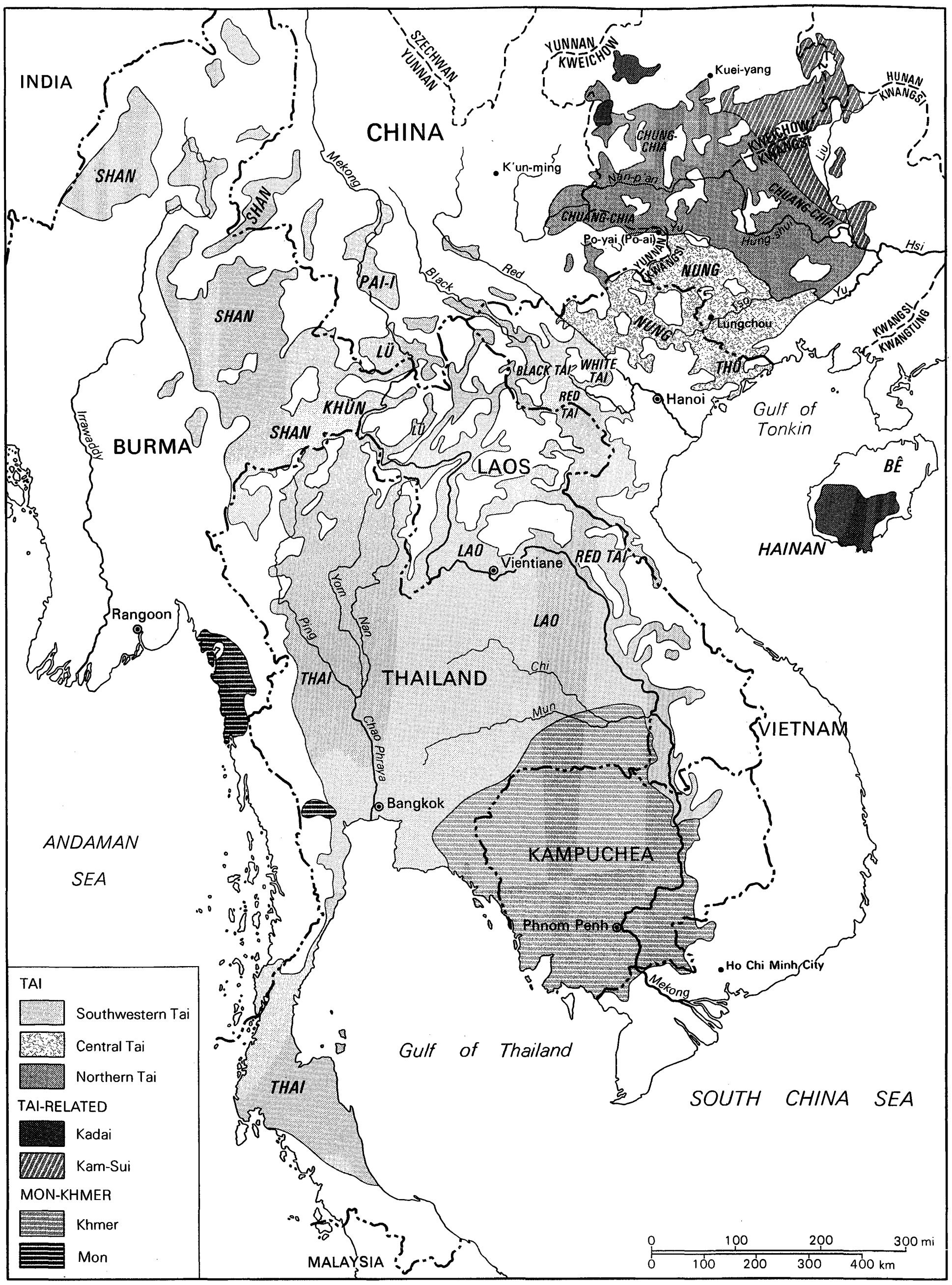


Figure 25: Major divisions of the Tai languages and neighbouring languages.

Adapted from Frank M. LeBar, Gerald C. Hickey, and John K. Musgrave, *Ethnic Groups of Mainland Southeast Asia;* New Haven, Human Relations Area Files Press, 1964

of sounds). According to these features the Tai languages are divided into the three groups mentioned above (see map). Languages of the Southwestern group are spoken in Thailand, Laos, northern Vietnam, Burma, and Yunnan, China; they include Thai, or Siamese, Lao, Shan, Khün, Lü, White Tai, Black Tai, etc. The Southwestern division, geographically the most widespread group, consists of two-thirds of the Tai-speaking population and represents an expansion in comparatively recent periods. To the Central group belong the Tho dialects spoken in northern Vietnam and the various dialects spoken in Kwangsi, such as Lungchow. The Chung-chia or Pu-yi dialects in Kweichow and the Chuang-chia dialects in Kwangsi belong to the Northern group. Some of the Northern dialects are also spoken in Yunnan and Vietnam, and one, called Saek, is spoken as far south as Laos and Thailand.

Main groups of Tai

The fairly large number of vocabulary items shared by languages of these three groups suggests their genetic relationship. There are also certain items that are shared by only two of the groups and not found in the other. For instance, the word for sky is shared by the Southwestern dialects (Siamese *fáa*) and the Central dialects (Lungchow *faa*), but another word is used in the Northern dialects (Po-ai *mïn*). Similarly, the word for beard is shared by the Central group (Lungchow *mum*) and the Northern group (Po-ai *mum*) but is replaced by another word in the Southwestern group (Siamese *nùat*); the term for knife is shared by the Southwestern (Siamese *mîit*) and the Northern groups (Po-ai *mit*) but not by the Central dialects. There are also vocabulary items that are found only in one group. In all, the evidence seems to indicate that there are three groups of dialects in the Tai family.

Different phonological features may be attached to some words according to the dialect group. For instance, the Southwestern forms for the verb "to be" (Siamese *pen*) are derived from a protoform **pɛn* (with the vowel like *e* in "egg"), whereas the Central dialect forms (Lungchow *pin*) and the Northern forms (Po-ai *pan*) come from a protoform **bɛn,* as indicated by the tone. (A protoform is the presumed or reconstructed ancestral form of a word; an asterisk [*] marks those forms that are unattested and reconstructed.) Similarly, the Southwestern and

Dialect groups

the Central forms for the classifier for animals (Siamese *tua,* Lungchow *tuu*) are derived from a protoform **tua,* whereas the Northern forms (Po-ai *tuu*) are attributed to a protoform **dua.* (A classifier is a term that indicates the group to which a noun belongs, like "animate object," or designates countable objects or measurable quantities, like "yards" [of cloth] and "head" [of cattle]). Such words as the forms for "to be" and the classifier for animals are good indications of dialect boundaries.

In phonological development, the Northern dialects differ from the rest in not maintaining the distinction between aspirated and unaspirated voiceless stops. That is, the dialects have lost the feature of aspiration, which sounds like a puff of breath accompanying a consonant. Aspiration may, however, be reintroduced in some dialects by later borrowing or secondary developments. The Central dialects differ from the other groups in the treatment of certain Proto-Tai consonant clusters, such as **tr-* and **thr-*. Although they have changed from the protoforms, these are usually kept distinct in the other groups—*e.g.,* in Siamese as *taa* ("eye") and *haaŋ* ("tail"), in Po-ai as *taa* and *lïïŋ*. In the Central dialects, however, they have merged into a single sound—*e.g.,* Tho *thaa* and *thaaŋ*, Lungchow *haa* and *haaŋ*.

Phonological characteristics. The sound system of the Tai languages may best be described in terms of its syllabic structure. Each syllable consists of an initial consonant or consonant cluster followed by a vowel or vowel cluster (long vowel or diphthong), which may be further followed by a final consonant, usually a nasal sound or an unreleased stop. (An unreleased stop is a consonant in which there is complete stoppage of the airstream from the lungs, and in which the tongue or the lips maintain the position of the consonant without opening the stoppage.) In addition, each syllable has a tone. As an illustration of this type of structure, the system of the Thai, or standard Siamese, language as spoken around the Bangkok area may be given. Labials are sounds using the lips as articulators; alveolars are formed by placing the tip of the tongue at the gums behind the upper teeth; sibilants, or fricatives, involve local friction and a hissing sound resulting from incomplete closure in the mouth. Velars are sounds made with the back of the tongue touching the soft palate (velum); glottal sounds involve constriction of the vocal cords; semivowels are gliding sounds with some of the properties of both vowels and consonants; and liquids are frictionless consonants, produced with incomplete closure in the vocal tract, that sometimes function as vowels.

Five tones

There are five tones: level (*a*), low (*à*), falling (*â*), high (*á*), and rising (*ǎ*); for example, *maa* "to come," *màak* "areca nut," *mâak* "much," *máa* "horse," and *mǎa* "dog" are differentiated by the various tones.

Grammatical characteristics. The statements about morphology and syntax refer particularly to the standard language of Thailand, though they are applicable in general to all of the Tai languages. Words or morphemes (word elements) are, for the most part, monosyllabic, but there are also many polysyllabic words, mainly compounds and loanwords from Sanskrit and from Khmer, an Austroasiatic language spoken in Cambodia. There are no inflections in Tai comparable to *-ed* in English "walked" or *-s* in "dogs," but derivations are common. Derivation is the formation of new words by the addition of endings, prefixes, or other words, often resulting in a change of the part of speech or meaning of a term—*e.g.,* English "unionize" or "unionization" from "union." Compounding is the chief Tai method of derivation—*e.g., nâa-taa* "countenance (face-eye)," *kèp-kìaw* "to harvest (gather-cut with a sickle)." Reduplication, the repetition of a word or part of it, is common—*e.g., díi-dii* "very good" from *dii* "good." Partial reduplication is also found, such as *sanùk-sanǎan* "to enjoy oneself" from *sanùk* "to have fun." Prefixes and infixes occur often in Indo-Aryan and Khmer loans—*e.g., pra-thêet* "country" (from Sanskrit *pradeśa*) and *d-amn-əən* "to proceed" (from Khmer *dae* "to walk") and *d-amn-aə* "process." There also developed some native prefixes that are abbreviated forms of what once were full words. For example, *ma-,* a prefix used in many names of fruits, such as *ma-phráau* "coconut," *ma-mûaŋ* "mango," is derived from *màak* "areca nut" (originally "fruit"). Similarly, in the word *sa-dïï* "navel," *sa-* is the reduced form of *sǎaj* "line, string," which refers to the umbilical cord. Old processes of derivation involved using the alternation of consonants or of tone or both, such as *níi* "this" and *nîi* "here," *n´ɔ̆ɔj* "small, little" and *n`ɔ̆ɔj* "a little bit," *khiaw* "sickle" and *kìaw* "to cut with a sickle." Such processes are, however, no longer active.

Use of nominal predicates

A sentence is usually formed by a noun phrase (subject) followed by a verb phrase (predicate). There are, however, nominal predicates as well, in which the verb is lacking and a noun is used instead; an example is *wan-níi wan-sǎw* "today (is) Friday." A noun phrase consists of a noun, which may be followed by its modifiers (another noun, an adjective, or a verb phrase), which, in turn, are followed by a numeral with a classifier, and finally, by a demonstrative. For example, *tûu* "cabinet" and *náŋsïï* "book" become *tûu-náŋsïï* "bookcase"; *tûu-kèp náŋsïï* is "bookcase" ("cabinet-keep-book"); *tûu-yen* is "icebox" ("cabinet-cool"), and *tûu-náŋsïi sɔɔŋ-bai níi* is "these two bookcases" ("cabinet-book two-classifier these"). Classifiers must be used with numerals and vary according to the nouns they are to enumerate. It is not uncommon that a noun may serve as its own classifier.

A verb phrase may consist of an adjective or of a verb often followed by its object or its complements or both. A common type of complement denotes the direction of action by using such words as *paj* "to go," *maa* "to come," *khiŋ* "to ascend," *loŋ* "to descend," and so on. Examples are *ˀaw náŋsïï paj* "Take the book away!"; *ˀaw náŋsïï maa* "Bring the book (here)!"; *aw náŋsïï khïn maa* "Bring the book up!"; and *aw náŋsïï loŋ paj* "Take the book down!". A number of particles, particularly those occurring at the end of a sentence, are used to indicate a question or a command, to express emphasis or uncertainty, to show politeness and the sex of the speaker, etc. For example, *kháw maa lɛ́ɛw* "He has come"; *kháw maa máj?* "Is he coming?"; *khun maa máj khráp*? "Are you coming?" (man speaking, polite); *khun maa máj khâ*? "Are you coming?" (woman speaking, polite).

Table 46: The Sounds of Thai (Siamese)

	initial consonants*					final consonants	
Labials	p	ph	b	m	f	p	m
Alveolars	t	th	d	n		t	n
Sibilants	c	ch			s		
Velars	k	kh		ŋ		k	ŋ
Glottals	ˀ				h	ˀ	
Semivowels	j	w				j	w
Liquids	l	r					

	vowels			long vowels			diphthongs		
High	i	ï	u	ii	ïï	uu	ia	ïa	ua
Mid	e	ə	o	ee	əə	oo			
Low	ɛ	a	ɔ	ɛɛ	aa	ɔɔ			

*Certain consonant clusters are also permitted initially, such as *kl*, *kr*, *khl*, *khr*, *tr*, *pl*, *pr*, *phl*, *phr*, *kw*, and *khw*.

Vocabulary. The various Tai languages have been in contact with many different languages spoken in the same area of Southeast Asia, and it is inevitable that words from different sources have been adopted by them. There is a basic core of vocabulary that is shared by most of the Tai languages and may be considered as the native vocabulary of the protolanguage (earlier form of the Tai languages). Nevertheless, items in this list may be found to resemble forms in other languages, such as Chinese or Indonesian, and they have given rise to different interpretations of the relationship of the Tai languages to other linguistic families. Loanwords in later periods from the Khmer and Indo-Aryan languages (Sanskrit and Pāli) are particularly common in Siamese and Lao, and loans from Chinese are abundant in the Central and the Northern dialects.

Loan words

Writing. There are two kinds of writing used among the Tai languages. One, ultimately derived from Chinese, is used by the Central and the Northern dialects; the other comes from Indo-Aryan sources and is used in many languages of the Southwestern group. The Chinese-based system, employed chiefly to write songs, consists of both

Table 47: The Thai Alphabet

consonants

middle		high		low			
ก k		ข kh		ค kh	ฆ kh	ง ŋ	
จ c		ฉ ch		ช ch	ฌ ch	ญ j	
ฎ d	ฏ t	ฐ th		ฑ th	ฒ th	ณ n	
ด d	ต t	ถ th		ท th	ธ th	น n	
บ b	ป p	ผ ph		พ ph	ภ ph	ม m	
		ฝ f		ฟ f			
		ศ s	ษ s	ย j	ร r	ล l	ว w
		ส s		ซ s			
อ ʔ		ห h		ฮ h	ฬ		

vowels*

short	long
—ะ -aʔ	—า -aa
—ั/ -a-	
—ิ -i(ʔ)	—ี -ii
—ึ -ï(ʔ)	—ือ -ïï
—ุ -u(ʔ)	—ู -uu
เ—ะ -eʔ	เ— -ee
แ—ะ -ɛʔ	แ— -ɛɛ
โ—ะ -oʔ	โ— -oo
— -o-	
เ—าะ -ɔʔ	—อ -ɔɔ
ไ— -aj	—า ย -aaj
ใ— -aj	
เ—า -aw	—า ว -aaw
ฤ rï(ʔ)	ฤ า rïï
ฦ lï(ʔ)	ฦ า lïï

vowel clusters†

—ัว -ua	—ว/ -ua-
เ—ียะ -ia	เ—ีย -ïa
เ—อ -əə	เ—ิ/ -əə-
เ—ย -əəj	—ำ -am

tonal markers (used with vowel symbols)

(1) — middle tone with middle and low consonants; rising tone with high consonants

(2) —่ low tone with middle and high consonants; falling tone with low consonants

(3) —้ falling tone with middle and high consonants; high tone with low consonants

(4) —๋ rising tone chiefly with middle consonants

*Pronunciation key (for symbols not used in the standard Roman alphabet): ɛ = *a* as in *bad;* ï = the *u*-sound in some pronunciations of *just* (*e.g.*, "just a minute"); ɔ = *o* as in *law;* ə = *a* as in *sofă;* ʔ = glottal stop (the catch sound for *tt* in Cockney or Brooklynese *bottle*). †For which the long-short distinction does not apply.

Chinese characters and modified Chinese characters, very much like the early writing in Vietnam. A specimen of this type of writing dates back perhaps to the 18th century, but it may have been in use much earlier. The script adapted from Indo-Aryan sources for the Tai languages dates perhaps from the 13th century. The earliest known example of such a writing system is the inscription of Ramkhamhaeng in northern Thailand from AD 1293. The Modern Thai alphabet (see Table 47) is a modified form of the original writing. It preserves the old distinction of voiced (low) and voiceless (middle or high) consonants, a distinction that is now lost, but leaves its effects on the tone. This system also provides an unambiguous method for indicating the vowels and tones. Similar types of writing are used in Lao, Lü, White Tai, Black Tai, etc. Another type of Indo-Aryan writing, used in Shan and Ahom, lacks the difference between high and low consonants, has an insufficient number of vowel signs, and provides no tone marks. Although the shapes of the letters in Shan differ greatly from those in Ahom, the basic principle is essentially the same.

(F.K.L.)

PALEO-SIBERIAN LANGUAGES

The collective term Paleo-Siberian is applied to four genetically unrelated language groups situated in northern Asia—Yeniseian, Luorawetlan (Luoravetlan), Yukaghir (Yukagir), and Gilyak. The Yeniseian group, whose only living member is Ket (or Yenisey-Ostyak), is spoken by about 800 persons in the Turukhansk region along the Yenisey River. Kott (Kot), Arin, and Assan (Asan), now extinct members of this group, were spoken to the south of the present-day locus of Ket. The Luorawetlan family consists of (1) Chukchi, spoken by 11,600 people in the northeasternmost parts of Siberia, west of the small enclave of Siberian Eskimo; (2) Koryak, also called Nymylan, with approximately 6,500 speakers, found to the south of Chukchi; (3) the more remotely related Kamchadal (or Itelmen), with a bare remnant of 500 speakers in southern Kamchatka; (4) Aliutor, perhaps a Koryak dialect, with a small and unknown number of speakers; and (5) Kerek, with about 100 speakers. Some Soviet scholars list Aliutor and Kerek as being more closely related to Chukchi and Koryak than to Kamchadal.

The Yukaghir group, whose only living member is Yukaghir proper (or Odul), is spoken by about 400 persons in two enclaves in the Yakut Autonomous Soviet Socialist Republic (Yakutskaya A.S.S.R.), near the estuary of the Indigirka and along the bend of the Kolyma River. Extinct languages belonging to the Yukaghir group (or perhaps dialects of an earlier form of Yukaghir proper) are Omok and Chuvan (Chuvantsy); these were spoken south and southwest of Yukaghir proper. Gilyak (or Nivkh) has about 2,200 speakers, 1,400 of whom live in the estuary of the river Amur and 800 on the island of Sakhalin.

Paleo-Siberian groups not genetically related

Classification of the languages. These four groups are *not* related to each other. They have been subsumed under the names Paleo-Siberian, Paleo-Asiatic, or, more rarely, Hyperborean, ever since the Baltic German zoologist and explorer Leopold von Schrenck surmised, in the middle of the 19th century, that they constituted the remnants of a formerly more widely dispersed language family that had been encroached upon by invading groups of Uralic, Turkic, and Tungus speakers. Schrenck's hypothesis is quite correct to the extent that as recently as the 17th century Yeniseian, Luorawetlan, and Yukaghir languages were spoken over much wider territories than they are today. For example, it is known that Samoyed languages (of the Uralic family) at one time in the past absorbed the languages of now extinct Yeniseian tribes, that Yukaghir was spoken as far west as the Taymyr Peninsula in the 17th century, and that the former domains of Chukchi and Koryak extended much further to the west. Little is known about the prehistory of Gilyak but it may be assumed that this language was also originally centred further to the west, perhaps in Manchuria. As far as can be determined with the help of the methods of the comparative linguistics, however, the four present-day Paleo-Siberian groups never formed a single family of languages in the accepted sense of that term. In fact, they may represent only a fragment of a possibly greater diversity of language families in prehistoric Siberia. Many of the languages spoken in the

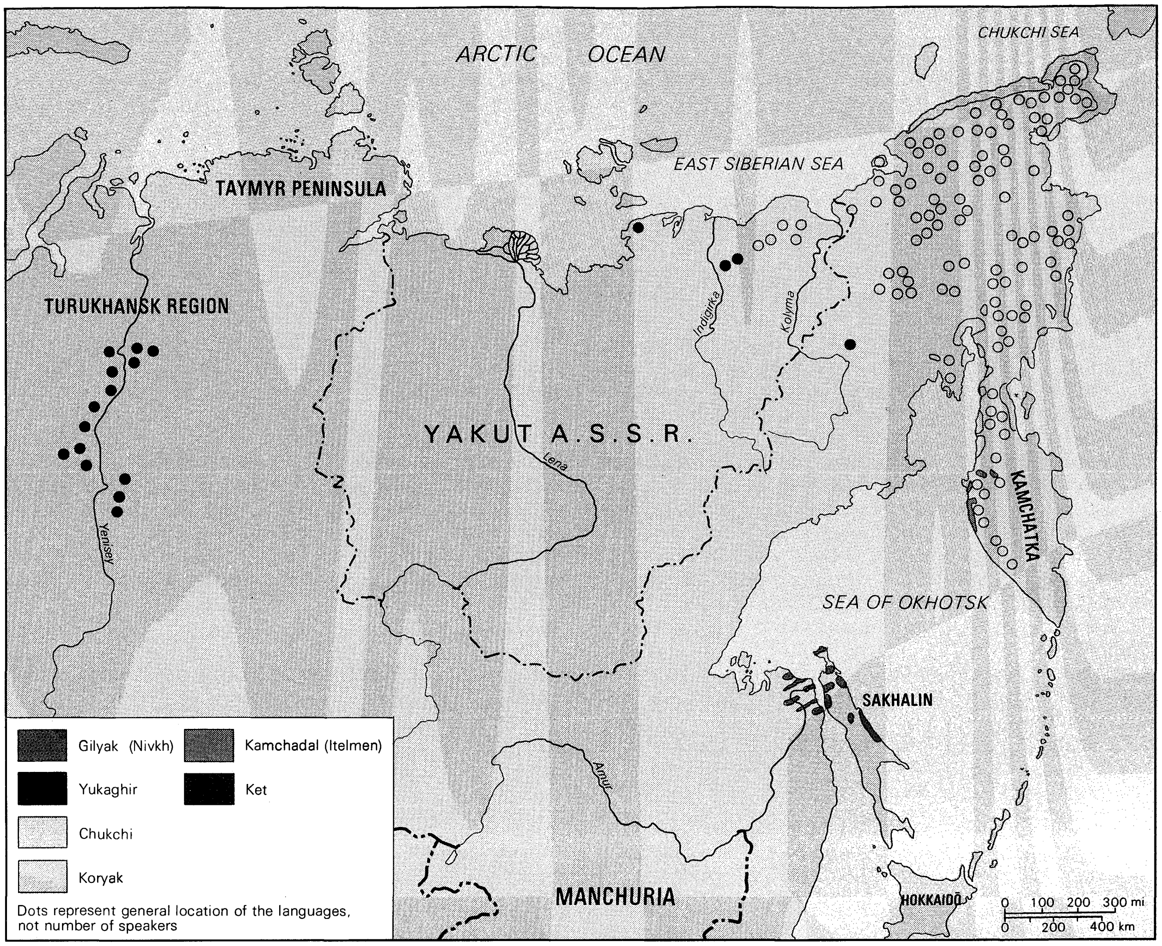


Figure 26: Distribution of Paleo-Siberian languages.

From M. Levin and L. Potapov (eds.), *The Peoples of Siberia,* translated by Scripta Technica, Inc., translation edited by Stephen Dunn, published by University of Chicago Press; © 1964 by The University of Chicago. All rights reserved. Published 1964. Printed in the United States of America by Scripta Technica, Inc.

area during earlier periods may have been swallowed up by the more recent and culturally vigorous intruders in Siberia that are now the neighbours of the Paleo-Siberian enclaves: mainly the Yakut (whose domains stretch as far as the Chukchi and Yukaghir areas) and various Tungus tribes (one or another of which borders on each Paleo-Siberian language).

Comparisons with other language families

Nevertheless, many attempts have been made to show that the four Paleo-Siberian families are either related to each other or to adjacent (or more distant) language families. Thus, Ket has been compared with the Sino-Tibetan family (Chinese, Tibetan) and with some of the languages of the Caucasus, and Yukaghir has been compared with Uralic. Some of these comparisons are fanciful experiments or completely unfounded (*e.g.,* the comparisons of Ket with Caucasian languages). Others are more reasonable but not compelling (*e.g.,* the Yukaghir-Uralic hypothesis) because the evidence adduced so far has been too unsystematic and fragmentary. It is therefore safest, at present, to consider Ket, Yukaghir, and Gilyak as language isolates that are unrelated to any known language or language family and to regard Luorawetlan as a family in its own right that is also unrelated to any other family or isolated language. Further research on the internal history of these four groups will eventually enable comparatists to reconstruct hypothetical earlier stages of the languages and to make comparisons with non-Paleo-Siberian languages more cogent, if at all feasible. Mere resemblances in grammatical or phonological traits between Paleo-Siberian and adjacent languages (such as between Chukchi and Eskimo, or between Gilyak and Korean or Japanese) are not indexes of genetic affinity, but are often the result of the diffusion of linguistic traits over large geographical areas. They may, however, provide clues to the linguistic prehistory of Siberia.

The cultures of the four Paleo-Siberian groups are, in general, similar in that they are all Arctic or subarctic. In detail, however, each group of speakers of a Paleo-Siberian language has its own characteristic cultural profile. These characteristics may on occasion even resemble the cultural profile of a non-Paleo-Siberian group very closely; *e.g.,* Ket culture resembles Selkup (Ostyak-Samoyed) culture more closely than it resembles that of any Paleo-Siberian group. (Selkup- and Ket-speaking groups are located in contiguous areas.)

Linguistic characteristics. *Grammar.* The grammatical structures of the four Paleo-Siberian groups differ considerably from each other. In a broad sense, Gilyak resembles Japanese in its grammatical categories and processes (in word order, heavy inflection of verbs, and use of enclitics—words closely connected with the word that precedes), whereas Yukaghir shares certain grammatical categories with some Uralic languages (the objective conjugation—*e.g.,* "he shot it" versus "he shot"—and the negative conjugation—*e.g.,* Yukaghir *tet mer-ai-mek* "you shot" versus *tet el-ai-yek* "you did not shoot"). A typical feature of Luorawetlan is its strong tendency toward complex compounding, also called incorporation, and circumfixation; *e.g.,* in Chukchi *ga* + *mor-ĭk* + *tor* + *orw-ĭma* "in our new sleigh," the entire unit is surrounded by the circumfix *ga-* . . . *-ĭma* "in" (compare *ga* + *mor-ĭk* + *orw-ĭma* "in our sleigh," without *tor* "new," and *ga* + *tor* + *orw-ĭma* "in the new sleigh," without *mor-ĭk* "our"). A characteristic feature of the Ket verb is its succinct complexity, involving such categories as gender, animateness, and type of event; *e.g., t-k-it-n-a* "I carved it up," which consists of *t-* "I," the verbal complex *k-* . . . *-a* "cut up (carve, split) into pieces once," *-it-* (feminine object marker "her, it"), and *-n-* (past-completed tense). All of the Paleo-Siberian languages are rich in devices for compounding words. In syntax, Luorawetlan favours ergative constructions in which markers indicate the agent or instrument of the action; *e.g., Father*

+ agent marker, *bear* (subject), *shoot* (main verb), "Father is shooting a bear."

Phonology. Typical phonological features of the Paleo-Siberian languages are post-velar consonants (*i.e.,* sounds formed further back in the mouth than *k,* usually represented as *q*), vowel harmony of various kinds (*e.g.,* the alternation of *e* and *i* in the form for "my" in Gilyak *ñe-řla* "my harpoon" and *ñi-řly* "my sky"), consonant alternations (*e.g.,* the alternation between *b, v,* and *f* in Gilyak *bal* "mountain," *ñ-val* "my mountain," *c-fal* "thy mountain"), and rich consonant clusters in all but Yukaghir.

Loanwords in Paleo-Siberian languages

Vocabulary. In addition to the stock of native words inherited from its ancestral language, each Paleo-Siberian language also has numerous loanwords, some of which are recent and from adjacent or recently adjacent languages, and others of which are ancient, from languages with which it no longer has contact. Some of the loanwords from ancient times are, of course, more difficult to identify. In general, Tungus, a branch of the Altaic family, is the source of most loanwords, but the Turkic languages (including Yakut) have also served as the sources of loans, and Ket has some words from Selkup. There are also more complicated loan relationships, such as are found in the reindeer terminology of Gilyak, which is borrowed from a Tungus language but seemingly not from any of the Tungus languages with which Gilyak is now in contact. South Sakhalin Gilyak also contains a considerable number of loanwords from Ainu (a language of northern Japan) and was, during 1905–45, hospitable to potential loans from Japanese; the Japanese loanwords never became acculturated because the Japanese hegemony over South Sakhalin ceased after World War II. Chukchi has some Eskimo loans.

The most viable source of technical and all of the other modern vocabulary has been the Russian language, the influence of which began with the first contact and continues to be strong. Each Paleo-Siberian language adjusts the Russian loans according to the dictates of its phonology and grammar but the most recent borrowed words tend to retain their original Russian form or one closely resembling it.

Writing. The Yukaghir had a tradition of pictographic writing (incisions on fresh birch bark) used by men for route maps and by young women for a type of love letter. Limited use of such a system among the Koryak speakers has also been reported.

Since the 1920s and 1930s each Paleo-Siberian language has had a literary language and a script now based on the Cyrillic alphabet (and formerly based on the Latin script). Because at one time these native languages were used in part in elementary education, primers and arithmetic books for the lowest grades were available. Some natives continue their education and acquire a good knowledge of Russian and of Soviet culture. This has led to the rise of bilingualism but has also contributed to the growth of modern literatures in the native languages, based on Russian models, especially among the Koryak and the Chukchi.

The native traditions and folklore of the Paleo-Siberian peoples have been collected since the last century, mainly by Russians and Westerners. Work in these fields is still continuing and is attracting a slowly emerging corps of trained native specialists. Such trained natives are also beginning to collaborate in the compilation of dictionaries.

(R.Au.)

CAUCASIAN LANGUAGES

The term Caucasian languages as used in this article includes groups of languages indigenous to the Caucasus region, between the Black and Caspian seas within the Soviet Union; it excludes the Indo-European (Armenian, Ossetic, Talysh, Kurdish, Tat) and Turkic languages (Azerbaijan, Kumyk, Noghay, Karachay, Balkar) and some other languages of the area, all of which were introduced to the Caucasus in historical times. The Caucasian languages are also referred to as Paleocaucasian and Ibero-Caucasian languages.

The Caucasian languages are found in the territory north and south of the main Caucasian mountain range; their number varies, according to different classifications, from 30 to 40. The concentration of so many languages in such a small territory is indeed remarkable. There are more than 6,600,000 speakers of Caucasian languages; their language communities range in size from only a few hundred people to large national groups of millions.

Three Caucasian language groups

The Caucasian languages fall into three typologically well-defined language families: the Northwest Caucasian, or Abkhazo-Adyghian, languages; the Northeast Caucasian, or Nakho-Dagestanian, languages; and the South Caucasian, or Kartvelian, languages (also called Iberian). From the typological point of view, the Northwest and Northeast Caucasian groups present opposite structural types, with South Caucasian holding an intermediary position.

The exact genetic relationships of the Caucasian languages are still unclear on many points, not only in regard to interrelationships of the three major groups but also to some internal groupings. Although the genetic relationships between Northwest and Northeast Caucasian seem probable, the interrelationships of North and South Caucasian are as yet uncertain because of the absence of any regular sound correspondences between them. At the present stage of comparative Caucasian linguistics, North Caucasian and South Caucasian must be viewed as separate language families.

The theories relating Caucasian with such languages as Basque and the non-Indo-European and non-Semitic languages of the ancient Near East also lack sufficient evidence and must be considered as inconclusive.

SOUTH CAUCASIAN (KARTVELIAN) LANGUAGES

Languages of the group. The Kartvelian (South Caucasian) language family comprises Georgian, Mingrelian (Megrelian), Laz (or Chan), and Svan. The speakers of these languages constitute the Georgian nation and numbered 3,647,000 in the early 1980s.

Georgian. Georgian (self-designation: *kartulis ena*), used as the language of literature and instruction, is the state language of the Georgian Soviet Socialist Republic. It is common to all speakers of the Kartvelian languages within the Georgian S.S.R. Beyond the Georgian republic, Georgian is spoken in the adjacent regions of the Azerbaijan S.S.R. and northeast Turkey. There are also 14 villages of Georgian speakers in the province of Isfahan, Iran.

The designation Georgian that is used in the European languages was coined during the Crusades; it is based on Persian *gorji* (Georgian), from which the Russian *gruzin* was also derived. The Greek term *ībēres* (Georgians) is connected with an Old Iranian name for Georgia.

The dialects of Georgian fall into two groups—East and West Georgian—divided by the Suram Mountains. These exhibit only slight differences.

Georgian literary tradition

Among the Caucasian languages, only Georgian has an ancient literary tradition, which dates back to the 5th century AD, when the oldest datable monuments were inscribed in an original script. This old Georgian script must have been derived from a local variety of Aramaic with influences—in regard to the order of the alphabet and the shape of some characters—from the Greek alphabet. The modern Georgian writing system is based on the round-form cursive, which was developed from the angular book script of the 9th century; the latter was a direct descendant of the old Georgian script. The Georgian writing system accurately reflects the distinctive sounds of the language.

During the Old Georgian period (from the 5th to the 11th century), original and translated literary monuments were produced, among them the Georgian translation of the Bible. The conventions of the New Georgian literary language, ultimately established in the middle of the 19th century on the basis of an East Georgian dialect, originated in the secular literature of the 12th century. New Georgian differs structurally in many respects from Old Georgian, but the old language is still comprehensible to the Georgians of today. Until the beginning of the 19th century, Old Georgian was still in use in religious services and theological writings.

Mingrelian. The Mingrelian language (self-designation: *margaluri nina*) is spoken in the territory north of the Rioni River and west of the Tskhenis-Tskali and along the Black Sea coast from the mouth of the Rioni up to the city of Ochamchire. The language is unwritten.

Laz. The Laz language (self-designation: *lazuri nena*) is spoken along the Black Sea coast from the Chorokh River (Georgian S.S.R.) to south of Pazar (Atina) in Turkish territory. The language is unwritten, Georgian being used as the literary language in the Georgian S.S.R. and Turkish in Turkey. In view of the structural closeness between Mingrelian and Laz, they are sometimes considered as dialects of a single language.

Svan. The Svan language (self-designation: *lušnu nin*), also unwritten, is located south of Mt. Elbrus, in the high valleys of the upper Tskhenis-Tskali and its tributary Kheledula and in the valleys of the upper Ingur. There are four fairly distinct dialects: Upper and Lower Bal in the Ingur region, and Lashkh and Lentekh in the Tskhenis-Tskali region.

Linguistic characteristics. Correspondences between sounds and meanings in words and word elements provide a basis for considering the Kartvelian languages as being closely related and descended from a common ancestral language (a protolanguage).

Vowel differences in Kartvelian

Phonology. The sound system of the Kartvelian languages is relatively uniform, with only the vowel systems exhibiting considerable differences. Apart from the five cardinal vowels *a, e, i, o, u,* which exist in all the Kartvelian languages, the Svan dialects show several additional vowels: the front (or palatalized) vowels, *ä, ö, ü,* and a high central vowel, *ə* (as the *a* in English "sofa"). All these vowels also have distinct lengthened counterparts, thus giving a total of 18 distinctive vowels in some dialects of Svan. Vowel length is not distinctive in the other Kartvelian languages.

Within the Kartvelian consonant system the stops and affricates have voiced, voiceless, and glottalized varieties. (Stops are produced by complete but momentary stoppage of the breath stream some place in the vocal tract; affricates are sounds begun as stops but released with local friction, such as the *ch* sounds in "church." Voiced sounds are made with vibrating vocal cords; in voiceless sounds, the vocal cords do not vibrate; glottalized consonants, indicated in phonetic transcription by dots below or above certain letters, are pronounced with an accompanying closure of the glottis (the space between the vocal cords). Fricative sounds (*e.g., s, z, v*), which are characterized by local friction, have only voiced and voiceless types.

Although most word roots begin with one or two consonants, instances of long consonant clusters in word-initial position occur quite frequently, especially in Georgian, in which such clusters may comprise up to six consonants; *e.g.,* Georgian *prckvna* "peeling," *msxverp̣li* "sacrifice," *brʒola* "fighting."

Grammatical characteristics. The Kartvelian languages exhibit a developed system of word inflection (*e.g.,* the use of endings, such as English "dish, dishes" or "walk, walks, walked") and derivation (word formation). Derivation is characterized by compounding, the combination of words to form new words, as well as by affixation, the addition of prefixes and suffixes; *e.g.,* Georgian *kartvel-i* "Georgian," *sa-kartvel-o* "Georgia"; Mingrelian *žir-i* "two," *mažir-a* "second."

The verb system distinguishes the categories of person, number (singular and plural, with differentiation of inclusive and exclusive plural in Svan), tense, aspect, mood, voice, causative, and version (the latter defines the subject–object relations). These categories are expressed mainly by the use of prefixes and suffixes, as well as by internal inflection (changes within the verb stem), which is frequently a redundant grammatical feature.

Multipersonal verbs

The system of verb conjugation is multipersonal; that is, the verb forms can indicate the person of the subject (the agent) and of the direct or indirect object by the use of special prefixes. (The subject of the 3rd person is marked by endings in Georgian and Mingrelo-Laz and by a lack of ending in Svan.) An example is Georgian *m-c̣er-s* "he writes to me," *m-xaṭav-s* "he paints me," in which *m* denotes the 1st person as object and *s* marks the 3rd person as subject. The finite verb forms fall into three series of tenses: the present tense, the aorist (indicating occurrence, usually past, without reference to completion, duration, or repetition), and the perfect or resultative (denoting an action in the past not witnessed by the speaker).

There is a developed system of preverbs, elements preceding the verb stem and attached to it, with local meaning indicating location of the action in space, as well as its direction (especially in Mingrelian and Laz). Simple preverbs are combined into complex ones. The preverbs are also used to mark the aspect (nature of the action indicated by the verb, with reference to its beginning, duration, completion), which is used for the formation of future and aorist forms; *e.g.,* Georgian *c̣er-s* "he writes" versus *da-c̣er-s* "he will write" and *da-c̣er-a* "he wrote."

The nominal (noun, pronoun, adjective) system is distinguished by less structural complexity than the verb system and has cases varying in number from six to 11. The six cases common to all the Kartvelian languages are: nominative, marking subject of the intransitive verb; ergative (see below), modified in Mingrelian and Laz; genitive, marking possession; dative, marking indirect objects; ablative–instrumental, expressing relations of separation and source and means or agency; and adverbial, expressing goal of the action—*e.g.,* "to make it." There are also some secondary local cases (in New Georgian, Mingrelian) that indicate location and direction toward the object as well as from the object (rendered in English by such prepositions as "in," "on," "to," "from," and so on). The nominal system does not distinguish gender, which is absent even in pronouns, and there are no special articles (such as English "a," "the").

Ergative sentence constructions

A basic feature of Kartvelian syntax is the ergative construction of the sentence. The subject of a transitive verb (the agent) is marked by a special agentive, or ergative, case, while the case of the direct object is the same as that of the subject with intransitive verbs, traditionally called the nominative case; *e.g.,* Georgian *ḳac-i* (nominative) *midis* and Svan *māre* (nominative) *esɤri,* "the, or a, man goes" but Georgian *ḳac-ma* (ergative) *moḳla datv-i* (nominative) and Svan *mārēm* (ergative) *adgär däšdw* (nominative) "the, or a, man killed the, or a, bear." A specific feature of the Georgian and Svan ergative construction is its restriction to the aorist series. In the present-tense series the subject (agent) of transitive as well as intransitive verbs is put into the same nominative case, and the direct object is in the dative; *e.g.,* Georgian *ḳac-ma* (ergative) *moḳla datv-i* (nominative) "the man killed a bear" (aorist), but *ḳac-i* (nominative) *ḳlavs datv-s* (dative) "the, or a, man kills the, or a, bear" (present tense). In Mingrelian the ergative case in *k* extends in the aorist series to the constructions with intransitive verbs and results in a formation of two distinct subject cases. In Laz, conversely, the case in *k* extends to the constructions with transitive verbs in the present-tense series.

Vocabulary. The genetic closeness of the Kartvelian languages is evidenced by a large number of structural correspondences and of common lexical (vocabulary) and grammatical items. Though the Kartvelian languages abound in ancient loanwords from Iranian, Greek, Arabic, Turkish, and other languages, it is nevertheless possible to single out the basic vocabulary and grammatical elements of original Caucasian origin, which exhibit a system of regular sound correspondences. The common Kartvelian vocabulary comprises the kinship terms, names of animals, birds, trees, and plants, the parts of the body, as well as different human activities, qualities, and states. The words for the numerals from one to ten and the word for hundred are also original common Kartvelian terms.

Proto-Kartvelian. A comparative study of the Kartvelian languages enables specialists to outline the general structure of the parent language, called Proto-

Ablaut in Proto-Kartvelian

Kartvelian, which yielded the known Kartvelian, or South Caucasian, languages. One of the most characteristic features of the Proto-Kartvelian language is the functional vowel alternation, or ablaut; different forms of a word root or word element appear either with a vowel (**e, *a, *o*), called full grade, or without a vowel, called zero grade. (An asterisk [*] indicates that the following form is not attested but has been reconstructed as a hypothetical ancestral form.) In a sequence of word elements (called morphemes) only one element may occur in full grade, the others being in either zero or reduced grade forms (*i.e.,* in a form with **i*). To a word root with a full-grade vowel, for example, a suffix in zero may be added, and vice versa: **der-ḳ-* (intransitive) "stoop, recline" and **dr-eḳ-* (transitive) "bend." When a full-grade ending is added to these stems, the preceding full-grade element is shifted to zero or a reduced grade; *e.g., *der-ḳ-* plus the ending **-a* becomes **dṛ-ḳ-a.* In such patterns the lengthened grade, a long vowel, may also appear.

These ablaut patterns, strikingly parallel to those of the Indo-European languages, and other linguistic features may have arisen in Proto-Kartvelian as a result of contacts with Indo-European at a comparatively early date. Such contacts between Kartvelian and Indo-European are further evidenced by a number of Indo-European loanwords in Proto-Kartvelian, such as Proto-Kartvelian **ṭep* "warm" (compare Indo-European **tep* "warm"), Proto-Kartvelian **ṃḳerd* "breast" (compare Indo-European **ḱerd* "heart"), and others.

In Mingrelo-Laz the ancient ablaut patterns were eliminated and new forms were set up with a stable, noninterchanging vowel in each word element. The ancient ablauting models were better preserved in Georgian and especially in Svan, in which new ablauting patterns, in addition to the old structures, were established.

The pronominal system of Proto-Kartvelian is characterized by the category of inclusive–exclusive (*i.e.,* there are two forms of the pronoun "we," one including the hearer, and the other excluding him), which survived in Svan but has been lost in other languages of the family. Along with it, Svan has preserved a certain number of archaic structural features of the Proto-Kartvelian epoch, setting it apart from Georgian and Mingrelo-Laz, which share a number of common lexical and grammatical innovations. Svan must have been separated fairly early from the rest of Proto-Kartvelian, which later yielded the Mingrelo-Laz and Georgian languages. (Th.V.G.)

NORTH CAUCASIAN LANGUAGES

The North Caucasian languages are divided into two groups: Abkhazo-Adyghian, or the Northwest Caucasian, languages, and Nakho-Dagestanian, or the Northeast Caucasian languages.

Abkhazo-Adyghian languages. The Abkhazo-Adyghian group consists of the Abkhaz, Abaza, Adyghian, Kabardian (Circassian), and Ubykh languages. Abkhaz, with about 93,000 speakers, is spoken in the Abkhazian A.S.S.R. (in the South Caucasus, Georgian S.S.R.). The other languages are spread over the western part of the northern Caucasus. Abazians (43,000) live in the Karachay-Cirassian Autonomous Oblast; Adyghians (111,000), in the Adyghe Autonomous Oblast; Kabardians or Kabardino-Circassians (332,000) dwell mainly in the Kabardino-Balkar A.S.S.R. Both Adyghians and Kabardians call themselves *adəge.* The Ubykh language was formerly found to the north of the area where Abkhaz is spoken, in the vicinity of Tuapse. In 1864 Ubykhians as well as a substantial part of the Abkhaz- and Adyghe-speaking population migrated to Turkey, where before long they lost their native tongue. In the early 1970s Ubykh was spoken by about 20 people living near the Sea of Marmara. The total number of people speaking Abkhazo-Adyghian languages is 579,000 (in the U.S.S.R.). Many Circassians speaking Abkhazo-Adyghian languages live in the countries of the Near East—Turkey, Jordan, and Iraq.

All Abkhazo-Adyghian languages, with the exception of Ubykh, are written. From the dialectological point of view, the Abkhazo-Adyghian languages are not widely differentiated, the differences being mainly of phonetic character. In Abkhaz two dialects are distinguished; Adyghian and Kabardian differentiate four dialects each. Abkhaz and Abaza are very close to each other and are considered by some scholars to be dialects of the same language. The same kind of affinity exists between Adyghian and Kabardian. Ubykh occupies an intermediate position between the Abkhaz-Abaza and Adyghe-Kabardian languages.

Few vowels, very numerous consonants

Phonology. A characteristic feature of the sound system of the Abkhazo-Adyghian languages is a rather limited number of distinctive vowels—*a* and *ə* (pronounced as the *a* in English "sofa"). Some scholars consider it possible to posit only one vowel, which, depending on the position, can be realized in different ways: *a, ə, i, o, e.* On the other hand, the languages are notable for a great diversity in their consonant systems. The number of consonants distinguished reaches about 70 (in the Abkhaz and Adyghian languages) or even 80 (Ubykh). Along with the consonants that occur in all the Caucasian languages, the Abkhazo-Adyghian languages are characterized by different sets of labialized consonants (formed by rounding the lips), strong (hard or tense) consonants, half-hushing consonants, and velarized consonants (formed with the back of the tongue approaching the soft palate).

Grammatical characteristics. The grammatical characteristics of the Abkhazo-Adyghian languages include an extremely simple noun system and a relatively complicated system of verb conjugation. There are no grammatical cases in Abkhaz and Abaza, and in the other languages only two principal cases occur: a direct case (nominative) and an oblique case, combining the functions of ergative, genitive, dative, and instrumental. In nouns, possession is expressed by means of pronominal prefixes; *e.g.,* Abkhaz *sarra s-č:ə* "my horse" (literally: "I my-horse"), *wara u-č:ə* "your horse" (pertaining to a man), *bara b-č:ə* "your horse" (pertaining to a woman), and so forth. (The symbol *:* indicates that the preceding consonant is a strong consonant.)

The Abkhaz and Abaza languages distinguish the grammatical classes of person and thing (the latter class includes all nouns denoting nonhuman objects). The class of person also differentiates between the subclasses of man and woman.

The verb in the Abkhazo-Adyghian languages has a pronounced polysynthetic character; that is, various words combine to form a composite word that expresses a complete statement or sentence. The most important verbal categories are expressed by prefixes, although suffixes also form tenses and moods. The principal verb categories are: dynamic versus static, transitivity, person, number, class, tense, mood, negation, causative, version, and potentiality. "Dynamic versus static" is a verb form expressing action versus state of being; "version" is a verb category denoting for whom the action is intended (compare Georgian *v-ç̣er* "I write," but *v-u-ç̣er* "I write for him"); "potentiality" is a category expressing the possibility of an action (*e.g.,* Abkhaz *s-zə-ɯuam* "I cannot write"). The verb is multipersonal and can denote up to four persons.

Adverbial relationships (such as "where," "when," "how") are expressed by prefixes following the personal markers. On the whole, the verb forms appear as a long string of word elements expressing the above mentioned categories; *e.g.,* Abkhaz *i-u-z-d-aa-sə-r-g-an* "that (thing)-you (a man)-for-them-hither-I shall make-bring" (*i.e.,* "I shall make them bring that for you"). In a sequence of prefixes, up to nine morphemes are possible.

Three sentence types

The simple sentence has three constructions: indefinite, nominative, and ergative (in Abkhaz and Abaza only indefinite). An indefinite construction has the subject in the indefinite case (*i.e.,* not marked with a special suffix); a nominative construction has the subject in the nominative case. The same personal markers, depending on their arrangement, can denote both the subject and various objects: *e.g.,* Abkhaz, *wara sara u-s-šwejṭ* "I kill you (a man)," *sara wara s-u-šweiṭ* "you (a man) kill me."

Nakho-Dagestanian languages. The Nakho-Dagestanian group consists of the Nakh and Dagestanian languages. Some investigators subdivide the Nakho-Dagestanian languages into two independent groups: Central Caucasian languages (Nakh) and East Caucasian languages (Dagestan), although the great proximity of these groups, and their equal remoteness from the Abkhazo-Adyghian languages, may justify regarding them as a common group of languages.

The Nakh languages consist of Chechen (792,000 speakers), Ingush (192,000), and Bats (or Tsova-Tushian, about 3,000 speakers). The Chechens and Ingushes live in the Chechen-Ingush A.S.S.R.; the Bats dwell in the village Zemo-Alvani in the Akhmeta district of the Georgian S.S.R. Both Chechen and Ingush, which are fairly close to one another, are written. The Bats language is unwritten; Georgian is used as the literary language for Bats speakers, and they consider Georgian as their mother tongue.

Groups of Dagestan languages

The Dagestan languages are numerous. The following groups can be distinguished:

1. The Avar-Ando-Dido languages: These occupy the central and western part of Dagestan and part of the Zakataly region of the Azerbaijan S.S.R. The member languages are the Avar language; the Andi subgroup of languages, including Andian, Botlikh, Godoberi, Chamalal, Bagulal, Tindi, Karata, and Akhvakh; and the Dido subgroup, including Dido, or Tsez, Khvarshi, Hinukh, Bezhta, and Hunzib, or Kapucha.

The Avar-Ando-Dido language with the most speakers (about 504,000) is Avar, a written language. All Andi-Dido languages are unwritten, and most of them are spoken by about 3,000 to 5,000 people each. From ancient times the Andi-Dido nationalities used the Avar languages for intertribal communication. Avar is still widely known and spoken among them. The Andi languages are phonetically and grammatically very close to each other. The same affinity is observed among the Dido languages, to the effect that Hinukh is considered by some specialists as a dialect of Dido, while Bezhta and Hunzib are viewed as two dialects of the same language. In respect to dialectology, the majority of Avar-Ando-Dido languages are widely differentiated.

2. The Lakk-Dargwa languages: Lakk (or Lakh, with 103,000 speakers) and Dargwa (or Khjurkili, with 301,000) are spoken in the central part of Dagestan. Both are written languages. The Lakk language is quite homogeneous with regard to its dialects; Dargwa, however, possesses several diversified dialects—sometimes considered as separate languages (*e.g.,* Kubachi). Some view Lakk and Dargwa as independent language groups.

3. The Lezgian languages: This group includes Lezgi (Lezghi), or Kuri (with 195,000 speakers in the Dagestan A.S.S.R. and about 163,000 in the Azerbaijan S.S.R.); Tabasaran (80,000); Agul (about 13,000); Rutul (about 16,000); Tsakhur (about 15,000); Archi (1,000); Kryz (or Dzkek, about 6,000); Budukh (1,000); Khinalug (about 1,000); and Udi (about 7,000). The majority of Lezgi languages are found in Dagestan occupying its southern part, but some of them (Kryz, Budukh, Khinalug, Udi, partly Tsakhur) are spoken in Azerbaijan; and one village of Udi speakers is located in Georgia. It must be noted that in Azerbaijan, as well as earlier in Russia, all Dagestanians—including Avars—called themselves Lezginians. Among the Lezgian languages, Lezgi and Tabasaran are written. The inclusion of Archi in this group gives rise to some doubt, and it has also been noted that the Khinalug language stands out from the Lezgian group in many respects. The Udi language is supposed to be one of the languages of ancient Caucasian Albania.

Phonology. The sound systems of the Nakho-Dagestanian languages are diverse. There are up to five vowels (*a, e, i, o, u*); in some languages *o* is only now becoming an independent distinctive unit. Along with these cardinal vowels, in a number of languages there are also long and nasalized vowels (the Andi languages), pharyngealized vowels (in Udi), and labialized vowels (in Dido). In the Nakh languages (Chechen) the vowel system is fairly intricate, the number of distinctive vowels amounting to 30 (including diphthongs and triphthongs).

Nakh and Dagestanian consonants

The consonant systems of the Nakh languages are relatively simple, coinciding, on the whole, with those of the South Caucasian languages (apart from a number of pharyngeal consonants characteristic of all the Nakh languages and a lateral sound peculiar to Bats). The opposition of strong and weak voiceless consonants is typical of the majority of the Dagestanian languages. This contrast has been lost in a number of languages and dialects; *e.g.,* in the Dido languages, in some dialects of Avar. The labialized clusters *kw, qw, sw,* and so on are widespread. In the Avar-Ando-Dido languages and in Archi there are fricative and affricate lateral sounds (*i.e.,* different types of *l*), with the maximum possible number being six (in Akhvakh).

All the Caucasian languages have a series of stops of three types—voiced, voiceless aspirated, and glottalized (*i.e.,* pronounced, respectively, with vibrating vocal cords; with vocal cords not vibrating but with an accompanying audible puff of breath; and with accompanying closure of the glottis [space between the vocal cords]). In some languages strong and weak consonants also contrast. Usually, in the languages with a strongly developed vowel system, the system of consonants is comparatively simple (*e.g.,* Chechen, Ingush, Dido), and vice versa (*e.g.,* Avar, Lakk, and Dargwa have complicated consonantisms and relatively simple vowel systems).

Grammatical characteristics. There are several common structural features in morphology (word structure), the most characteristic being the existence of the grammatical category of classes (eight classes in Bats; six in Chechen and Andi; five in Chamalal; four in Lakk; three in Avar; two in Tabasaran).

In a number of languages (Lezgi, Udi), noun differentiation by classes has disappeared. The class of "thing" is distinguished from the "person" class, which can be differentiated into the subclasses of man and woman. Compare, for example, Avar *emen w-ač̣ana hani-w-e* "father has come here" (in which *w* is equivalent to the marker of the class of man), *ebel j-ač̣ana hani-j-e* "mother has come here" (in which *j* is equivalent to the marker of the class of woman), and *ču (kaɤat) b-ač̣ana hani-b-e* "a horse (a letter) has come here" (in which *b* is equivalent to the marker of the class of thing). In the plural there are usually fewer grammatical classes denoted.

Nouns have many cases, both in singular and in plural; there are cardinal cases (nominative, ergative, genitive, dative) and local cases that denote the location of a thing ("on," "in," "near," "under"), with a specification of movement ("where," "which way," "from where," "over what"). The ergative case, the case of the real subject of transitive verbs, is present in all the Nakho-Dagestanian languages. Nouns have different stem forms in the nominative and the oblique (non-nominative) cases; *e.g.,* Avar *gamač̣* "a stone" (nominative), *ganč̣-i-c:a* (ergative), and *ganč̣-i-da* "on the stone." In pronouns the category of inclusive–exclusive is distinguished; *e.g.,* Avar *niĺ* "we with you," *niž* "we without you."

The class of the noun in the nominative case (*i.e.,* in the case of the subject of intransitive verbs and of the direct object of transitive verbs) is reflected in the verb; *e.g.,* Avar: *was* (nominative, class I) *w-ač̣ana* "the boy has come," *jas* (nominative, class II) *j-ač̣ana* "the girl has come."

In the Lezgi language, a characteristic structural feature is agglutination, the combination of various elements of distinct meaning into a single word. A typical feature of Nakho-Dagestanian syntax is the presence of the ergative construction of the sentence (the subject of transitive verbs is put in the ergative case and the real object in the nominative case). Complex sentences are usually formed with participial and adverbial–participial construction; *e.g.,* Avar *haniwe wač̣araw či dir wac: wugo* "the man who arrived here is my brother" (literally, "the here arrived man my brother is").

Vocabulary, writing, and alphabets. The original vocabulary of the North Caucasian languages has been fairly well preserved in the modern languages, although there is a substantial number of words borrowed from Arabic (through Islām), the Turkic languages, and Persian. There are also loanwords from the neighbouring languages (Georgian, Ossetic). Russian has played a major part since the late 19th century and is currently the main source for new words, especially technical terminology.

Sources of loanwords

The written languages of the area are the state languages. Newspapers, magazines, and books, as well as radio and television programs, use the local languages, and children in primary schools are taught in their mother tongue.

The alphabets of the North Caucasian written languages (Abkhaz, Abaza, Adyghe, Kabardian, Chechen, Ingush, Avar, Lakk, Dargwa, Lezgi, Tabasaran) are based on the Cyrillic alphabet, which was introduced for these languages in 1936–38 (in Abkhaz, from 1954). Previously, from 1928, the modified Latin alphabet was used; it superseded the Arabic script, which was adapted to the local languages in the Soviet period, at a time when a number of North Caucasian languages became literary languages.

Some written attempts had been made earlier. In the 18th century an insignificant number of monuments were created (with the use of Arabic writing) in Lakk and Avar. Stone crosses with Old Georgian–Avar bilingual inscriptions, dating from not later than the 14th century AD, have been preserved in central Dagestan. (T.E.G.)

HAMITO-SEMITIC LANGUAGES

The Hamito-Semitic languages, a family of genetically related languages, developed from a common parent language that presumably existed about the 6th–8th millennia BC and was perhaps located in the present-day Sahara. Also known as the Semito-Hamitic, Erythraean, Afro-Asiatic, and Afrasian language group, it is the main language family of northern Africa and southwestern Asia and includes such languages as Arabic, Hebrew, Amharic, and Hausa. The total number of speakers is estimated to be more than 200,000,000.

The term Hamito-Semitic, or Semito-Hamitic, was introduced by a German Egyptologist, Karl Richard Lepsius, in the 1860s. Although it has become traditional, it is an unfortunate label in suggesting that the family is divided into a group of Semitic and a group of Hamitic languages; in fact, the family has at least four other branches of the same order as the Semitic languages. The term Erythraean is inappropriate in implying that the family originated on both shores of the Red Sea, an assumption that cannot be proved; and Afro-Asiatic (proposed by a U.S. linguist, Joseph Greenberg, in 1950) may be too comprehensive insofar as it suggests that all the languages of Africa and Asia are included. Igor Diakonoff, a Soviet linguist, has suggested the term Afrasian, meaning "half African, half Asiatic," which corresponds to the area of the actual distribution of the languages of this family since at least the 5th millennium BC.

The five branches of Hamito-Semitic

The languages belonging to the Hamito-Semitic family can apparently be subdivided into branches representing dialects of the original parent language—namely, Semitic, Egyptian, Berber, Cushitic, and Chadic. Some linguists deny the genetic affinity of the Chadic languages with the other branches of Hamito-Semitic, while others (*e.g.,* Joseph Greenberg) accept it. Certain scholars have expressed doubts concerning the Hamito-Semitic character of some of the Chadic languages but not of others. Among the linguists who classify the Chadic languages as Hamito-Semitic there is some hesitation as to the degree and character of their affinity with the languages of the Cushitic branch, especially with West Cushitic. On the basis of the low percentage of vocabulary items held in common between the West Cushitic languages and the other Cushitic languages, some scholars classify West Cushitic as a separate branch of Hamito-Semitic, called Omotic. There is, however, a probability that the parent language common to Omotic and the Cushitic languages proper is not the Common Hamito-Semitic protolanguage but a later dialect (namely, Common Cushitic) and that Omotic (West Cushitic) is thus, nevertheless, a subgroup of Cushitic. Others connect Omotic with the Chadic group.

Some linguists have suggested that the Hamito-Semitic languages are related to the Indo-European languages; others have favoured the existence of a superfamily, including the Hamito-Semitic, Indo-European, Altaic, Finno-Ugric (Uralic), Kartvelian, and Dravidian languages; but most scholars regard such far-flung genetic ties as unproven and, indeed, hardly provable.

Because there has been a considerable difference of opinion as to the criteria to be applied when identifying a language as Hamito-Semitic, the basic principles of linguistic classification as applicable in this case should be stated. The only real criterion for classifying certain languages together as a family is the common origin of their most ancient vocabulary as well as of the word elements used to express grammatical relations. A common source language is revealed by a comparison of words from the supposedly related languages expressing notions common to all human cultures (and therefore not as a rule likely to have been borrowed from a group speaking another language) and also by a comparison of the inflectional forms (for tense, voice, case, or whatever). If, as a result of a step-by-step reconstruction of forms having existed at earlier periods, scholars arrive at an identical original phonological structure for each of the words or word elements compared in several different known languages, then such original forms can be ascribed to a common language, which, in the case of the languages here discussed, is conventionally termed Common Hamito-Semitic (or Proto-Hamito-Semitic). It also stands to reason that wherever one parent language has existed the daughter languages must to some degree reflect some of its grammatical characteristics.

Reconstruction of Common Hamito-Semitic

Despite the work of several scholars, only an approximate and provisional reconstruction of the parent language forms of Hamito-Semitic has so far been made. More work, however, has been done in comparing the language typologies.

COMMON CHARACTERISTICS

Certain typological features seem to have been common to all Hamito-Semitic languages at an early stage of their development. Among the phonological features are (1) a six-vowel system (*a, i, u, ā, ī, ū* that is, short and long *a,i,u*), perhaps developed from an earlier two-vowel system (of **a,* and **ə* [pronounced as the *a* in "sofa"]; an asterisk before a sound or a word-form indicates that it is not attested but is reconstructed hypothetically); (2) pharyngeal fricative consonants, indicated by the symbols ʿ (voiced) and *ḥ* (voiceless) and produced in the region of the pharynx; (3) the functioning of the glottal stop (articulated by closing the glottis, the space between the vocal cords) as a separate distinctive sound (phoneme)—this is conventionally indicated by ʾ; (4) the use of the semivowels *u̯* (*w*) and *i̯* (*y*) in the structural role of consonants; and (5) three types of consonants: voiceless, voiced, and "emphatic," the last type being phonetically realized either as voiceless consonants combined with a glottal stop, as pharyngealized voiceless or voiced consonants, or as consonants in which the air is drawn into the mouth (injective [preglottalized], or implosive), consonants in which the tongue is retracted from the usual position (velarized), or in which the tongue tip is curled upward toward the hard palate (retroflex or cerebral).

Common morphological features include (1) word bases for verbs and for nouns derived from verbs consisting of two elements that interweave with one another, a "root" consisting of consonants, and a "scheme" consisting of vowels (for examples see below); (2) a predominance of word roots consisting of three consonants over roots of two consonants; (3) a strongly developed system of infixation—*i.e.,* the insertion of elements within the root of a word to show grammatical changes and form new words with related meaning; and (4) a comparatively poorly developed system of prefixes and suffixes.

Morphological similarities among the languages

In the area of morphological typology, there are numerous similarities among the Hamito-Semitic languages, such as a system of declension of the noun and pronoun with at least three cases (nominative, genitive, accusative, with

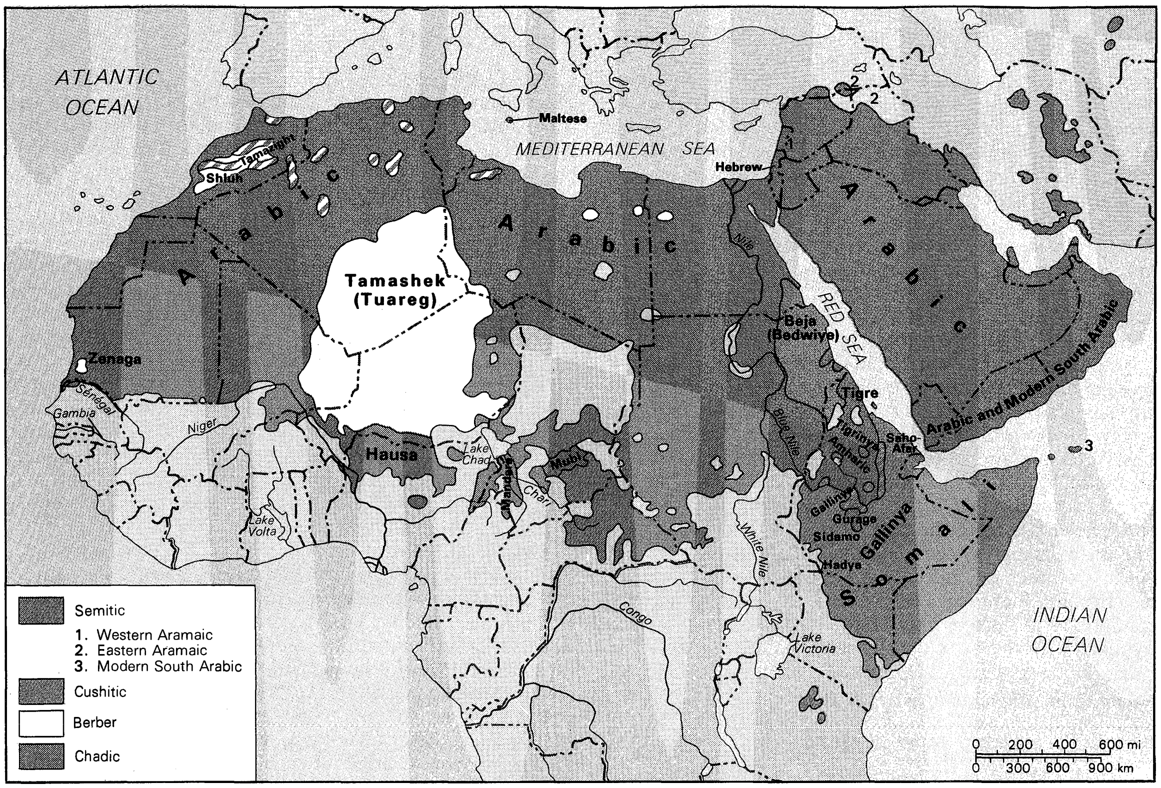


Figure 27: Distribution of the modern Hamito-Semitic languages.

traces of a still earlier system including only the agentive [ergative], and unmarked [zero] cases, or agentive, genitive, and unmarked). There are three numbers in the noun, pronoun, and verb—singular, dual, and plural. An event considered from the point of view of the resulting state, as opposed to the point of view of the action itself, is expressed by a special predicative (zero) form of the noun that later developed into a new verbal "tense." In addition, there is a well-developed binary system of verbal aspects, indicating the mode of an action (*i.e.,* punctual contrasts with durative, or perfective [completed action] contrasts with imperfective [ongoing action]), but tenses and voices of the verb remained undeveloped until the later stages. Pronominal possession markers and object markers in the form of suffixes are another common Hamito-Semitic feature, as are the prefixing of certain actor markers to the verb and a two-gender system in the noun, pronoun, and verb, perhaps developed from a still earlier system of many genders. In syntax, the Hamito-Semitic languages show certain favoured types of attributive constructions, among other common characteristics.

The above inherited Hamito-Semitic characteristics are listed, for each linguistic level, in the approximate reverse order of their stability. Languages retaining all or most of these features can be classified as belonging to the Ancient Stage of Hamito-Semitic; those that retain no less than two-thirds of the ancient consonantal system and about one-half to two-thirds of the above-listed other features belong to the Middle Stage; those that have lost more than half of these characteristics belong to the New Stage. At the New Stage, however, there are usually enough of these features still preserved to identify the language as belonging to the Hamito-Semitic family, and most of the other features can, as a rule, be reliably reconstructed for one of the former stages of its development. Moreover, the original form of the word elements that express the typical Hamito-Semitic grammatical features is usually apparent in all languages of the family. All modern Hamito-Semitic languages except Literary Arabic and Hebrew belong to the New Stage.

Hamito-Semitic verbal systems

The character of the relationship between the five branches of the Hamito-Semitic family—Semitic, Egyptian, Berber, Cushitic, and Chadic—can best be seen by comparing their systems of verbs and pronouns. There are several types of verbal systems in Hamito-Semitic, but all of them (with the exception of the Egyptian, which has developed in a quite different direction) can apparently be traced back to one single system. In this system the action (including intransitive action) is expressed by a verbal form proper, with a prefixed actor marker (singular: 1st person *'*a*-, 2nd **ta*-, 3rd **ya*-) probably deriving from a separate personal pronoun in an oblique case; the state is expressed by a form of a noun used as a predicate, plus a personal pronoun in the direct case (this is called stative). Hamito-Semitic apparently developed from a protolanguage with an ergative type of sentence construction (in which there is a special case denoting the agent of an action but no marker for the subject of a state and the direct object of an action) to a language of the nominative type (in which the subject both of an action and a state is always in the nominative case and the direct object is in the accusative case). At the same time, the predicate of state (the so-called stative) either developed into a perfective aspect (marking completion of the action of the verb) or a past tense of the verb, or it disappeared altogether. There are, however, enough traces of its existence in all branches of the family (*e.g.,* in Egyptian, in Kabyle of the Berber branch, in Sidamo of the Cushitic branch, in Mubi of the Chadic branch, and in all Semitic languages) to see that the form goes back to the parent language.

As for the verbal forms that express action and have a prefixed actor marker, there is some discrepancy of opinion. Some scholars posit for the parent language only one form. It may be, however, that there were two forms for the transitive, a perfective and an imperfective type, and possibly only one form for the intransitive type.

In several languages of the New Stage, new verbal types have developed for all aspects and tenses, particularly in the languages of the Cushitic branch (the Northern, Eastern, and Central groups, in part; and the Southern and Western [Omotic] group, always), the Chadic branch (in most languages), and the Semitic branch (typically in Neo-Syriac). These verbal forms consisted originally of a noun (for the most part, derived from a verb) plus an auxiliary verb with a prefixed actor marker. Everywhere, as a rule, the perfective aspect (or the past tense) is formed from bases of the auxiliary verb with a reduced vowel scheme in the verbal base, while the imperfective aspect (or the present/future tense) is formed from bases with a full vowel scheme (_cf._ the Akkadian perfective form *_yaprus_ "he divided," with a reduced vowel scheme, and the imperfective form *_yaparras_ "he divides," with a full vowel scheme). (There are also forms based on the participle of the auxiliary verb; _e.g._, Neo-Syriac _biktā-vövin_ "I am writing" from *_bi-ktābā-hāwē-'ănā_ "in-write-being-I.")

In that the Central Semitic verbal system (which has the imperfective with a reduced vowel scheme, as in Arabic, Hebrew, and Aramaic) is restricted to only two groups of languages inside only one branch of the entire family, it is improbable that it is this verbal system that is descended from the parent language.

Stem modification

A typical feature of the Hamito-Semitic verbal system is the existence of so-called stem modifications—_i.e._, groups of systematically related verbal stems deriving from a single root, each having its own type of semantics—that variously characterizes the action or state from the point of view of its quality, quantity, frequency, causal relations, direction, and so on. In Hebrew, for example, _šābar_ "he broke," _šibbēr_ "he broke to pieces," _hišbīr_ "he let (him) break out," and _nišbar_ "he was broken, destroyed, stranded" all are from the root _šbr_.

The pronominal systems in the different branches of Hamito-Semitic are more or less alike. Some pronouns are virtually identical everywhere; _e.g._, the possessive pronouns (2nd person masculine—"your": Semitic *_-ka_, Egyptian _-k_, Berber _-k_, reconstructed Cushitic _-ka_ or *-k^wa, Chadic [Hausa] _-ka_). Suffixed pronouns expressing the object of the verb are very similar to the possessive.

Table 48: Common Hamito-Semitic Vocabulary Items

	Semitic	Egyptian	Berber	Cushitic	Chadic
"bone"	_qōṣ_ ("thorn," Hebrew)	_qs_	_i-ghəs_	*_m-ḳḳac_	*_ḳasi_ (Hausa)
"to die"	*_mūt_	_m(w)t_	_əmmət_		_mutu_ (Hausa)
"dog"	_kal-b-_			*_kala-_ _kara-_	*_kala-_ _kara-_
"eye"		_ỉr. t_		*_yil_	*_yil_
"heart"	*_libb-_	_ỉb_		*_libb-_	
"I"	_'anāku_ _'anā_ _'anī_	_ỉnk_ _anok_ (Coptic)	_n(ə)ki_	*_(')ani_ *_ta-(')ani_ _anu_ _ana_	_an_ (Sura) _nī_ (Hausa) _n-ani_ (Kanakura)
"jackal" ("wolf," "dog")		_wnš_	_uššən_	*_waẑẑ_ *(from _wanẑ-_?) (Highland Eastern Cushitic)	
"man"	*_mut-_	*_mt_			_mito_ (Jegu) _miji_ (Hausa)
"name"	*_šim-_		_i-səm_	*_sim-_	*_sim(m)-_ _summ-_ _sūn-_
"thou"	_'anta_ _'atta_	_ntk_ _əntok_ (Coptic)		*_'atta_	
"tongue"	*_laš-ān-_	_ns_ (pronounced _las_)	_i-ls_		*_ha-ls(e)_ (Hausa)
"tooth"	*_šinn-_	_sn_ ("harpoon")	_-sin_		*_sinn-_(?)
"two"	*_t͟hin-_	_sn_	_sin_	*_čan-_("two equal parts")	
"water"	*_mā'-_ _may-_	_m(y)-w_	_a-ma-n_	_mā-n_ ("sea") (Somali)	

*Reconstructed form.

Linguistic substrata

The diverging of the branches and the individual languages of the Hamito-Semitic family from the common ancestral language, although mainly explained by the internal development of the languages after loss of contact, also results to a great extent from the influence of different linguistic substrata. Thus, the ancient Hamito-Semitic language had in many cases probably spread to originally alien populations. This view is supported by the different racial types of the speakers. In some cases the substratum language (_i.e._, that of the original population) can be identified —_e.g._, Sumerian, Hurrian, and others for North Semitic; Nilo-Saharan and East Sudanese for Cushitic; East Sudanese and possibly some others for Chadic. The least substratum influence seems to have been experienced by the Berber branch.

SEMITIC LANGUAGES

Languages of the group. The Semitic languages can be subdivided into four groups: the Northern Peripheral, the Northern Central, the Southern Central, and the Southern Peripheral.

Akkadian

Northern Peripheral Semitic. The Northern Peripheral group, from the Ancient to Middle Stage, includes Akkadian with its dialects of Babylonian and Assyrian, spoken in Mesopotamia from about 3200 BC to the beginning of the Christian Era. Typical features are stative verb forms conjugated with suffixes and two verbal forms with a prefixed actor marker for the imperfective and perfective (with full and reduced vowel schemes, respectively; later a new "perfect" with an infixed _-ta-_ in the stem developed). Originally there were five cases of the noun, plus an unmarked form for the nominal predicate and the noun without grammatical relations. Later three cases remained but were lost in the 1st millennium BC. Loss of the _gh, ḥ, ʿ,_ and _h_ sounds occurred from _c._ 2000 BC. The vowels were _a, e, i, u_ (both long and short).

Hebrew

Aramaic dialects

Northern Central Semitic. The Northern Central Semitic group includes the Canaanite, Ugaritic, and Amorite languages of the Ancient Stage, which were spoken in Palestine, Phoenicia, Syria, and Mesopotamia from the 3rd to the 2nd millennium BC. To the Middle Stage belongs Phoenician-Punic, spoken in Phoenicia, on islands of the Mediterranean, and in North Africa, from the 2nd millennium BC to the 1st millennium AD. Also to the Middle Stage belong Hebrew, Moabite, Ya'ūdī, and Old Aramaic. Hebrew, originally spoken in Palestine from the 13th century BC to the 2nd century AD, later spread all over the world as a written language. At present there are about 2,600,000 Hebrew speakers in Israel. Moabite and its kindred dialects in the Transjordan were alive in the 1st millennium BC but are now extinct. Ya'ūdī, spoken in northern Syria in the 9th century BC, is also extinct. Old Aramaic, from Syria and Mesopotamia, existed from the 14th century BC(?) through the 15th century AD. Its oldest written texts date from the 9th century BC. The dialects of Aramaic include Ancient Aramaic proper; Imperial Aramaic (the official language of Assyria and Achaemenid Persia, including also Biblical Aramaic, or Chaldean); Western Aramaic, with Palmyrenean, Nabataean, Palestinian, Galilean, and other varieties; Eastern Aramaic, including Syriac (Edessan, with subdialects), Babylonian Talmudic, and Mandaic. Most Aramaic dialects gave way to Arabic beginning with the 7th century AD.

The New Stage of Northern Central Semitic is represented by New West Aramaic, or Maʿlūla, in Syria, with a small number of speakers, and Neo-Syriac, or "Assyrian," in Iraq (al-Mawṣil), Turkey (Ṭūr-ʿAbdīn), Iran (Urmia), the Soviet Union, and the United States, with about 200,000 speakers.

Typical features of the Northern Central Semitic group are the perfective aspect with suffix conjugation and the imperfective aspect with prefix conjugation and stems with a reduced vowel scheme. (An entirely new verbal system developed in Neo-Aramaic.) The group is also characterized by the article _ha-_ (prefixed, in Hebrew and Punic) or _-ā_ (suffixed, in Aramaic). The system of declension was lost from the Middle Stage on. In this group the number of vowel qualities increased beyond just _a, i, u,_ while the number of consonants diminished considerably from the Middle Stage on. The sounds _p, t, k, b, d, g_ became aspirated after vowels—_i.e._, pronounced with an accompanying puff of breath (and are now pronounced as _f, t_ or _s, k͟h, v, d, g_ in Modern Hebrew).

Arabic

Southern Central Semitic. The Southern Central group includes Classical, or Literary, Arabic belonging to the Ancient and Middle stages. Originally spoken in Arabia, Classical, or Literary, Arabic is now found from the Indian to the Atlantic Ocean and has been attested from the 5th century BC to the present time. From the New Stage come the modern Arabic dialects, some of them mutually unintelligible. They have about 130,000,000 speakers all over northern Africa, on the Arabian Peninsula, in Jordan, Israel, Lebanon, Syria, and Iraq, and in some districts of Turkey, Iran, and the Soviet Union. Maltese, on the island of Malta, has developed into a separate language, spoken by about 300,000. A typical feature of the group is a verbal system that is very similar to that of the Northern Central group (with minor differences) but that developed tenses instead of aspects from the late Middle Stage (*e.g.,* in the Egyptian dialect, from prepositional constructions). There were three cases of the noun, but declension was lost at the late Middle and New Stage. Also characteristic of South Central Semitic are the article *al-* and a strongly developed system of internal inflection with the plural mostly of the *pluralis fractus* type ("broken plural," in which plurality is shown by means of internal vowel changes). The Proto-Semitic phonological system has been on the whole well preserved, but **ś* has become *s*, **ṯh* has become *ẓ,* and other similar changes.

Southern Peripheral Semitic. To the Southern Peripheral group of the Ancient to Middle Stage belong the South Arabian dialects, Sabaean (*cf.* Sheba), Minaean, Qatabānian, and Ḥaḍramawtian, spoken from the 1st millennium BC to the 1st millennium AD. The Middle Stage is represented by Geʿez (Geez or Gəʿəz), or Ethiopic, found in northern Ethiopia in the 1st millennium AD; and the New Stage by the South Arabian group, including Mahrī, Shaḥrī (Eḥkalī), Ḥarsūsī, and Baṭḥarī on the Arabian shore of the Indian Ocean, and Suquṭrī (possibly a dialect of Mahrī) on the island of Socotra, with the total number of speakers probably being around 50,000. Also of the New Stage is the Ethiopic group, consisting of three subgroups: North Ethiopic, Central Ethiopic, and the Gurage subgroup. North Ethiopic, nearest to Geʿez, includes Tigrinya (Tigrai) and Tigre, spoken in northern Ethiopia and Eritrea by 4,500,000 speakers. Within the Central Ethiopic group (10,000,000 speakers) are Amharic, the official language of Ethiopia, the near-extinct Argobba language, and the entirely extinct Gafat. The Gurage cluster of languages is found south and east of Addis Ababa and has 650,000 speakers. In all, there are somewhat more than 15,000,000 speakers of Southern Peripheral Semitic languages.

Ethiopic group of languages

Typical features of the group include traces of two types of verbal forms with prefixed actor marker (one type with full vowels and the other with reduced vowel schemes). In other respects the verbal system is as in South and North Central Semitic; considerable innovations, however, have developed at the New Stage, especially in the Ethiopic group (with a Cushitic substratum). Declension was lost from the Middle Stage. Phonetic development is as in Arabic, but more of the ancient consonants were lost. The Ethiopic group has lost most of the pharyngeal and laryngeal consonants.

Table 49: The Arabic Alphabet and Numerals

consonants					equivalents		approximate pronunciation, classical Arabic
alone	initial	medial	final	name	EB preferred	alternatives	
ا	ا	ـا	ـا	alif	*		*
ب	بـ	ـبـ	ـب	bāʾ	b		*b*a*b*y
ت	تـ	ـتـ	ـت	tāʾ	t		*t*ie†
ث	ثـ	ـثـ	ـث	thāʾ	th	t͟h	*th*in
ج	جـ	ـجـ	ـج	jīm	j	d͟j	*j*ob
ح	حـ	ـحـ	ـح	ḥāʾ	h	ḥ	‡
خ	خـ	ـخـ	ـخ	khāʾ	kh	k͟h	Ger. Bu*ch*§
د	د	ـد	ـد	dāl	d		*d*i*d*†
ذ	ذ	ـذ	ـذ	dhāl	dh	d͟h	*th*en
ر	ر	ـر	ـر	rāʾ	r		e*rr*or (trilled)
ز	ز	ـز	ـز	zāʾ	z		*z*one
س	سـ	ـسـ	ـس	sīn	s		*s*and
ش	شـ	ـشـ	ـش	shīn	sh	s͟h, š	*sh*y
ص	صـ	ـصـ	ـص	ṣād	ṣ	ṣ	
ض	ضـ	ـضـ	ـض	ḍād	ḍ	ḍ	‖
ط	طـ	ـطـ	ـط	ṭāʾ	ṭ	ṭ	
ظ	ظـ	ـظـ	ـظ	ẓāʾ	ẓ	ẓ	
ع	عـ	ـعـ	ـع	ʿayn	ʿ		¶
غ	غـ	ـغـ	ـغ	ghayn	gh	g͟h	Fr. *r*ien
ف	فـ	ـفـ	ـف	fāʾ	f		*f*i*f*ty
ق	قـ	ـقـ	ـق	qāf	q	ḳ	♀
ك	كـ	ـكـ	ـك	kāf	k	k	*k*in
ل	لـ	ـلـ	ـل	lām	l		*lil*y†
م	مـ	ـمـ	ـم	mīm	m		*maim*
ن	نـ	ـنـ	ـن	nūn	n		*n*o†
ه	هـ	ـهـ	ـه	hāʾ	h		*h*at
و	و	ـو	ـو	wāw	w		*w*atch♁
ى	يـ	ـيـ	ـى	yāʾ	y		*y*et □
ء				hamzah	initial, omit; medial and final,ʾ		*

vowels, diphthongs, and special diacritical marks		equivalents		approximate pronunciation, classical Arabic
letter	name	EB preferred	alternatives	
ـَ	fatḥah	a	e	*a*t
ـُ	ḍammah	u		f*oo*t
ـِ	kasrah	i		*i*f
ـَا	long fatḥah (alif)	ā		*a*dd, f*a*ther◇
ـُو	long ḍammah (wāw)	ū		f*oo*d
ـِي، ى	long kasrah (yāʾ)	ī		*e*ve
ـَوْ	fatḥah wāw sukūn	aw		*ou*t
ـَىْ	fatḥah yāʾ sukūn	ay	ai, ei	*i*ce
ى	alif maqṣūrah	ā	á	*a*dd, f*a*ther◇
ة	tāʾ marbūṭah	-ah or -at	-a or -at	▲
ٱ	hamzat al-waṣl	restore a	ʾ	✚
آ	alif maddah	ā	a	*a*dd, f*a*ther◇
يَّ ، ـيَّ	yāʾ shaddah followed by short vowel	īy-	iyy-	*e*ve⊕
وَّ	wāw shaddah followed by short vowel	ūw-	uww-	f*oo*d⊕

numerals					
Arabic	westernized Arabic	Arabic	westernized Arabic	Arabic	westernized Arabic
١	1	١٢	12	٢٣	23
٢	2	١٣	13	٢٤	24
٣	3	١٤	14	٢٥	25
٤	4	١٥	15	٢٦	26
٥	5	١٦	16	٢٧	27
٦	6	١٧	17	٢٨	28
٧	7	١٨	18	٢٩	29
٨	8	١٩	19	٣٠	30
٩	9	٢٠	20	١٠٠	100
١٠	10	٢١	21	١٠٠٠	1,000
١١	11	٢٢	22		

**Alif* has no consonantal value of its own; in transliteration, omit at the beginning of a word. *Hamzah* is pronounced as a glottal stop, as in Cockney or New York "bottle." As a vowel, *alif* is pronounced *a*dd. †Pronounced dentally. ‡A pharyngeal fricative. §A velar fricative. ‖It is impossible by English examples to indicate the difference from pronunciation of *ṣ, ḍ, ṭ,* and *ẓ*. The back of the tongue is raised and the pharynx is constricted (velarization), in addition to the regular articulation of these consonants. ¶A contraction of the throat (a pharyngealized vowel that is considered a consonant in Arabic). ♀A uvular stop; a *k* sound produced farther back in the throat than any English *k*. ♁As a consonant, *w*atch; as a vowel, f*oo*d. □As a consonant, *y*et; as a vowel, *e*ve. ◇ Pronunciation varies from place to place and also depending on the accompanying consonant. ▲Pronounced as *t* in certain grammatical constructions; otherwise pronounced as silent *h*. ✚Initial vowel elision; that is, in pronunciation the *a* is simply omitted and the preceding vowel is pronounced with the following consonant. ⊕A long vowel sound followed by appropriate consonant sound. See also *footnotes* ♁ *and* □.

Table 50: The Hebrew Alphabet

consonants			numerical value†	transliteration		approximate Israeli Sefardic pronunciation
printed*	written*	name		EB preferred	alternatives	
א	א	alef	1	ʾ; omit at beginning of word		glottal stop or silent
בּ	בּ	bet	2	b		*b*oy
ב	ב	vet		v	bh, b̲	*v*end
ג	ג	gimel	3	g		*g*irl
ד	ד	dalet	4	d		*d*ove
ה	ה	he	5	h; omit at end of word unless written with dot		*h*ow; silent at end of word
ו	ו	waw (vav)	6	w‡	v	*v*end
ז	ז	zayin	7	z		*z*ebra
ח	ח	ḥet	8	ḥ	h̲, ḫ, ch	Ger. Bu*ch*
ט	ט	ṭet	9	ṭ	t	*t*oy
י	י	yod	10	y§		*y*et
כּ	כּ	kaf	20	k		*k*ey
כ (ךְ)	כ (ךְ)	khaf		kh		Ger. Bu*ch*
ל	ל	lamed	30	l		*l*eg
מ (ם)	מ (ם)	mem	40	m		*m*ember
נ (ן)	נ (ן)	nun	50	n		*n*ow
ס	ס	samekh	60	s		*s*o
ע	ע	ʿayin	70	ʿ		glottal stop‖
פּ	פּ	pe	80	p		*p*aper
פ (ף)	פ (ף)	fe		f	ph	*f*an
צ (ץ)	צ (ץ)	tzade	90	tz	z̲, z, ẓ, ṣ, ts	pe*ts*
ק	ק	qof	100	q	k, ḳ	*k*ey
ר	ר	resh	200	r		Fr. *r*ien: or trilled *r*
שׁ	שׁ	shin	300	sh	š	*sh*oe
שׂ	שׂ	sin		s	ś	*s*o
תּ	תּ	taw	400	t		*t*oy
ת	ת			t	th	*t*oy (Ashkenazi *s*o)

vowels		transliteration		approximate Israeli Sefardic pronunciation
sign¶	name	EB preferred	alter-natives	
◌ַ	pataḥ	a		*f*a*ther*
◌ֲ	ḥatef pataḥ	a	ă	
◌ָ	qametz gadol ♀	a	ā	
◌ֶ	segol	e	ä	p*e*t
◌ֱ	ḥatef segol	e	ĕ	
◌ֵ	tzere ḥaser	e	ē	
◌ֵי	tzere male	e	ê	th*e*y
◌ְ	shewa naʿ	e	ᵉ(above line) (silent)	

vowels		transliteration		approximate Israeli Sefardic pronunciation
sign	name	EB preferred	alter-natives	
◌ִ	ḥireq ḥaser	i		f*ee*t
◌ִי	ḥireq male	i	î	
◌ֳ	ḥaṭef qametz	o	ŏ	c*o*rd
◌ָ	qametz qaṭan	o		
◌ֹ	ḥolem ḥaser	o	ō	
וֹ	ḥolem male	o	ô	
וּ	shureq	u	û	s*oo*n
◌ֻ	qibbutz	u		

*Final forms in parentheses. †Hebrew has a "ciphered" numeral system in which the letters of the alphabet have numerical value and are used as numbers. For example, numbers 11 through 19 are written with the letter *yod* (10) plus the letter for 1, 2, etc. (except 15 which is written *ṭet* [9] plus *waw* [6], and 16 which is written *ṭet* [9] plus *zayin* [7]); the 20s are combinations of *kaf* (20) plus the letter for 1, 2, etc; 500 is written *taw* (400) plus *qof* (100). ‡Functions as both a consonant and a vowel. See *ḥolem male* and *shureq* in "vowel signs" table above. §Functions as both a consonant and a vowel. See *tzere male* and *ḥiriq male* in "vowel signs" table above. ‖ In formal pronunciation, pronounced as a pharyngeal voiced fricative sound. ¶The long horizontal line represents the consonant; vowel signs are placed above, below, or to the left of it, as shown. ♀In Ashkenazi pronounced as in "ball." *Note:* In transliteration, a letter is usually doubled in medial position to indicate a strong *dagesh* (*dagesh* is indicated in Hebrew by a point in the middle of a letter). *Pataḥ* under final *he, ḥet,* and *ʿayin* is pronounced before the consonant under which it is written.

Historical and cultural background. Glottochronological methods, which attempt to measure degrees of differences between related languages by comparing a list of basic vocabulary items, indicate that the first group to separate from the Common Semitic ancestral language was Akkadian (Northern Peripheral group, *c.* 3300 BC) and the second was the Southern Peripheral group (second half of 3rd millennium BC). The Northern Central group had contacts for a long time with the Southern Central languages, and linguistic division within the North Central group is dated at the beginning or middle of the 2nd millennium BC. The relative position of Arabic to the other Semitic languages is not quite clear, probably because of its uninterrupted contacts with Aramaic and other nomadic Semitic groups for many centuries.

Semitic writing forms

The oldest of the attested Semitic languages, Akkadian, was the vehicle of a great ancient literature written in a logosyllabic cuneiform writing system of Sumerian origin. Records of other ancient Semitic languages exist in various forms. Amorite, another ancient Semitic language, is known from proper names; Ugaritic was written in a quasi-alphabetic cuneiform script unconnected with the Akkadian. The Canaanites of Phoenicia used a still undeciphered syllabic script, the Proto-Byblian, in the 2nd millennium BC, while those of Palestine and the Sinai Peninsula employed another undeciphered writing, the Sinaitic script, which may be alphabetic in nature. All the other Semites used and, for the most part, still use consonantal quasi-alphabets with no means or only imperfect means to distinguish the vowels. All such alphabets—of which the more important are the Hebrew, the Syriac, and the Arabic—are descended from the Phoenician linear quasi-alphabet of 22 signs, first attested at Byblos and externally similar to the Proto-Byblian script. (All the European alphabets are descendants of the Phoenician, and all the Asiatic alphabets are descendants of the Aramaic variants of the Phoenician.) From a South Arabian variant of the earliest Semitic alphabet the Ethiopians developed a syllabic writing still in use for the languages of Ethiopia. Maltese uses the Latin alphabet.

Two of the later Semitic languages, Hebrew and Arabic, have been the languages of great religions, Judaism and Islām. The religious significance of Hebrew explains why the language, although already partly replaced by an Aramaic vernacular in the everyday life of Palestine in the late 1st millennium BC and early 1st millennium AD, was still preserved as a literary language by the Jews who were expelled from Palestine by the Hellenistic kings and the Romans between the 3rd century BC and the 2nd century AD. It has been revived in a modernized form as a spoken and written language in Israel. Classical Arabic has been preserved as a literary language since the Arabic conquest of North Africa and the Near and Middle East in the 7th and 8th centuries AD. The language was used for literary purposes by Muslims of different nations all through the Middle Ages and is still used as a language of the school and the administration and as the spoken language of the educated in all Arabic countries, although the vernacular New Stage Arabic idioms are to a great extent mutually unintelligible. There is a great amount of literature—scholarly, religious, scientific, and fiction—both in Hebrew and in Arabic. Of the other Semitic languages, Syriac was and is the language of certain Eastern Christian sects and was the means by which the Greek tradition was passed on to the Arabs. Another Christian sect, that of the Monophysites of Ethiopia, used Geʿez (Ethiopic) and still retains it in ecclesiastical use, but the literary and other secular remains are less important.

Common Semitic sounds

Linguistic characteristics. The Semitic branch of this language family is characterized by several general features.

Phonology. In phonology, the emphatic stop consonants *ṭ* and *q* (from **ḳ*) were retained but not **ṗ*. The affricates of the parent language (which are begun as stops and released as fricative sounds), if they ever existed, were lost or replaced by sibilant and interdental sounds (which are symbolized as *s, ṣ, z; th, ṭh, dh*); the lateral sounds and the interdentals were subsequently lost in most languages. The labialized velar sounds (except in the Ethiopic group) and all postvelar stops were lost. As for the glottal, pharyngeal, and laryngeal consonants, six of them (*g̲h̲*, *k̲h̲*, ʿ, *ḥ*, ʾ, *h*) are retained in Arabic and were retained in the other Semitic languages at the Ancient Stage. Hebrew and Aramaic retain ʿ, *ḥ*, ʾ, and *h* (but only *k̲h̲*, and *h* in Modern Hebrew and in most New Aramaic dialects); later Ethiopic and Punic retained only ʾ and *h*, and Akkadian only *k̲h̲*, and ʾ (but *a* became *e* in words that formerly included *g̲h̲*, ʿ, or *ḥ*). The original six-vowel system changed everywhere as early as the Middle Stage; Arabic preserved it the longest.

Word formation. Word formation is achieved by an intricate system of vowel infixation, sometimes accompanied by a few suffixes or prefixes. Each pattern of infixation ("scheme"), in combination with a consonantal root, plus the affixes, has its own peculiar type of meanings. The Arabic noun *ma-KTaB-*, for example, means "place of writing, school," and *KaTTāB-* means "writer, scribe"; *KāTiB-*, a participle, means "writer, [the one] writing"; *ya-KTuB-u*, the imperfective form, is "he writes"; *yu-KaTTiB-u*, another imperfective, is "he writes, he teaches to write"; and *KaTaBa*, the perfective, means "he wrote." (The capital letters indicate the sounds of the consonantal root.) The need to correlate these diverse patterns with the basic meaning of the root resulted in the absence of important positional changes in the Semitic consonant sounds as well as in the comparative scarcity of borrowed terms, especially of verbs. The primary nouns, those not derived from verbal forms, are not included in this system of patterns, except, by analogy, in Arabic and the other Southern languages.

Morphology. In regard to morphology, the masculine gender marker is zero (*i.e.*, it has no structural marker), but traces of a *-u* can be detected; the feminine gender marker is *-a* or *-ā* or more usually *-(a-)t-*, although *-t-* belonged originally to another series of gender markers (in which there were more than two genders). The declension of the noun and pronoun was retained in the Ancient Stage of Semitic, with nominative, genitive, accusative, dative-locative, and locative-adverbial cases. The dative-locative ending was lost in Arabic, and the locative-adverbial form appeared only in Akkadian. There are traces of an earlier suffixed definite article, *-m*(*a*) or *-n*(*a*), retained in Arabic as the marker of the indefinite form of the noun. Later new definite articles developed. The dual number was expressed at the Ancient Stage but became lost in the later stages. The plural of the noun is formed in North Semitic by lengthening the singular form. This means of expressing the plural exists also in South Semitic, but here it is to a great extent replaced by the so-called *pluralis fractus* "broken plural" (*e.g.*, Arabic *kalb-* "dog," *kilāb-* "dogs").

Dual and plural forms

The West Semitic languages (except, in part, for Ethiopic and the South Arabian dialects) have lost the old imperfective form with a full vowel scheme and have replaced it by the form next in frequency, namely, the subjunctive mood (*ya-qtul-u*) with a reduced vowel scheme. The stative verb form, still preserved in Akkadian, developed into a new perfective (Arabic *qatala* "he killed," *mariḍa* "he was ill"), leaving the form **ya-qtul*, which was originally the perfective and jussive (a form expressing a wish or command), for the jussive only.

Syntax. Typical of the Semitic languages are attributive constructions: (1) a construction in which the governing noun appears in a shortened form before the governed noun in the genitive, as well as (2) a construction in which the two nouns, each in its complete case form, are connected by a pronoun (*e.g.*, Old Akkadian *t̲h̲u*, Aramaic *d̲h̲ī*).

Vocabulary. As mentioned above, the system of word formation in Semitic does not favour borrowings, especially verbal ones. There are, however, a number of nouns borrowed from Sumerian in Akkadian; from Akkadian, Iranian, and Greek in Aramaic; from Persian and Turkic in Arabic; and from the Agau and other Cushitic languages in the Semitic languages of Ethiopia.

EGYPTIAN

The Egyptian branch of the Hamito-Semitic family includes only one language (with local dialects), namely, Egyptian. It can be differentiated into several stages. The

Ancient Stage, Old Egyptian, extended from before 3000 to *c.* 2200 BC; the transition from the Ancient to the Middle Stage, Middle Egyptian, lasted from *c.* 2200 to *c.* 1600 BC and, as a dead literary language, until *c.* 500 BC. The Middle Stage, Late Egyptian (also called Neo-Egyptian), is dated from *c.* 1550 to after *c.* 700 BC and the Demotic language between *c.* 700 BC and some time after AD 400. Finally, the New Stage, called Coptic, began in about the 2nd century AD and lasted at least to the 17th century and possibly, in some villages, until the 19th century. Thus, five literary dialects are differentiated. All these language periods refer to the written language, which often differed greatly from the spoken dialects. Coptic is still in ecclesiastical use (along with Arabic) among the Arabic-speaking Monophysite Christians of Egypt.

Characteristic Egyptian sound changes

Phonology. The Egyptian hieroglyphic writing was not adapted for expressing vowels. By the Coptic period, when an alphabetic writing came into use, the Egyptian vowel system had undergone so radical a change that the original vowels can be reconstructed only very approximately. In the consonantal system the loss of the emphatics (except *p* from **ṗ* and *q* from **ḳ*) is characteristic, as are the changes of **-r* (at end of syllable) to -ʾ, **li-* and **lu-* to *ỉ-*, **ki-* and **ku-* to *ṯ* (pronounced as *tch*), **gi-* and **gu-* to *ḏ* (pronounced *dj*). In some cases *ṯ* and *ḏ* apparently reflect the affricates of the parent language. In addition, the original lateral sounds were lost as well as the postvelar stops and labialized velars, and the system of spirants was simplified. Beginning with Middle Egyptian, *d, ḏ,* and *ṯ* developed gradually to *t,* and many final consonants (*e.g.,* *-t, -r*) were dropped.

Word formation, morphology, and syntax. Word formation in Egyptian was similar to the Semitic type, although probably less consistent. As for the inflection, there may have been only two cases of the noun in Old Egyptian. The actor case coincided with the genitive, and this may have been responsible for a drastic rearrangement of the entire verbal system. Of the original Hamito-Semitic verbal forms only the stative ("pseudo-participle") is preserved; its subject, when a pronoun, is in the ancient direct (zero) case. Verbal forms expressing action were replaced by attributive and prepositional constructions, with the person of the actor being expressed by a suffixed possessive pronoun or by a noun in the genitive. Stem modifications are less developed in Egyptian than in Semitic; a habitative form with reduplication (repetition) of the third consonant of the root exists along with the normal imperfective of the main stem. In the later periods a new complicated system of secondary verbal tenses developed. Masculine gender was marked by zero (the absence of any ending) or **-aw-*, feminine gender by *-at-*, plural masculine by **-ā-w-*, and plural feminine probably by **-ā-w-āt-*.

Typical of Egyptian syntax are a construction in which two nouns, each in its complete case form, are connected by a pronoun, called the *nota genitivi;* and a *status constructus,* in which the governing noun appears in a shortened form before the governed noun in the genitive.

Hieroglyphic, hieratic, and demotic writing

Writing. The ancient Egyptian writing was a logosyllabic one, having symbols representing either complete words or syllables of words; identical signs were used for syllables with identical consonants but different vowels. According to the external form of the signs, the writing is classified as hieroglyphic when it is found on inscriptions on stone, metal, and other hard surfaces and as hieratic and later demotic when it is used for cursive writing on papyrus manuscripts. Typologically the three forms of writing are identical. Coptic was written in an alphabet based on Greek and partly on Demotic. There is a considerable literature in Egyptian and in Coptic (in the latter, mostly of a religious nature; see also WRITING: *Hieroglyphic writing*).

BERBER LANGUAGES

The Berber (Berbero-Libyan) branch is represented by a multitude of New Stage Berber dialects distributed all over North Africa, from the Siwa Oasis in the Arab Republic of Egypt to Senegal (about 11,000,000 speakers). The more important dialect clusters are Tamashek (Tuareg), in the central Sahara and south of the Niger; Shawia and Kabyle (Zouaouah), both in Algeria; Rif and Tamazight, predominantly in Morocco; Shluh (Tashelḥayt or Shilha), in Morocco and Mauritania; and Zenaga, in Senegal. Little is known of ancient Libyan, also called Numidian. It is attested by inscriptions found in Tunisia, Algeria, and elsewhere, dating from the times of the Roman Empire and written in a native consonantal quasi-alphabetic script still surviving in a modified form among the Tuaregs of Sahara. Whether the extinct language of the Guanches in the Canary Islands and of the Iberians of Spain belonged to the Berber branch or even to Hamito-Semitic is doubtful.

Ancient Numidian

Phonology. In the phonologies of these languages the vowels **a, *i, *u* were lost or reduced to *ə*, and **ā, *ī, *ū* became *a, i, u;* **w* and **y* may appear both as consonants and as vowels, and the emphatics are represented by *ḍ, gh* (but in reduplication *ṭṭ, qq*), and *ẓ.* The system of spirants has been simplified but retains *š* (*sh*) and *ž* (*zh*) sounds. Interdentals, laterals, and affricates were lost.

Word formation and morphology. Except in the verb, there are only traces of the internal inflection type of word formation characteristic of the Semitic branch. Among grammatical features, a former article no longer retaining its determinative function (masculine **hā-*, plural **hī,* feminine **tā-,* plural **tī-*) is prefixed under certain conditions to the noun, displacing the prefixed markers of gender, *w-* and *t-*. These latter gender markers are at present used in a form of the noun as an attribute or as a subject of a verb when following the predicate in the sentence. The plural of the noun is masculine *-ən* and *-an* and feminine *-in.* A *pluralis fractus* "broken plural" has also developed (mostly an infixation of *-a-*). The perfective of the main verbal stem also has a habitative form (reduplication of the second consonant of the root, or prefixation of *-tt-* to the word base). Tamashek has several verbal tenses.

There is little or no intelligibility between the dialects, except for historically neighbouring ones. A great number of Arabic borrowings are evident in most dialects; there are also a number of borrowings from Punic, Latin, and from the languages south of Sahara.

CUSHITIC LANGUAGES

The Cushitic branch goes back to a reconstructed Common Cushitic parent language; this, according to the Soviet scholar A.B. Dolgopolsky, was the dialect of Common Hamito-Semitic that best preserved the original phonological system. Whether or not West Cushitic (Omotic) is a descendant of Common Cushitic is not clear. The Cushitic languages are all from the New Stage and have about 16,000,000 speakers.

The five Cushitic language groups

Languages of the group. The Cushitic languages, including the West Cushitic group, can be subdivided into five groups. The Northern group is represented by Beja, or Bedawiye, spoken mainly in The Sudan close to the Red Sea and also in Eritrea; it has about 1,300,000 speakers. Typical linguistic features include the scanty representation of affricates, velars changing partly to ʾ, and postvelar consonants changing to *h.* Two or, in some cases, three verbal forms with prefixed actor markers exist ("strong conjugation"), but many verbs are conjugated by suffixes (developed from an auxiliary verb with prefix conjugation). There are stem modifications similar to those of Berber in the strong conjugation, formed by suffixes in the other verbs (this is also typical of the other Cushitic languages). In addition, declension of the noun, with traceable relics of the ancient type similar to the Semitic, also can be seen.

The Eastern group has several subgroups. The highland languages, spoken east of Addis Ababa, include Hadya-Libide (900,000 speakers), Kambata (300,000 speakers), Sidamo (1,100,000 speakers), Darasa (300,000 speakers), Burji, and some related languages. The total number of speakers of this subgroup is about 2,600,000. The other subgroups include Saho-Afar in Eritrea, northeast Ethiopia, and Djibouti, with 750,000 or more speakers; the Somali subgroup, with Somali, Bayso, Rendile, and other languages in Somalia, eastern Ethiopia, and eastern Kenya, having a total of more than 5,000,000 speakers; the Gallinya subgroup, comprising Oromo (Gallinya, with several dialects) in western, central, and southern Ethiopia and northern and eastern Kenya; the subgroup of Arbore, Dathanaic (Geleba), and other languages, together having about 13,000,000 speakers; Konso, Gidole, and related dialects, with about 80,000 speakers in western Ethiopia; Warazi (Warize) and related languages also in western Ethiopia, with about 50,000 speakers; and Mogogodo of northern Kenya. Typical features of the group are the presence of emphatic affricate sounds and the change of postvelar sounds to ʾ and *h;* in some languages the older **l* sound is represented as *j* or *r,* and **r* as *r, d,* or *n.* The number of verbs of the "strong conjugation" is very small in some languages and nonexistent in others. In addition, there are grammatical genders differing from the ancient type.

The Central, or Agau, group includes languages or dialects dispersed over Ethiopia. They include Bilin, Khamta, Awngi, and Kemant (Qimant), among others, and are spoken by more than 100,000 people. The Quara dialect, spoken formerly by the Falashas, an Ethiopian Jewish ethnic group, is now extinct. Although the vocabulary of all Agau dialects is very similar, there is little mutual intelligibility as a result of the dissimilarity in the phonetic reflexes of the Proto-Cushitic sounds and the strong influence of Ethiopic and Amharic.

The Southern group, located in Tanzania, south of the Equator, includes Iraqw and its related dialects, Asa, Ngomwia, and others. Characteristic of the group is the loss, for the most part, of emphatic consonants. The laterals, however, are partly preserved, as are the pharyngeal ʿ and a few of the affricates. Both **l* and **r* are reflected as *l-* and *-r-*. In spite of numerous innovations as a result of substratum influence, there are many similarities with Eastern Cushitic in grammar.

The Omotic languages

The Western group, also called the Omotic branch by some scholars, encompasses Ometo, a dialect cluster including the Walamo language, with about 1,600,000 speakers; Janjero (Yamma), Bworo (Shinasha, Gonga), Anfillo (Southern Mao), Benesho-She (Gimira), Ari-Banna, and others, all of which are languages with small numbers of speakers, perhaps about 120,000 in all; and Kafa (Kaficho)-Mocha, with more than 200,000 speakers. All of these languages are spoken along the western border of Ethiopia and in northern Kenya. Typical features include the change of **s* to *š,* the preservation of most affricates, but the loss of laterals and of all postvelar, pharyngeal, and laryngeal fricatives. The sonants **l,* **r,* and **n* are usually represented alike as *n-* and *-r-*. Some languages have tones that serve to differentiate word meaning. Also characteristic are drastic innovations in the pronoun and the verb. Traces of the genders are usually represented as masculine *-ō* (from **-aw*) and feminine *-ā* and *-ē* (perhaps from **-at* or **-ay*).

There have been some attempts to create a written language for Oromo and especially for Somali on the basis of Ethiopic, Latin, or Arabic writing. An original Somali writing system was invented in the beginning of the 20th century, but at present Somali is written in the Arabic alphabet.

Linguistic characteristics. *Phonology.* All the Hamito-Semitic groups of consonants were preserved in Common Cushitic, and separate reflexes of each group can be traced in the different Cushitic languages. Because of an imperfect development of the system of word formation by vowel infixation, however, the stability of the consonantal root was not as necessary for correlation of forms as in Semitic. The reflexes of the sounds of the protolanguage in the individual Cushitic languages therefore depend to a great extent on positional circumstances; thus, a Proto-Cushitic **c,* pronounced as *ts,* may have developed into a *d-* in an initial position and an *-s-* in an intervocalic or final position, and so forth. Emphatics are mostly preserved (*ḍ, ç* or *č̣, ḳ,* etc.); **ṗ* is distinguished from **p* by different reflexes (Omotic partly retains *ṗ*). Affricates (and also *d, š, s,* etc.) represent what in Semitic are interdental consonants.

Morphology. Verbal conjugation by means of prefixed actor markers is preserved only in a part of the verbs or else in traces; in most of the verbs it is replaced by a new system of conjugation (originally a combination of verbal noun plus prefix conjugation of an auxiliary verb). Two genders (masculine **-w,* feminine **-t*) and traces of noun declension can be observed; partial and sometimes complete reduplication of stems is used as a means of expressing the plural, along with the means known from the other branches. The pronominal system (except in Omotic) is very close to that of Semitic. In vocabulary, there are many borrowings from Ethiopic, Amharic, Arabic, and Nilo-Saharan.

CHADIC LANGUAGES

Hausa

Languages of the group. Of the Chadic (Chado-Hamitic) branch the most important language by far is Hausa. Its approximately 22,000,000 speakers live in northern Nigeria, in the Republic of the Niger, in the northern part of Ghana, in Cameroon, and in parts of Togo, Benin (Dahomey), the Chad Republic, and the Central African Republic. Hausa is also spoken as a second language by many speakers of other African languages. All the other languages of the Chadic branch are spoken by smaller ethnic groups; for example, the Bura, with about 1,500,000 people; the Mandara, with about 670,000; the Angas, with about 500,000; the Bolewa and Karekare, numbering about 220,000; and the Kanakuru, with about 150,000. In regard to the classification of the Chadic languages, some units are at times classified as languages but should preferably be treated as dialects, and vice versa, and the information on many languages (dialects?) is very scanty. Thus, all attempts at classification must be regarded as provisional only. The more important of the approximately 150 "languages" follow.

1. Western group: Hausa, Gwandara, Ngizim, Bedde (Bade), and related languages; Warjawa, Afawa (Paʾa), and related languages; Gezawa, Seiyawa, Barawa of the Dass region; Bolewa, Karekare, Kanakuru, and related languages; Angas, Sura, Ankwe, Gerka, and related languages; Maha (?).

2. Ron group: Fyer, Bokkos, Daffo-Butura, Sha, and Kulere.

3. Kotoko group: Logone, Buduma, Afade, Gulfei, and related languages and dialects.

4. Musgu (Musgum).

5. Masa group: Masa (Banana), Bana, Kulung; Mussoi (?); Marba (?); and Dari (?).

6. Eastern group: Somrai and related languages; Gaberi and related languages and dialects; Sokoro and related languages; Modgel; Tuburi (Kera); Mubi; Dangla-Jegu, Jonkor, and related languages.

The affiliation of the following languages and dialects to the Hamito-Semitic family has been questioned by some scholars:

7. Tera, Jera; Hona, Gaʾanda, and related languages; Bata, Gundu, and related languages; Margi, Bura, Chibak, and related languages; Higi (Hiji) and related languages; Laamang (Hidkala); Mandara (Wandala), Glavda, Yawotatakha, Sukur, and related languages.

8. Daba, Hina, Gauar, Musgoi (?); Matakam, Mofu, Gisiga, and related languages.

9. Gidder.

The total number of speakers of Chadic languages is probably about 26,000,000.

It is probable that the linguistic area of the Chadic branch formerly extended farther to the east, thus contacting the Omotic (Western Cushitic) group.

No Chadic language except Hausa has been reduced to writing; for Hausa, Arabic writing began to be used in the 16th century, and now a modification of the Latin alphabet is used.

Linguistic characteristics. In relation to the original phonological system of Proto-Hamito-Semitic, there are many missing sounds in the individual Chadic systems, but a comparison of the different Chadic groups shows that all the distinctive sounds of the parent language are represented in one way or another. Typical of all Chadic languages are tones serving to differentiate meaning in otherwise identical words.

The verb is for the most part a combination of an auxiliary verb (prefix conjugation) followed by a verbal noun (the reverse order is typical for Cushitic); the nominal part often has a vowel suffix pointing out certain qualities of the verb (transitiveness, intransitiveness, and so on). Verbal stem modifications exist, being expressed by such devices as reduplication and suffixes, but in most languages they are not strongly developed. Most of the languages have more or less clearly expressed genders but no declension. Plural is shown as in Cushitic and Berber. Chadic languages have innovations in the pronominal system. In vocabulary there are borrowings from English, Arabic (especially in Hausa), Fulani, and East Sudanic.

(I.M.D.)

KOREAN LANGUAGE

Korean is spoken by more than 55,000,000 people on the Korean peninsula and its coastal islands, and many among the approximately 558,000 Korean residents in Japan still speak the language. As the language of a sizable population occupying a strategic position in east Asia between China and Japan, Korean is important both historically and culturally. A considerable body of belles lettres and other literature is written in it.

Origin and classification. The direct precursor of the modern Korean language was the language of the ancient kingdom of Silla (*c.* 57 BC–AD 935), the original territory of which was in the southeastern part of the peninsula. The language of Silla spread over most of the peninsula when, in the 7th century, that kingdom conquered and annexed the territories of its rival states, Paekche (*c.* 18 BC–AD 660) in the southwest and Koguryŏ (*c.* 37 BC–AD 668) in the north. It is believed that the people of Silla and Paekche spoke dialects of the same language, while the people of Koguryŏ spoke a different, though related, language.

Possible relationship to Altaic

The most fruitful hypothesis concerning the affinity of Korean to other languages is that it belongs to the Altaic family, which includes the Manchu-Tungus, Mongol, and Turkic groups of languages. Like them, Korean is a language of the agglutinative type, and it shares with them many features of phonology and grammar. (Agglutinative languages combine into a single word several components, each of which remains relatively distinct in form and meaning). This theory is supported by a number of cognates and some sets of sound correspondences. The relationship is, however, probably a remote one.

Some scholars have proposed that Korean and Japanese are related. Their grammatical structures are similar in many ways, but their phonologies (sound systems) differ greatly. It is still an open question whether the sound correspondences demonstrated so far rigorously support a genetic relationship.

A standard spoken and written language is taught in the schools and used as the norm throughout Korea. This form of Korean was defined in the 1930s and is based primarily on the speech of the middle class of Seoul, with certain peculiarities excluded. Regional dialects, which are still vigorous, can be grouped into six major divisions: central, northeastern, northwestern, southeastern, southwestern, and Cheju-do, off the southern coast. These regional dialects differ from one another most conspicuously in intonation (the pitch pattern of a sentence), vocabulary, and the form of some grammatical endings.

Knowledge of the early stages of the language that existed before the introduction of the native Korean alphabet in the middle of the 15th century is still slight. The most extensive records of ancient Korean extant are 25 texts of songs, called *hyangga,* from the kingdom of Silla; these are quoted in works written in Chinese during the succeeding Koryŏ period (935–1392). In these writings Korean is written in Chinese characters used as phonetic and semantic symbols of Korean words.

Ancient Korean records

There are no generally agreed upon dates for the stages of development of the language. A possible dating is as follows: Old Korean (before the 12th century), Middle Korean (beginning of the 12th century to the end of the 16th century), Modern Korean (from the beginning of the 17th century).

The most penetrating external influence exerted on the language has been that of Chinese, from which many words were borrowed over the centuries. This influence has been limited largely to vocabulary. In addition, until the end of the 19th century, the prevailing written language of serious Korean literature was Chinese (*hanmun*) or, later, a deliberately Koreanized version of it.

Standardization of the spoken and written language became an urgent problem toward the end of the 19th century, when Korea undertook large-scale reforms to meet the social needs occasioned by internal changes and by pressures from the West. Efforts by Korean scholars, educators, and writers to develop a standard spoken and written language culminated in the publication by the Chosŏnŏ hakhoe (Society for the Study of the Korean Language) in 1933 of its *Han'gŭl mach'umppŏp t'ongiran* ("A Proposal for Unifying the Orthography") and its *P'yojunmal moŭm* ("Collection of Standard Forms of Words") in 1936. These standards became official only in 1945, upon Korea's liberation from Japanese rule. Some divergencies in the spelling system and in word usage have since developed in the north and in the south.

Native Korean alphabet

The writing system. The native Korean alphabet was introduced in 1446, after centuries of the use of cumbersome methods (known as Idu) to transcribe Korean with Chinese characters. The new set of 28 letters (not an adaptation of an existing alphabet) was designed by a group of scholars commissioned by Sejong (reigned 1419–50), the fourth king of the Yi dynasty (1392–1910). Although the alphabet was meant to be a "script for the people," as its original name *Hunmin-jŏngŭm* implies, it did not win acceptance as a respectable form of writing among the literati, or scholar-officials, who continued to write Hanmun. This alphabet, also called Ŏnmun (the common script), did, however, give rise to a body of premodern popular literature. Finally, in the beginning of the 20th century, the alphabet received general recognition as the national writing system. Today it is known as Han'gŭl in South Korea and as Chosŏn muntcha in North Korea.

Two methods of writing Korean are in use today: the purely alphabetic method and the "mixed script" method, in which Sino-Korean words (Chinese loanwords) may be written in their original characters and read in their Korean pronunciation. In North Korea only the alphabetic method has been in use since 1949; in South Korea both methods are in use, though it has been the government policy to dispense with Chinese characters in gradual stages. Four of the original 28 letters have gone out of use. The accompanying alphabet chart includes combinations of letters that are units in alphabetizing. Letters are grouped into syllable blocks, which contain at least an initial consonant plus a vowel. If the phonetic syllable has no initial consonant, a special letter is used as a filler for the initial consonant space in the syllable block and is silent in that position.

Linguistic characteristics. *Phonology.* The McCune-Reischauer system of transcription is used for the Korean in this article; see the alphabet chart for the phonetic values of the transcription symbols. In the phonology, the consonants show an unusual three-way distinction among a series of lenis (soft) consonants, *p, t, k, ch, s;* a fortis (hard), unaspirated series, *pp, tt, kk, tch, ss;* and a fortis, aspirated series, *p', t', k', ch'*. (Aspirated consonants are pronounced with an audible release of air.) These consonants are all typically voiceless (*i.e.,* pronounced without vibration of the vocal cords), but the lenis consonants *p, t, k,* and *ch* are voiced (*i.e.,* pronounced with vibrating vocal cords) when they come between voiced sounds, thus becoming pronounced as *b, d, g,* and *j* (as in "jam").

Three-way consonant distinction

The central characteristic of the Korean phonological system is that the consonants are more severely restricted as to position of occurrence within a word than within a morpheme (*i.e.,* within the stems and affixes of words). For example, the sound [s] cannot be pronounced at the end of a word, although some word stems do end in *s;* or such sequences as [s] + [k] cannot be pronounced in Korean,

Table 51: The Korean Alphabet (Han'gŭl)

	McCune-Reischauer romanization*			approximate equivalent sound in English		McCune-Reischauer romanization	approximate equivalent sound in English
	initial	medial	final†				
Consonants					**Vowels and Diphthongs**		
ㄱ	k	g	k‡	*c*ut, a*g*ain, ba*ck*	ㅏ	a	d*o*t
ㄲ	kk	kk	k§	s*k*y	ㅐ	ae	b*a*ck
ㄴ	n	n	n	*n*ine	ㅑ	yaδ	*ya*
ㄷ	t	d	t‡	*t*ake, i*d*eal, ho*t*	ㅒ	yae	*ya*mmer
ㄸ	tt	tt	t§	s*t*and	ㅓ	ŏ	c*u*t
ㄹ	r	r	l‖	wa*t*er, *l*eap	ㅔ	e	s*e*t
ㅁ	m	m	m	*m*an	ㅕ	yŏ	*you*ng
ㅂ	p	b	p‡	*p*ut, a*b*out, ho*p*	ㅖ	ye	*ye*t
ㅃ	pp	pp	p§	s*p*end	ㅗ	o	l*aw*
ㅅ	s	s	t‡	*s*o	ㅘ	wa□	*wa*nder
ㅆ	ss	ss	t§	me*ss s*ergeant	ㅙ	wae	*wa*x
ㅇ	(silent)	ng	ng¶	si*ng*er, si*ng*	ㅚ	oe	Ger. können, or English *we*t
ㅈ	ch	j	t‡	*ch*in, a*dj*ust	ㅛ	yo	*yaw*l
ㅉ	tch	tch	t§	mea*t ch*opper	ㅜ	u	m*oo*n
ㅊ	ch'	ch'	t♀	a*ch*ieve	ㅝ	wŏ	*wo*n
ㅋ	k'	k'	k♀	a*cc*ount	ㅞ	we	*we*t
ㅌ	t'	t'	t♀	a*tt*end	ㅟ	wi	*wee*p
ㅍ	p'	p'	p♀	a*pp*ear	ㅠ	yu	*you*
ㅎ	h	h		*h*ome	ㅡ	ŭ	b*oo*k
					ㅢ	ŭi	
					ㅣ	i	n*ea*t

*The system of phonetic transcription of Korean generally used for scholarly purposes in the United States. †Only seven consonant sounds (*k*, *t*, *p*, *n*, *l*, *m*, *ng*) are pronounced at the end of a word or before a consonant. Others are phonetically "reduced" to one of these seven when they occur in these positions; consequently, different letters may be romanized in the same way. The same letter may be romanized differently in different phonetic contexts because of sound changes. ‡Lenis (lax or softer than the corresponding tense consonant), voiceless, and slightly aspirated at the beginning of a word and typically voiced, except the sound *s*, between voiced sounds; *s* is palatalized as in English "she" when it comes before *i*, or *y* sounds. At the end of a word or before a consonant the sounds *k*, *t*, *p* are unreleased. §Tense (fortis or "reinforced"), voiceless, and unaspirated. ‖In South Korean published writing this letter is not found at the beginning of words except in some recent loanwords; in North Korean published writing it is found at the beginning of both Korean and foreign words. This letter, represented by *r*, is pronounced like the *t* in "water" when it comes between vowels or semi-vowels and at the beginning of words; when it comes at the end of a word or before a consonant, it is pronounced like English initial *l*. There is also a long *l* sound, romanized *ll*, that occurs between vowels or semi-vowels; this is the result of sound assimilations involving *l*. There is no special letter in the alphabet for it. ¶The sound of this letter in medial position is to be distinguished from *n'g*, which is the combination of *n* + *g* sounds. ♀Voiceless, and strongly aspirated. δThe *y* part of a diphthong is represented in the Korean orthography by an extra stroke added to the letter for a vowel. □The *w* part of a diphthong is represented in Korean orthography by the letter for the vowel *o* or the letter for the vowel *u*. ◇This diphthong has no close approximation in English.

but they can result from adding a suffix beginning with *k* to a stem ending in *s*. (The square brackets indicate that included letters stand for sounds pronounced rather than the conventional spelling.) The language has an extensive system of sound changes to resolve these conflicts between combining forms and permissible pronunciations. The standard orthography spells words in terms of the morphemes (constituent units) that comprise them, as long as the sound changes are covered by a regular phonological rule. Thus, most morphemes preserve their orthographic identity, although they may not be pronounced as spelled; *e.g.*, the noun stem for "five" is always spelled *tasŏs*, whether in the form *tasŏs-i* "five (as subject)," pronounced [tasŏsi], or in the form *tasŏs* "five (as a word by itself)," pronounced [tasŏt], with a final [t]. Forms based on the verb stem *pis-* "to comb," are spelled *pis-ŏ, pis-ko*, but are pronounced as [pisŏ] and [pikko] respectively. If sound changes are not covered by a phonological rule of the language, the principle is to spell words phonetically.

Grammar. Aside from interjections and primary (not derived) adverbs, a word in Korean typically consists of a stem plus an ending. Stems are of two types: (1) nouns and (2) verbs and adjectives. Nouns take endings that mark their syntactic role in the sentence—such as subject, topic (often equivalent to English "as for," as in "As for Korea, it's a peninsula."), object, genitive—or that express such meanings as "to, in, from." They are indifferent as to grammatical number and gender. Verbs and adjectives have much the same pattern of forms and enter into many of the same kinds of constructions. They may take one or a combination of two tense suffixes, roughly designated as past and future tense; the lack of a tense suffix indicates present tense or tenselessness. To the stem or to the stem plus the tense suffix is added an ending that concludes the sentence (called a sentence-conclusive form), or a conjunctive, nominal (noun), adnominal (*i.e.*, attributive), or adverbial form.

Basic clause structure

The basic structure of a clause is subject + predicate; in the predicate the main verb or adjective comes last. An example is [abŏjiga^1 ŏje^2 segŭmŭl^3 naesiŏssŭmnida^4] "Father^1 paid^4 his taxes^3 yesterday^2." This order applies to conjunctive clauses (*i.e.*, clauses similar to "but he left early" in "He came, but he left early.") as well as to independent clauses, and it holds, too, for independent clauses regardless of mood—declarative, interrogative, imperative, and so on; *e.g.*, [abŏjiga^1 ŏje^2 segŭmŭl^3 naesiŏssŭmnikka^4?] "Did Father^1 pay^4 his taxes^3 yesterday^2?"

The mood of the Korean sentence is specified by an ending on the verb or adjective that concludes the sentence. This ending expresses a complex of meanings—not only whether the sentence is a statement, a question, or a command, but also whether the speaker is reporting something from personal observation or from hearsay, and whether the speaker is certain or hesitant, and so on. At the same time, this ending specifies one of several possible levels of address for the sentence, reflecting the social relationship, conventionally defined, between the speaker and the person spoken to. For example, "Where^{2} is^{3} the post office^{1}?" is [uch'egugi^{1} ŏdi^{2} issŭ*mnikka*^{3}?] when spoken to a senior or a stranger, [uch'egugi ŏdi iss*ŏ*?] when asked of a friend, and [uch'egugi ŏdi in*ni*?] when addressed to a child. The relationship between speaker and person spoken to is one of three dimensions of a highly developed system of honorifics (forms showing deference). A second dimension, occurring between the speaker and the person spoken about (*i.e.,* the subject of the verb or adjective), requires a stem marked by the element [(ŭ)si-] when speaking of a senior. Thus, [abŏjiga^{1} sanpporŭl^{2} ka*si*nda^{3}] "Father^{1} is going^{3} for a walk^{2}" contrasts with [ch'ŏlsuga sanpporŭl kanda] "Ch'ŏlsu (a boy's name) is going for a walk." The third dimension concerns the speaker in speaking of himself before a senior.

Pronouns, especially those for 2nd person ("you"), are used sparingly. The pronoun subject or object may be left out and understood from the linguistic or nonlinguistic context.

Modifying elements precede that which is modified. The most characteristic way to modify nouns is by the use of adnominal (attributive) forms of verbs and adjectives or of whole clauses; examples are [chohŭ*n*^{1} nalssi^{2}] "weather^{2} that is good^{1}," or, "nice^{1} weather^{2}"; [naega^{1} tanidŏ*n*^{2} hakkyo^{3}] "the school^{3} that I^{1} attended^{2}"; [ka*l*^{1} ttae^{2}] "time^{2} to go^{1}." Adnominal + noun constructions play a variety of important roles in Korean syntax; *e.g.,* in nominalization processes (processes by which nouns or nounlike expressions are formed from verbs or verbal expressions, such as English "unifying" from "unify").

In multiclause sentences the clauses are attached one to the other in linear fashion. The last clause in the sentence is the main clause, containing the verb or adjective with the sentence-conclusive ending. The conjunctive relation is expressed by the conjunctive ending on the main verb or adjective in the preceding clause; *e.g.,* [mari tomada chokkŭmssik tarŭ*jiman*^{1} hakkyoesŏnŭn p'yojunmarŭl karŭch'inda^{2}] "Speech varies a bit from province to province, *but*^{1} in the schools they teach the standard language^{2}."

Vocabulary. Among the Sino-Korean words, which make up more than half of the vocabulary, there are many ordinary as well as learned words. Sino-Korean elements and now also Sino-Japanese ones are the primary resources for new technical terminology. Most Sino-Korean elements are used as nouns or nounlike units in Korean, regardless of their original part of speech. They can be made into verbs or adjectives by adding certain stems that mean "to do," "to be," "to become." This device is also used to form verbs and adjectives from native Korean nouns. (F.L.)

JAPANESE LANGUAGE

Japanese is the language of more than 121,044,000 persons on the islands of Japan, including the Ryukyus. In addition, there are more than 558,000 Koreans and about 44,000 Chinese in Japan who speak the language. Outside Japan a considerable number of people of Japanese parentage can speak the Japanese language with some degree of proficiency; these include 701,000 persons in the United States, 550,000 in Brazil, 60,000 in Peru, and 47,000 in Canada. In Korea and Taiwan the older generations speak Japanese as a second language.

For centuries there have been many dialects in Japan differing from each other to such an extent that some of them are mutually unintelligible. Since the Meiji Restoration, which ended with the promulgation of the Japanese constitution in 1889, the rapid development of elementary education has eliminated illiteracy in the country, and a common written language has been established, based on the dialect of the residential sections of Tokyo. At present, people of the various parts of Japan can speak the common language, although with their own accents, in addition to their own dialects. Since World War II, a common spoken language based on the same dialect of Tokyo has been exerting more and more influence upon the speech of the younger generation all over Japan via radio and especially via television; as a result, the local dialects are disappearing more rapidly than before.

The genetic relationship of Japanese to other languages has not been linguistically established. It is, however, probably related to Korean and possibly to the Altaic languages, which include the Manchu-Tungus, Mongolian, and Turkic families; all have similarities in their phonological and grammatical structures. Some lexical (vocabulary) and other resemblances, however, have been pointed out between Japanese and other East Asian languages and language families—*e.g.,* Austronesian (Malayo-Polynesian), Austroasiatic, Tibeto-Burmese, and Ainu.

Two major dialect groups

Dialects. Japanese can be divided into two major dialect groups: those of the mainland and those of the Ryukyu Islands. The mainland dialects are divided by some scholars into three groups—Eastern, Western, and Kyushu. In other systems, however, they are classified into the Eastern division and the Western division, which is then split into the Kansai (including the Chūgoku and Shikoku dialects) and the Kyūshū dialects. The Tokyo dialect belongs to the Eastern group, Westernized during the last two or three centuries; except for the accent and other features, it is not very different from that of Kyōto, a Kansai dialect, which was the most influential central dialect for more than 1,000 years. The peripheral dialects (*e.g.,* the dialects of remote areas—north Tōhoku and Kagoshima of south Kyūshū—as well as the Ryukyuan dialects) are very different from those of Tokyo and Kyōto and are incomprehensible in those cities.

There is evidence that there may have been a language spoken by the people of the Yayoi culture in north Kyūshū about 2,000 years ago that became the Proto-Japanese tongue—*i.e.,* the ancestor of all the present-day Japanese dialects. The Yayoi culture with its rice cultivation was brought from the Asian continent to Japan. If it is assumed that the people of this culture also brought the Proto-Japanese language, then it must also be assumed that the language of the remainder (if any) of the Proto-Japanese-speaking people on the Asian continent was extinguished by another language or languages; because there is no evidence of a language on the continent that is not only similar to Japanese but also differs from it enough to reflect the 2,000 years of separation.

A more probable hypothesis is that the Japanese language became separated from Korean 5,000 or more years ago and was spoken in Japan thousands of years before the Christian Era. In that case, Proto-Japanese was probably only one of the Japanese dialects spoken contemporaneously in most of Japan, including Kansai and the area east of it. The Yayoi culture is known to have diffused rapidly eastward from north Kyushu and apparently later stimulated the development of the Tumulus culture in Kinai (*i.e.,* the "Home Provinces," including Yamato, one of the centres of which was the city of Nara). The dialects that were contemporary with and different from Proto-Japanese must have been absorbed during the subsequent 2,000 years by the dialects that branched out from Proto-Japanese.

In the 8th century, however, people noticed that the Eastern dialects were remarkably different from those around Nara (and perhaps also from the Western dialects); some of their peculiarities were recorded. These peculiarities may have been non-Proto-Japanese features, which at that time were still resisting the influence of the central dialects but which have disappeared from the present-day Eastern dialects except for the dialect of a remote island, Hachijō-jima (about 300 kilometres south of Tokyo), that preserves some of the old peculiarities.

The differentiation of social dialects has been slight, but some peculiarities of the speech of the former samurai have been reported in several cities.

Phonological characteristics. Japanese is a polysyllabic language. Simple (*i.e.,* not compound) nouns consist of one or, more often, two or three syllables, to which various particles of one or more syllables are often suffixed. Various "inflectional forms" of simple verbs and adjectives usually consist of two or more syllables and some have various endings or particles or both suffixed as well. ("Inflectional forms" in Japanese are such forms [stems] as negative, preterite, conditional, imperative, and so forth.) The structure of the syllable is rather simple; syllables are ordinarily open (*i.e.,* they consist of one consonant and one vowel that is either short or long, with or without an intervening *y* after the consonant). The syllable can, however, be closed with a nasal sound (indicated by linguists with the symbol /N/) or a checked sound (conventionally symbolized as /Q/), which then acquire similar phonetic characteristics or become identical to the sound that follows. For example, /N/ becomes *m* before *p, b, m; n* before *t, d, n; ng* before *k, g, ng;* nasalized ĩ before *y;* nasalized ũ before *w;* and so forth. /Q/ is always followed by a consonant and is changed to an implosive *p, t, k, s,* etc., which forms a geminate sound (*i.e.,* a double consonant like the *t*'s in English "cattail") with the following explosive consonant. There are five vowels, *a, i, u, e, o,* similar to those of Italian. Short vowels are pronounced very short; the short *i* and *u* between voiceless sounds are devoiced (pronounced without vibrating vocal cords—voiceless) or omitted in the Tokyo speech. This tendency, remarkable in Tokyo, is less prevalent in some Western dialects.

Structures of the syllable

Japanese has the following consonants: *p, t, k, b, d, g, ts* (*ch*), *s* (*sh*), *z* (*j*), *m, n, r, h, y, w.* Basically, any vowel or *ya, yu, yo* can follow any of the above consonants. But there are restrictions to this rule that include several unacceptable combinations of sounds, among them *ti, tu, di, du, si, zi, wi, we,* and others. *Chi, tsu, ji, zu, shi, ji, i, e,* etc., occur in place of them, respectively. In pronunciation, the Japanese *p, t,* and *k* are not as aspirated as the initial *p, t,* and *k* of English (*i.e.,* they are not pronounced with a strong accompanying puff of breath). *B, d,* and *g* are fully voiced (pronounced with vibration of the vocal cords) as in French; *g* between vowels is usually a nasal *ng* (as in English "sing") in many Kinki (central) and Eastern dialects. Unlike the English sounds (which are formed with the tongue touching the gum ridge behind the upper teeth), *t, d,* and *n* are articulated against the teeth; *ch, sh,* and *j* are pronounced with the front of the tongue, not with the tip of the tongue as in English; initial *z* is pronounced like *dz* in English "adz" in Tokyo; and *r* is a flapped sound like the American *t* sound in "city."

Pitch accent

The majority of dialects, including those of Tokyo and Kyōto, have a word pitch accent. In Tokyo, for example, *hashi* with a high-low accent means "chopsticks," but with a low–high accent pattern it denotes "bridge"; in Kyōto, on the other hand, *hashi* means "edge, end" with high–high accent, "bridge" with high–low, and "chopsticks" with low–high. There are various patterns of pitch accent, and their geographical distributions are very complicated. The dialects of Tokyo and Hiroshima and those of Kyōto and Ōsaka have patterns quite different from each other. Some dialects in Tōhoku and Kyūshū, among others, have no pitch contrast at all.

It is a common feature of all the dialects, however, that they have no word stress accent (as occurs in English—*e.g.,* hăppy, fŏreigner, charactĕrístics). The sound of Japanese gives a very different impression from that of English, and it is said to be spoken with even stress and rhythm, as if a metronome were very rapidly ticking off each syllable.

Grammatical characteristics. Nouns have neither number (singular and plural) nor gender (masculine and feminine) and take no article (such as "the," "a," "an"). Case distinctions that show such grammatical features as subject and object are expressed with particles added to the ends of words; for example, *kodomo-ga* "a (the) child (children) [usually nominative]," *kodomo-no* "of a child," *kodomo-o* "a child [accusative]," *kodomo-ni* "to a child," *kodomo-kara* "from a child." Verbs have no person, no number, and no gender and are conjugated with the use of endings—*e.g., kaku* "write, writes, will write," *kakanai* "do (does, will) not write," *kake* "write [imperative]," *kakō* "I'll write, let's write," *kaite* "having written, writing," *kaita* "wrote, has (have) written," *kakeba* "if . . . write (writes)." Adjectives, which have no number, gender, or case, are also "conjugated" with suffixes—*e.g., shiroi* "is (are, am) white," *shiroku-nai* "is not white," *shiroku-te* "is white, and . . . ," *shirokatta* "was (were) white."

There are no relative pronouns. Demonstrative pronouns have a three-way distinction rather than the two-way division of English "this" versus "that"—*e.g., kore* "this," *sore* "that around you," and *are* "that far from both of us." Personal pronouns have a complex and intricate system: for the 1st person singular (English "I, me, my"), for example, there are *watakushi, watashi, washi, atakushi, atashi, temae, boku, ore,* and others; and for the 2nd person singular (English "you, your") there are *anatasama, anata, anta, kimi, omae, kisama, temé,* and others. The use of these forms depends on such factors as the social relationship of the addresser and the addressee(s) and the degree of intimacy between them, the formality of the speech, and the sex and age of the speaker. Verbs and adjectives also have conjugated forms corresponding to some of these distinctions. For example, *kakimasu, kakimasen, kakimashita* are used instead of *kaku, kakanai, kaita,* respectively, when the utterance is addressed to those who are superior or not intimate to the addresser. On the other hand, if the person who is the performer of the action *kaku* ("to write") is a superior to the speaker, then *oka-ki-ni naru* and *okaki-ni narimasu* are used instead of *kaku* and *kakimasu,* respectively. If the action is done for a superior, then *okaki suru* and *okaki shimasu* are used instead.

Order of words in sentences

The predicate stands at the end of a sentence. The subject, the object, adverb, adverbial phrase, and other elements of the sentence precede the predicate and can be omitted when possible. For instance, *Kaita,* which is a sentence consisting only of a predicate, can be translated as "I (he, they, etc.) wrote (have, has written) it," according to the context or the situation. The complement precedes the copula (a linking verb such as English "is"); *e.g., gakusei-da* "I (he, they, etc.) am (is, are) a student (students)" is composed of the complement *gakusei* "student" and the copula *da* "am, is, are." Sentences of the structure A-*wa* B-*da* (*wa* means "as for") have various structural meanings—*e.g., Are-wa daigaku-da* "That is a university" (*daigaku* = "university"), *Watashi-wa daigakuda* "I am for (from, in, etc.) the university," *Gakusei-wa gakusei-da* "A student is nothing but a student." The expressions that modify a noun directly precede it; *e.g., chiisana hitsuji* "a small sheep," *ōkina ōkami* "a large wolf,"

ōkami-no	*kutta*	*hitsuji* "the sheep that a wolf ate,"
wolf-[genitive]	ate	sheep

hitsuji-o	*kutta*	*ōkami* "the wolf that ate the sheep."
sheep-[accusative]	ate	wolf

Vocabulary. Although the greater proportion of the basic Japanese words are native words, a large percentage of the whole vocabulary is composed of Chinese loan elements—comparable to the loanwords from Greek, Latin, and French in English. In early times Japanese apparently borrowed a number of cultural words from Korean, but from the 6th century through the 9th century the direct contact with Chinese culture had a much greater influence on Japanese; the phonology, grammar, and basic vocabulary, however, were not as strongly affected. During those centuries several Sanskrit words entered into Japanese through Chinese in connection with Buddhism—*e.g., danna* "patron, master" (from *dannapati*), *hachi* "bowl" (from *pātra*), *kawara* "tile" (from *kapāla*). Later, in the 12th and 13th centuries, Zen priests introduced several words from Middle Chinese, such as *manjū* "bean-jam bun," *yōkan* "sweet paste," *udon* "noodle,"

Table 52: Japanese Kana*

simple *kana* symbols

No.	H	K	E	No.	H	K	E	No.	H	K	E	No.	H	K	E	No.	H	K	E
1.	あ	ア	a	2.	い	イ	i	3.	う	ウ	u	4.	え	エ	e	5.	お	オ	o
6.	か	カ	ka	7.	き	キ	ki	8.	く	ク	ku	9.	け	ケ	ke	10.	こ	コ	ko
11.	さ	サ	sa	12.	し	シ	shi	13.	す	ス	su	14.	せ	セ	se	15.	そ	ソ	so
16.	た	タ	ta	17.	ち	チ	chi	18.	つ	ツ	tsu†	19.	て	テ	te	20.	と	ト	to
21.	な	ナ	na	22.	に	ニ	ni	23.	ぬ	ヌ	nu	24.	ね	ネ	ne	25.	の	ノ	no
26.	は	ハ	ha	27.	ひ	ヒ	hi	28.	ふ	フ	fu	29.	へ	ヘ	he	30.	ほ	ホ	ho
31.	ま	マ	ma	32.	み	ミ	mi	33.	む	ム	mu	34.	め	メ	me	35.	も	モ	mo
36.	や	ヤ	ya					37.	ゆ	ユ	yu					38.	よ	ヨ	yo
39.	ら	ラ	ra	40.	り	リ	ri	41.	る	ル	ru	42.	れ	レ	re	43.	ろ	ロ	ro
44.	わ	ワ	wa													45.	を	ヲ	o
																46.	ん	ン	n(m)‡
47.	が	ガ	ga	48.	ぎ	ギ	gi	49.	ぐ	グ	gu	50.	げ	ゲ	ge	51.	ご	ゴ	go
52.	ざ	ザ	za	53.	じ	ジ	ji	54.	ず	ズ	zu	55.	ぜ	ゼ	ze	56.	ぞ	ゾ	zo
57.	だ	ダ	da	58.	ぢ	ヂ	ji	59.	づ	ヅ	zu	60.	で	デ	de	61.	ど	ド	do
62.	ば	バ	ba	63.	び	ビ	bi	64.	ぶ	ブ	bu	65.	べ	ベ	be	66.	ぼ	ボ	bo
67.	ぱ	パ	pa	68.	ぴ	ピ	pi	69.	ぷ	プ	pu	70.	ぺ	ペ	pe	71.	ぽ	ポ	po

digraphs representing single syllables

No.	H	K	E	No.	H	K	E	No.	H	K	E
1.	きゃ	キャ	kya	2.	きゅ	キュ	kyu	3.	きょ	キョ	kyo
4.	しゃ	シャ	sha	5.	しゅ	シュ	shu	6.	しょ	ショ	sho
7.	ちゃ	チャ	cha	8.	ちゅ	チュ	chu	9.	ちょ	チョ	cho
10.	にゃ	ニャ	nya	11.	にゅ	ニュ	nyu	12.	にょ	ニョ	nyo
13.	ひゃ	ヒャ	hya	14.	ひゅ	ヒュ	hyu	15.	ひょ	ヒョ	hyo
16.	みゃ	ミャ	mya	17.	みゅ	ミュ	myu	18.	みょ	ミョ	myo
19.	りゃ	リャ	rya	20.	りゅ	リュ	ryu	21.	りょ	リョ	ryo
22.	ぎゃ	ギャ	gya	23.	ぎゅ	ギュ	gyu	24.	ぎょ	ギョ	gyo
25.	じゃ	ジャ	ja	26.	じゅ	ジュ	ju	27.	じょ	ジョ	jo
28.	ぢゃ	ヂャ	ja	29.	ぢゅ	ヂュ	ju	30.	ぢょ	ヂョ	jo
31.	びゃ	ビャ	bya	32.	びゅ	ビュ	byu	33.	びょ	ビョ	byo
34.	ぴゃ	ピャ	pya	35.	ぴゅ	ピュ	pyu	36.	ぴょ	ピョ	pyo

hiragana trigraphs containing long vowels §

No.	H	E	No.	H	E
2.	きゅう	kyū	3.	きょう	kyō
5.	しゅう	shū	6.	しょう	shō
8.	ちゅう	chū	9.	ちょう	chō
11.	にゅう	nyū	12.	にょう	nyō
14.	ひゅう	hyū	15.	ひょう	hyō
17.	みゅう	myū	18.	みょう	myō
20.	りゅう	ryū	21.	りょう	ryō
23.	ぎゅう	gyū	24.	ぎょう	gyō
26.	じゅう	jū	27.	じょう	jō
29.	ぢゅう	jū	30.	ぢょう	jō
32.	びゅう	byū	33.	びょう	byō
35.	ぴゅう	pyū	36.	ぴょう	pyō

*H = *hiragana*; K = *katakana*; E = equivalent. Some *kana* undergo a change in pronunciation in specific situations.

†*Tsu* is also used to indicate a doubled consonant. In such cases the *tsu kana* is written slightly below the line (in horizontal texts) or slightly to the right of the line (in vertical texts) and sometimes also in slightly smaller script. Other *kana* are also positioned in this manner when they serve special functions.

‡Romanized *m* before *b*, *p*, and *m*.

§All five vowels have long forms, which in a Japanese text can be indicated in one of several ways. Romanized long vowels are indicated by macrons, except *i*, which is written *ii*.

isu "chair," and *futon* "bedding." Around the end of the 16th century, Japanese borrowed several words from Portuguese, like *pan* "bread," *kasutera* "sponge cake," *rasha* "woollen cloth," and *karuta* "cards." During the 18th and 19th centuries the language acquired several words from Dutch—*e.g., buriki* "tinplate," *garasu* "glass pane," *rappa* "trumpet," and *zukku* "duck, canvas."

A large part of the Chinese loanwords in contemporary Japanese are compound terms or derivative words coined in Japan since the Meiji Era; these combine two or more Chinese morphemes (word elements) that were borrowed in ancient times. *Kisha* "train," for example, is a Japanese compound of this kind that consists of *ki* "steam" and *sha* "car"; the Chinese equivalent, *ch'i ch'e,* means "automobile." The newly coined Japanese compound *jidōsha* "automobile" consists of three Chinese loan elements—*ji* "self," *dō* "to move," and *sha* "car"—but Chinese has no corresponding compound such as *tzu tung ch'e.* Some of these new Japanese compounds have been borrowed back into Chinese through the medium of Chinese characters.

Influence of Chinese characters on word formation

The Chinese characters play a very peculiar but important role in the word formation of written Japanese, which naturally has an influence on the spoken language. Every character usually has two readings: the *kun,* which is an indigenous Japanese word, and the *on,* which is an old Chinese loan morpheme. These two readings are closely associated with each other and alternate freely in word formation. For instance, *Keiō Daigaku* "Keio University" is abbreviated to *Kei-dai,* using the original *on* reading; whereas *Waseda Daigaku* "Waseda University" becomes *Sō-dai,* formed from the *on* reading for the indigenous Japanese word *wase* in *Waseda. Wase,* which means "early-ripening variety of rice," is shown by a combination of two Chinese characters meaning "early" and "rice," respectively. The *on* reading of the character meaning "early" is *sō,* hence *Sō-dai.*

In the vocabulary of contemporary Japanese, all the words beginning with *p-* are either onomatopoetic or recent loanwords from European languages—*e.g., pan* "bread" (from Portuguese), *pen* "pen" (from English) —because the initial *p-* of Old Japanese has changed into *h-* in Modern Japanese. In addition, there are very few indigenous words ending in a nasal sound. Such phonological peculiarities are now utilized by pharmaceutical and other companies, who coin for their new products names containing a *p* or a nasal sound so that they sound like something modern or imported. Especially since World War II, commercialism has deluged the language with English and other European words, many of which are short-lived. In general, however, almost all the foreign words have been adapted to Japanese sound patterns (*e.g., baiorin* "violin," *bisuketto* "biscuit," *rejā* "leisure"), and only a few have brought new sound combinations such as *ti, di* (*e.g., pātī* "party," *birudingu* "building").

Early records of Japanese

History. There is no evidence that the Japanese had their own script before they adopted the *kanji* (Chinese characters) early in the Christian Era. The earliest records of Japanese consist of several words found in a Chinese book of history, the *Wei Chih,* of the late 3rd century. A few words written with *kanji* are found on the swords and the mirrors of the 5th and 6th centuries, but the earliest extant Japanese documents of any length are the *Kojiki* (712) and the *Manyōshū* (later than 771) of the Nara period. From the 9th century on, records, mainly of the Kyōto dialect and the common written language, abound. The history of the language is usually divided into Old Japanese (to the 8th century), Late Old Japanese (9th–11th centuries), Middle Japanese (12th–16th centuries), and Modern Japanese (from the 17th century). Old Japanese was considerably different from Modern Japanese in phonology, morphology, and vocabulary, but not so much in syntax (the arrangement of words and word elements in sentences).

From Old Japanese to Modern Japanese there have been numerous sound changes, among them the shift of initial *p-* to *h-* in most of the modern dialects and the loss of three vowels, usually represented as *ï, ë,* and *ö.* Some remnant of vowel harmony was seen in Old Japanese. In vowel harmony certain vowels are restricted by the language structure from occurring in successive syllables of a word. Thus, in one word root or stem *ö* never co-occurred with *o,* and rarely with *u* and *a; e.g.,* the *-kö* ending in *kökö* "here," *sökö* "there" appears as *-ku* and *-ko* in *iduku* "where" and *miyako* "metropolis (*miya* "palace"). These words have become *koko, soko, doko, miyako* in Modern Japanese.

In Old Japanese and Late Old Japanese there was a distinction between "finite" forms (which occur at the end of a sentence as the predicate) and noun-modifying forms of verbs and adjectives—*e.g.,* the finite forms *uku* "receive(s)," *oku* "get(s) up," *shiroshi* "is white" differ from the corresponding noun-modifying forms *ukuru, okuru, shiroki* (later *shiroi*). In Middle Japanese the latter began to replace the finite forms (*uku, oku, shiroshi*), which ultimately disappeared from the spoken language. In Old Japanese and Late Old Japanese there was a rule of peculiar syntactic agreement: when the predicate verb or adjective was preceded by a word suffixed with such a particle as *zo* (emphatic), *ka* (interrogative), and so forth, the sentence was finished with the noun-modifying form instead of the finite form. This syntactic restriction disappeared, however, as the result of the above-mentioned change in syntax. Later, on the analogy of the forms such as *ukete* "having received" and *okite* "having got up," *ukuru* and *okuru* changed into the present forms—*ukeru* "to receive," *okiru* "to get up."

Kanji and *kana*

Writing systems. During the several centuries after the adoption of the *kanji,* the Japanese apparently used classical written Chinese as their formal written language. As they became accustomed to the characters, however, they tried to write Japanese with them, and in the process the *on* and *kun* of every *kanji* became established. Each *kanji* represents a Chinese word or morpheme, which has its own sound and meaning. The *on* is a Japanese imitation of the Chinese sound; *e.g.,* the 8th-century Chinese forms *pat* "eight" and *tap* "answer" became Japanese *pati* and *tafu,* respectively. The *kun* of a *kanji* is an indigenous Japanese word with a meaning similar to that of the Chinese; it is this reading that the Japanese are accustomed to give the *kanji.* Because the Chinese word *pat* means "eight," the *kun* reading of the *kanji* for this word is *ya,* which is a Japanese term for eight. The *kanji* for the Chinese word *pa* "wave" was read by the Japanese either as *pa* (an *on*) or as *nami* (a *kun*). When a *kanji* is used to represent a Japanese syllable by means of its *on* or *kun* without reference to its meaning, it becomes a kana, a phonogram. For instance, when the *kanji* for the Chinese word *pa* "wave" is used for the sound *pa* in such Japanese words as *pana* "flower" and *payashi* "fast," it becomes a *kana.* As early as the 6th century, there are examples of *kanji* used as *kana.* Although the *Kojiki* and the *Manyōshū* of the 8th century are written exclusively in *kanji,* the language represented is Japanese, not Chinese. The *Kojiki,* however, uses more *kun* readings, mixing them with *kanji* used according to the Chinese syntax, while the *Manyōshū* uses many more *kana.* This is why *kanji* used as *kana* are called *manyōgana* (*kana* often becomes *gana* in compounds).

As it was a toilsome task to write Japanese with the *kanji* (which are squarish and complicated in shape), the Japanese began to write them in such a cursive and simplified way of their own that the symbols retained little or no vestige of their original shape. The resulting syllabic characters, called *hiragana,* "common *kana*" (but known as *onna-de* "letters for women" in the Heian period), began to appear in the 9th century. These simplified characters were used extensively by women, who wrote many poems, diaries, and novels during that period. There is, however, evidence that men also learned and used the script, although they wrote their diaries in *kanji.* Parallel with this script another system of syllabic writing developed in the 9th century; it was called *katakana* (*kata* "one side, one of a pair"). When the priests of the temples in Nara read Chinese texts, especially Buddhist scriptures, they would translate into Japanese as they went along, and would jot down beside the *kanji* for their own memory the Japanese particles, endings, and so forth that were lacking in Chinese. This was done with symbols made for private use, The growing popularity of hiragana

mainly symbols formed by abbreviating the strokes of the *kanji.* Originally every sect, and sometimes every person, had a special system, so that there were various symbols for one and the same syllable. In the 10th century, however, more common features began to appear, indicating that the symbols were becoming more common and popular in use. In this way Japanese began to be written with *kanji* and *katakana* intermixed and sometimes only with *katakana.* In the 15th century, however, the *hiragana* symbols, which were cursive and fine in shape, became the more popular script; literary works were written in *hiragana,* while scholarly or practical books were written in *katakana.* Even some literary works written with *kanji* and *katakana* in earlier days were rewritten with *kanji* and *hiragana.* The present orthography is in *kanji* and *hiragana,* and only European loanwords and onomatopoetic words are written with *katakana.* Only *katakana,* however, is used in telegrams and in notes typed or printed with machines in companies and offices.

Orthographic reform

In the beginning of the Meiji Era, the grammar of the language written with *kanji* and *kana* still was archaic, based mainly on that of Old Japanese and early Late Old Japanese. Toward the end of the 1880s, however, famous writers such as Futabatei Shimei, Yamada Bimyō, and Ozaki Kōyō began a successful movement to write in the colloquial style. The orthography, with its thousands of *kanji* symbols, was difficult to learn and to use. Therefore, after World War II, the government carried out a series of reforms that had been advocated for many years. The *kana* spelling, based on the sounds of early Late Old Japanese, was changed to conform to the contemporary pronunciation; and the *kanji* characters, which had been used without restriction, were limited to 1,850 symbols for official and daily use, and their shapes and strokes were greatly simplified.

There also have been a number of advocates of romanization since the Meiji Era, but such a program presents many difficulties. Old Japanese syllables had a much simpler structure than those of the contemporary Chinese, which is a monosyllabic language without word endings. Therefore, many different Chinese words or morphemes became homophones (words pronounced alike) in Japanese as early as the 8th and 9th centuries; *e.g.,* the Japanese imitated the Chinese *k, k', x* (and *g,* γ) sounds with only a single *k* sound. Moreover, sound changes in Japanese during the succeeding 12 centuries produced a great number of homophones; for example, the sound sequences *au* (from Chinese *au* and *ang*), *afu, ou,* and *ofu* that were distinct from each other in Old Japanese and early Late Old Japanese have all become the same *ō* today. Accordingly, Japanese now has a great many homophonous *kanji*, so that the Chinese loan morphemes that clearly convey certain meanings when written with *kanji* would very often become incomprehensible when romanized. Moreover, *kanji* have more characteristic configurations than would the same words written with roman letters, and thus enable rapid reading. In addition, a high literacy rate prevails throughout the entire Japanese nation, and the people are presumed to be so accustomed to the *kanji* and *kana* that such a great change in writing as romanization would undoubtably encounter very strong resistance. (S.Ha.)

AUSTRONESIAN LANGUAGES

The Austronesian language family, also called Malayo-Polynesian, consists of languages spoken in almost the whole of Malaysia and the Indonesian Archipelago, all of the Philippines, parts of Vietnam, Kampuchea (Cambodia), and Taiwan (Formosa), Madagascar, and on all of the main island groups of the South and Central Pacific (except for Australia and a large part of New Guinea, which contain languages belonging to other stocks). In terms of the number of its languages and of their geographic spread, the Austronesian language family is among the world's largest.

Austronesian languages are generally divided into two primary subgroups: a Western, or Indonesian, branch contains perhaps 200 languages, including such well-known tongues as Malay, Indonesian, Javanese, and Pilipino, the national language of the Philippines (based on Tagalog); and an Eastern branch, more commonly termed Oceanic, comprises about 300 small languages, scattered throughout the South and Central Pacific, the best known of which are the Polynesian group and Fijian. The classification of a small residue of languages is uncertain. While around 170,000,000 people speak languages belonging to the Western Austronesian branch, only about 1,000,000 speak Oceanic languages.

The challenge of piecing together the complex history of a family of languages with such an enormous distribution, and with speakers of such great cultural and physical diversity, has attracted scholars to the study of Austronesian languages. Recently, the Austronesian languages have also received attention because of their structural characteristics and, in addition, have served as a testing ground for subgrouping methods and theories of linguistic change.

HISTORY AND CLASSIFICATION

Eastern and Western Austronesian groups

Resemblances between the languages of Madagascar, the East Indies, and Polynesia were first pointed out in 1706 by Hadrian Reland, a Dutch scholar. It was not until the second half of the 19th century, however, that the languages of the intervening island groups of Melanesia and Micronesia were recognized as belonging to the same family. At that point it became customary to divide Austronesian languages into four groups coinciding with the geographic regions known as Indonesia, Melanesia, Micronesia, and Polynesia. More recently, it has become clear that such a classification is not linguistically valid. Though arguments continue over the position of certain languages, it is generally agreed that most Austronesian languages fall into only two groups. The majority of the languages formerly assigned to the Melanesian and Micronesian divisions, together with the Polynesian group, form a single, although internally very diverse, subgroup: Eastern Austronesian (more often called Oceanic). The Western division coincides fairly closely with the old Indonesian grouping, comprising most, and possibly all, Austronesian languages spoken west of New Guinea, together with two found in Micronesia.

The main areas of disagreement in classification concern the position of the Formosan languages, of certain languages located in eastern Indonesia and on or near the western tip of New Guinea, and of the more divergent languages in Melanesia. The Formosan languages are usually treated as Western Austronesian, though some scholars have suggested that they represent a third primary branch of Austronesian. The languages of east Indonesia, including the western end of New Guinea, more diverse than those of west Indonesia and the Philippines, are sometimes treated as a single subgroup of Western Austronesian, and sometimes regarded as composed of a number of primary branches, each coordinate with Oceanic and with a group comprising the remaining members of Western Austronesian. Similarly troublesome are a number of languages of the north coast of New Guinea and certain regions of insular Melanesia. These languages share very few related words with each other and with other languages in the family, although in grammar they usually show quite strong resemblances to members of the Oceanic subgroup. They are generally regarded as aberrant Oceanic languages, but other theories have been advanced to explain their divergent character.

Proto-Austronesian. A protolanguage is a parent language, or early form of a language or group of languages. It can either be a hypothetical reconstruction or may actually exist in written records, as does Latin, the protolanguage of the Romance tongues. There are no actual records of Proto-Austronesian, but it has been the subject of research and reconstruction by linguists. Because it is

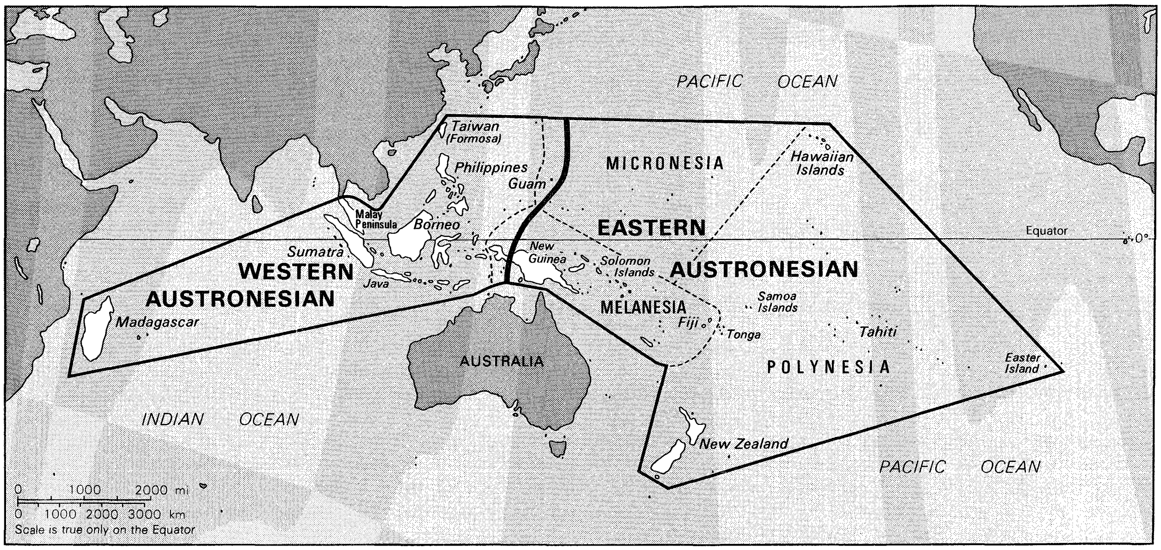


Figure 28: Major divisions of the Austronesian languages.

clear that the Proto-Austronesian speech community possessed agriculture and may have been responsible for its introduction—along with that of several other important cultural innovations—into the Pacific, the reconstruction of Proto-Austronesian and the history of the dispersal of Austronesian languages has been of considerable concern to culture historians as well as linguists.

Location. The location of Proto-Austronesian has been the subject of much speculation but little systematic investigation. At various times in the past the parent tongue has been placed somewhere in the Southeast Asian mainland, South China, and even in India and Mesopotamia. There is increasing evidence of an archaeological, geographic, and linguistic nature, however, that the homeland lay in the region of Indonesia and New Guinea.

Chronology. The best evidence for dating and determining the directions of the dispersal of Austronesian languages comes from Oceania, where it is easier to correlate linguistic and archaeological findings than in the west. It is clear that Austronesian languages were already in Fiji and Tonga, near the eastern margin of Austronesian territory, by 1000 BC. Archaeological excavations indicate that the coastal area of Fiji was widely settled by that date, that by then Tonga also had been settled by a people with a material culture essentially identical to that of the first inhabitants of Fiji, and that Samoa was inhabited a few centuries later. In each place there is continuity of material culture—and, one may assume, continuity of language—right through to the period of European contact. The first Europeans found Fiji and Polynesia (and indeed the whole of the neighbouring island groups—the New Hebrides, New Caledonia, and Micronesia) to be occupied exclusively by Austronesian-speaking peoples. Glottochronology, a technique for dating the division between languages (known as linguistic splits) based on the assumption that there is a stable rate of basic vocabulary replacement in languages, places the separation of the Fijian and Polynesian subgroups at between 3,000 and 4,000 years ago. The divergence of Proto-Polynesian into separate branches is dated at between 1,800 and 2,500 years ago.

These are relatively recent branchings, far down on the Oceanic limb of the Austronesian family tree, and it follows that the separation of the common ancestor of Fijian and Polynesian from more distant branches of Oceanic must have occurred somewhat earlier. Glottochronological estimates indicate that diversification of Austronesian languages began around 4,000 to 5,000 years ago in the New Hebrides, in New Caledonia, and in the Solomons, and earlier still in the region of New Guinea. Very little archaeological work has been done in these areas, but assemblages of artifacts similar to those found in early Fijian and Tongan sites have been unearthed in New Caledonia and the central New Hebrides and dated at 800 BC and 600 BC respectively. The general trend, at least, is clear. The dispersal of Austronesian languages in Oceania cannot have begun later than around 2000 BC, with 3000 BC appearing to be a more realistic estimate.

Glottochronological computations suggest that the differentiation of the Western Austronesian languages was well advanced by 1000 BC.

Relationships to other families. Many different proposals have been made to link Austronesian with other language groups—Mon-Khmer, Munda, and Vietnamese of the Austroasiatic language family, Tai-Kadai, Sino-Tibetan, and Indo-European, among others. None has been convincing. Ultimately, no doubt, Austronesian languages, like every other family in Oceania, must derive from ancestral stages spoken in Asia at some remote period. Discovery of such distant connections, however, will have little bearing on the question of where the ancestral Austronesian language itself developed.

Reconstructions of Otto Dempwolff

Reconstruction. Many scholars have worked to reconstruct the Proto-Austronesian sound system and word stock. The reconstructions of Otto Dempwolff, a German ethnologist and linguist, published between 1920 and 1938, have remained the point of departure for all subsequent comparative studies. Dempwolff attributed to Proto-Austronesian a four-vowel system, consisting of a low vowel *a* and three higher vowels—*i* (front), *e* (central), and *u* (back). The reconstructed consonants are the voiced stops *b, d, D, j, g* (a stop is a sound made with complete stoppage of the breath from the lungs); the matching voiceless stops *p, t, T, c, k* and the glottal stop *q;* the nasal consonants *m, n, ñ,* ŋ (nasals are pronounced with the breath going through the nose); the semivowels *w* and *y;* plus *l, r, R, h, s, z,* and *Z.* (Phonetic value of phonemes represented by capital letters, and that of certain other reconstructed symbols, cannot be precisely determined.) Most word bases consisted of two syllables, the commonest shape being consonant-vowel-consonant-vowel-consonant or consonant-vowel-consonant-consonant-vowel-consonant. Clusters of consonants were restricted to a few types and occurred only in the middle of words and possibly in initial position. Words with initial vowels or final vowels or both were probably more common than Dempwolff's reconstructions allow.

Although some languages, particularly certain members of the Oceanic group, have changed this sound system drastically, it is still reflected fairly faithfully by many Western Austronesian languages. Indeed, in both Western Austronesian and Oceanic languages, many words seem to have persisted in almost unchanged form, a condition unparalleled in the Indo-European languages, for example. Table 53 compares the forms of some Proto-Austronesian words with the cognate terms in four modern languages.

Systematic reconstruction of Proto-Austronesian grammar has scarcely begun. Structural features retained by a majority of languages in both major branches include a

Table 53: Some Proto-Austronesian Terms and Their Related Forms in Several Modern Languages

	Proto-Austronesian	Tagalog	Malay	Fijian	Samoan
two	*Duwa	dalawa	dua	rua	lua
four	*e(m)pat	apat	empat	vā	fā
five	*lima	lima	lima	lima	lima
six	*enem	anim	enam	ono	ono
bird	*manuk	manok	manu	manu-manu	manu
eye	*mata	mata	mata	mata	mata
road	*Zalan	daan	jalan	sala	ala
pandanus	*panDan	pandan	pandan	vadra	fala
coconut	*niuR	niyog	nior	niu	niu

*Form that is not actually found in any document or living dialect; it is a reconstructed, hypothetical form.

fairly constant form for each grammatical and vocabulary element, with boundaries between elements in words being clearly definable, and a relatively simple morphology of verbs and nouns. In addition, in the verb phrase, a number of elements indicating tense, aspect, and voice are present; they were evidently prefixes and infixes (particles inserted within the body of a word) in Proto-Austronesian. Reduplication, the repetition of a word or a portion thereof, occurs in the case of the verb root and has several functions. Most roots are capable of being used either as nouns or verbs. Adjectives, numerals, and markers indicating negatives can act as verbs. Noun subclasses include personal names, marked by a personal article; common nouns, marked by a common article; locatives, place names, and directionals, marked by a locative particle; and temporals, which are not marked. Personal pronouns include distinct forms for 1st person including the hearer and 1st person excluding the hearer; pronouns marking subject or possessor differ in form from those marking object or focus.

Current research. Since World War II the descriptive and comparative study of Austronesian languages has expanded considerably, with centres at universities in the Philippines, Indonesia, Australia, New Zealand, the United States, and Europe. Reasonably good dictionaries and grammars are now available for most of the better known Western Austronesian languages and many of the Polynesian languages. Several large areas remain poorly known, however, including most of Melanesia and much of Borneo and east Indonesia. A good deal of recent research has concentrated on classification and associated problems in the methodology of subgrouping and the theory of linguistic change. A major work has been a lexicostatistical classification of over 200 Austronesian languages prepared by the American linguist Isidore Dyen (1965).

Lexicostatistical classification

A lexicostatistical classification uses statistics to compare the vocabularies of two or more related languages; the method is similar to glottochronology (see above) but assumes a constant rate of change only within a given language family.

WESTERN AUSTRONESIAN (INDONESIAN)

The Austronesian languages lying west of New Guinea, together with the Chamorro and Palauan tongues of Micronesia, are often called Indonesian. The term is unfortunate in that it does not do justice to the wide geographic distribution of the languages concerned and can be confused with the name of Indonesia's national language. The commonly used alternative, Western Austronesian, is, therefore, employed here.

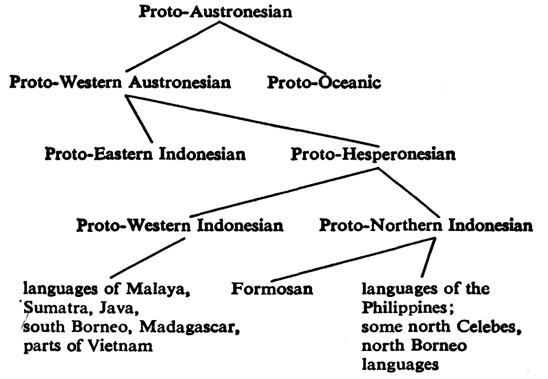
The classification of Western Austronesian languages is not completely agreed upon. There is, however, fairly general recognition of one very large grouping containing most of the languages of west Indonesia and of Malaysia, all the languages of the Philippines and Madagascar, some languages of the northern Celebes, and the Chamic group of Vietnam and Kampuchea. The name Hesperonesian has recently achieved some acceptance as a convenient label for this grouping. Disagreement centres around the place of the Austronesian languages of Taiwan and east Indonesia, including the western end of New Guinea. Many linguists regard these languages as being most closely related to Hesperonesian, and, in particular, treat the Formosan languages of Taiwan as a branch of Hesperonesian with closest relatives in the Philippines. In this view, the term Western Austronesian refers to a primary subgroup of Austronesian, contrasting with Oceanic, as indicated in the family tree in Table 54.

Other scholars regard the east Indonesian–west New Guinea languages as forming a number of primary subgroups, each coordinate with Hesperonesian and Oceanic. Similarly, a case recently has been made for treating the Atayalic group in Taiwan as a primary branch of Austronesian. The Atayalic languages share a very low percentage of basic vocabulary with all other members of the family, including other Formosan languages. There is, then, some support for an alternative family tree, with several primary branchings, representable roughly as in Table 55.

Branches of the Hesperonesian group

The Hesperonesian group is generally regarded as dividing into two main branches (see Table 54). One, Western Indonesian, includes the languages of Malaya, Sumatra, Java, Madura, Bali, and Lombok, south Borneo, Madagascar, and probably the Chamic group of Vietnam and Kampuchea. The other, Northern Indonesian, encompasses the Philippine languages, some of the languages of north Borneo and the northern Celebes, and possibly the Formosan languages.

Table 54: Austronesian Family Tree (Theory 1)

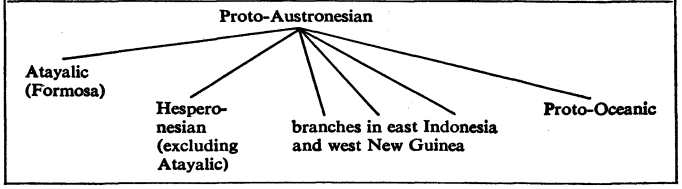

Proto-Austronesian
Proto-Western Austronesian
Proto-Oceanic
Proto-Eastern Indonesian
Proto-Hesperonesian
Proto-Western Indonesian
Proto-Northern Indonesian
languages of Malaya, Sumatra, Java, south Borneo, Madagascar, parts of Vietnam
Formosan
languages of the Philippines; some north Celebes, north Borneo languages

Regional division of Western Austronesian languages. In the following sections, the main Western Austronesian languages are discussed in more detail under the appropriate regional headings.

Western Indonesia and Malaysia, Madagascar, Vietnam. The west Indonesian–Malaysian region includes the islands of Sumatra, Java, Madura, Bali, Lombok, and Borneo, and the Malay Peninsula. Except for a few Mon-Khmer languages in the interior of Malaya, its indigenous languages are all Austronesian. There have been important influences on the languages and cultures from mainland Asia, however, and, more recently, from Europe. Hindu culture reached Java and Sumatra in the 1st century AD, and from the 4th century diffused to Borneo and the Celebes. Many languages of western Indonesia contain extensive borrowings from Sanskrit. Islāmic influence began in the 7th century and spread over much of Indonesia and the Philippines. European colonization, beginning in the 16th century, added a third layer of borrowings to the Sanskrit and Arabic.

Malay in its several dialects is the native language of some 11,000,000 people occupying both sides of the Strait of Malacca. Its expansion throughout Malaya is apparently recent; until the 13th century it was confined to

Table 55: Austronesian Family Tree (Theory 2)


Proto-Austronesian
Atayalic (Formosa)
Hesperonesian (excluding Atayalic)
branches in east Indonesia and west New Guinea
Proto-Oceanic

Bahasa Indonesia

east and south Sumatra and the facing coastline of the Malay Peninsula. For some centuries prior to European contact, a simplified form of Malay, Bazaar Malay, had been the lingua franca in coastal regions of the whole Indonesian Archipelago. This status was attained because of the strategic position of Malacca on the trade routes and its function as a dispersal centre of Islām, and also because of the seafaring and trading skills of the Malay speakers. Malay was in turn adopted by the Dutch and British administrations as their lingua franca. In the present century, the language, now the national tongue of the Indonesian state, is called Bahasa Indonesia, or Indonesian. The choice of Malay as a politically acceptable national language was facilitated by the fact that it was the first tongue of only a small minority of Indonesians (in contrast to such languages as Javanese and Sundanese) but was, at the same time, widely used in Indonesia as a second language. The Indonesian variant of Malay has undergone spelling reforms, and a large new vocabulary has been coined to cover modern technical concepts. Many words have also been borrowed from European languages and Javanese.

Malay is closely related to most of the other languages of Sumatra, including Minangkabau, Kerintji (Kinchai), Rejang, and Achinese, and to Madurese of Madura Island. Other well-known Sumatran languages are Gayo and Toba-Batak, spoken in the north and somewhat less closely related to Malay. In all, Sumatra contains between 12 and 15 languages (the boundaries between dialects and languages not being clear-cut) spoken by more than 13,000,000 people. Engganese, used on a small island off southeast Sumatra, is an extremely divergent Indonesian language that shares only about 10 percent of basic vocabulary with the rest.

Styles, or registers, of Javanese

Java, the most populous island of Indonesia, is linguistically one of the most uniform, with only three indigenous languages. Javanese, with about 54,000,000 speakers, is numerically the largest Austronesian language after Indonesian (which is known by over 100,000,000 people, but mainly as a second language). Javanese dialects are spoken throughout central and east Java and in sections of western Java. Old Javanese is known from inscriptions dating back to the 9th century AD. The language was extensively influenced by Sanskrit between the 9th and 15th centuries, during the time of the Indianized kingdoms of east Java. Thoroughly documented by Dutch linguists in a series of grammars and dictionaries produced over the last 250 years, Javanese is now generally written in the roman (Latin) alphabet and continues to flourish alongside Indonesian as a literary language and as the daily medium of communication of many Javanese newspapers, magazines, and radio stations. Several registers, or styles, distinguishing degrees of respect and marked by vocabulary differences, have arisen out of the elaborate stratification of Javanese society.

Sundanese, with about 17,000,000 speakers, is found throughout west Java, with the exception of small islands of Javanese speakers along the north coast of Banten and in the region of Indramayu. The language has been written in various scripts since the 14th century, roman now being in general use. The dialect spoken around Bandung is considered to be the most prestigious, or the standard form. Madurese is the language of Madura Island and smaller offshore islands, but about half of its 8,700,000 speakers now reside in east Java. There are two principal variants of Madurese—Eastern and Western; an Eastern dialect, Sumenep, has been adopted for educational purposes as the standard dialect. Madurese shows greater similarities to the Malay–Sumatran group than to Javanese or Sundanese. Balinese, with about 2,600,000 speakers, is spoken on Bali and the western part of Lombok; Sasak is the language of eastern Lombok.

Linguistically the most diverse of the islands of the western Indonesian region, Borneo is also the least known. It is sparsely populated, and most coastal areas have been occupied in relatively recent times by speakers of Malay-type languages, such as Iban (Sea Dayak), Brunei Malay, Kutai Malay, Banjarese, and Sambas Malay, all of which are closely related dialects or languages. The most important Bornean language is Ngadju (Ngaju), spoken as the native language in southwest Borneo in the region of the Barito, Kapuas, Kahayan, Kaitingan, and Sampit (Mentaya) rivers; it is more widely used as the lingua franca of most of south Borneo. In northeast Borneo there are a number of languages with close relatives in the Philippines, including Illanum (Lanum), Bajau, Sulu, and the various Murut dialects. Maanyan and closely related languages spoken in south Borneo are apparently the closest relatives of the Malagasy dialects of Madagascar.

The Malagasy dialects

The Malagasy dialects probably derive from Bornean traders who journeyed to Indian, Arabian, and East African ports and settled the previously uninhabited island of Madagascar. A quite close correspondence in basic vocabulary (45 percent agreement) between Maanyan and Malagasy indicates separation around 2,000 years ago. The Malagasy dialects show sufficient internal diversity to justify classification into two or three different languages. The Merina dialect, now official in Madagascar, has around 8,800,000 speakers. A small group of Austronesian languages spoken in Vietnam and Kampuchea—the Chamic languages—probably falls into the Western Indonesian subgroup. Cham, with 76,000 speakers in Vietnam and until 1975 about 90,000 speakers in Kampuchea, has been the subject of several studies. Rade, Curu (Chru, Cru), and Jarai, spoken in Vietnam, may be closely related to Cham.

The Philippines and Taiwan. At least 70 languages are spoken in the Philippines in a land area a little larger than that of Great Britain. There are two main subgroups. A Central (Mesophilippine) division includes many of the languages of central and southern Luzon, Mindoro, Palawan, and the Visayan Islands. In the early 1980s Tagalog was the mother tongue of about 11,400,000 people in central and southwest Luzon, including the region of the capital city, Manila, and, parts of coastal Mindoro. A standardized form of Tagalog was selected as the national language after the Philippines gained independence in 1946. This language, officially called Pilipino, is now spoken as a first or second language by a considerable proportion of the 50,000,000 Filipinos. Although it faces competition from English and other important local languages, it is increasingly used in education and government and in the popular press, radio, and literature of the Philippines. Closely related to Tagalog are the Bikol or Bicol dialects (over 3,500,000 speakers in southeast Luzon and smaller adjacent islands). Also in the Central branch are Cebuano, or Sugbuhanon, with about 12,400,000 speakers, and Hiligaynon, or Ilongo, with more than 4,800,000 speakers, the two main members of the Bisayan subgroup. Spoken primarily in Cebu, Bohol, west Leyte, and the eastern third of Negros, Cebuano is used as a trade language throughout Mindanao. Hiligaynon is spoken in Panay and on smaller adjacent islands and in Negros. Pampangan (Kapampangan) is another important language, spoken in Luzon on the northwest flank of the Tagalog speech area.

Northern Philippine subgroup

The second major Philippine group, called Northern or Cordilleran, contains most of the languages of northern Luzon. Chief among them is Ilocano (Iloko), the lingua franca in northern Luzon, which has more than 5,400,000 speakers in much of northwest and northern Luzon. Other members of this branch include the languages of the Igorot mountain tribes, including Bontoc, Ifugao, and Tinggian, the Gaddang group, Isneg, Isinay, and Kalinga. Pangasinan, spoken by 1,100,000 in the province of the same name, probably belongs to the Northern subgroup.

A number of Philippine languages fall outside the two large subgroups. Maranao, Tiruray, Bilaan, and Tagabili are languages of important groups in the southern Philippines, while Luzon contains several unclassified tongues—Ilongot, Casiguran, and others. These languages share around 30 percent of basic vocabulary and many structural characteristics with other Philippine languages but are not closely related to any large subgroup. The Sulu Archipelago has Tau Sug and Samal; both are also spoken on Borneo, where Tau Sug is known as Sulu. The major Philippine languages—Tagalog, Cebuano, and Ilocano—have extensive literatures, and together with other important languages, are widely used in broadcasts and newspapers.

Prior to the arrival of the Chinese, Austronesian languages were probably spoken over most of Taiwan. The surviving languages are now confined to small communities in less accessible regions. The Formosan languages of Taiwan are internally fairly diverse, but there is some evidence for treating them as a single subgroup of Western Austronesian. The Atayalic or Northern group contains the Atayal dialects and Seedik. Although some central, eastern, and southern languages share less than 25 percent of basic vocabulary with each other, most of them are generally assigned to a single group (Central or East Formosan). Its members include Ami, Paiwan, Bunun, and Thao. Tsou, Saaroa, and Kanabu (central Taiwan), though they probably may form a third subgroup, are ordinarily treated as members of the Central division.

Two languages of Micronesia, Chamorro (Mariana Islands) and Palauan (Palau, western Carolines), are generally regarded as deriving from the Philippines or Indonesia.

Austronesian tongues in eastern Indonesia and west New Guinea

Celebes, eastern Indonesia, west New Guinea. Eastern Indonesia is centred in the Lesser Sunda Islands and the Moluccas. The large island of Celebes occupies a position intermediate between Borneo, the Philippines, and eastern Indonesia. Approximately 100 different Austronesian languages are spoken in eastern Indonesia, in places on the coast of west New Guinea, and on the Aru Islands south of the Doberai (Vogelkop) Peninsula, west New Guinea. Papuan languages are spoken in parts of Timor in the Lesser Sundas and Halmahera in the Moluccas.

Dutch and Indonesian linguists have long recognized that this region contains many disparate Austronesian groups but have generally assigned them to a relatively small number of divisions: Bima–Sumba, Ambon–Timor, Sula-Batjan, south Halmahera–west New Guinea, and several Celebes groups. Very thorough lexicostatistical comparisons provide little support for most of these larger groupings and indicate instead that east Indonesia–west New Guinea harbours many small, genetically diverse subgroups, most of which are members of or have their closest relationships with the Hesperonesian group and with each other, with a residue not at present assignable to any large subgroup of Austronesian.

The Celebes, with more than 40 languages, is the most diverse area. Buginese, with about 3,400,000 speakers, is the most important language of the group. Gorontalo and Suwawa, in the north Celebes, and Sangir (Sangihe [Sangir] Islands, north of the Celebes) show closer agreement with Philippine languages than with other Celebes tongues. Other numerically large speech communities in or near the Celebes are the Sidjai, Duri, and Mandar (Andian), Kendari, Muna (Muna Island), and Butung (Butung Island). Sikka of Flores Island and Solor of nearby Solor Island appear to be quite closely related to each other, with a 37 percent common basic vocabulary. Endeh of Flores is more distant. Havunese, spoken on Sawu and Raijua, between Flores and Timor, and Sumba (Sumba Island) appear to be fairly distant from all other eastern Indonesian languages. A large number of similar isolated languages exists on smaller islands in this region; one such small group includes Ambonese of Ambon Island and the adjacent coastal area of Seram. Buli and Minyafuin, of Halmahera, may belong to a group that also includes As, spoken on the western tip of New Guinea, and Biga, from Wakde Island in northwest New Guinea. The Bomberai Peninsula group in west New Guinea has close relatives on east Seram. Most of the Austronesian languages on the north coast of west New Guinea, as far east as Sorera Bay, however, show no close relationship to the Hesperonesian, eastern Indonesian, or Oceanic languages.

Great internal diversity of western Austronesian

Characteristics of Western Austronesian. The Western Austronesian languages are so diverse internally that little can be said about their common characteristics apart from what has already been said of Proto-Austronesian. Most of the members of this division maintain distinctions in the Proto-Austronesian sound system that are lost in Oceanic; *e.g.,* the distinction between **b* and **p.* (An asterisk indicates an unattested, hypothetical, reconstructed form.) The original verb morphology is also generally retained more completely than in the Oceanic division, but it is difficult to isolate common innovations that mark off Western Austronesian languages in the absence of detailed reconstructions of Proto-Austronesian grammar.

The Philippine languages, typologically the most uniform of the large subgroups of Western Austronesian have been the best described. No doubt owing to their long coexistence in a single region, as well as to their relatively close relationship, members of this subgroup of Hesperonesian are more alike in their sound systems (though not in vocabulary) than the languages of the West Indonesian subgroup. Consonants usually include three voiceless stops, *p, t, k,* and the glottal stop *q;* three voiced stops, *b, d,* and *g,* and the nasals *m, n,* and ŋ; plus *l, r, h, s, w,* and *y.* In many languages, borrowings from Spanish have resulted in the addition of two new vowels, *e* and *o,* to the original four (*a, i, u,* and ɨ or ə) retained from Proto-Austronesian. The morphology of the verb and, to a lesser extent, of the noun is fairly complex. The use of affixes and reduplication distinguishes several tenses (past, future, and present or general), various aspects (progressive, distributive, causative, etc.), and two modes (obligatory and indicative).

Philippine voice and case systems

The grammar of Philippine languages is most famous for its complex voice and case systems. (Voice, in grammar, indicates the relationship of the subject of the verb to the action that the verb expresses. A grammatical case is a form of a word that shows the relationship of the word to other words.) Besides distinguishing so-called active (*e.g.,* "John washed the dishes") and passive (*e.g.,* "The dishes were washed by John") constructions, Philippine languages possess at least three kinds of passive. The noun phrase occurring as "topic" or "focus" in the sentence can represent the actor (subjective or active constructions), the goal or object (goal or object-focus passives), the referent (benefactive or location-focus passives), or the instrument (instrumental-focus passives). The distinctions are expressed by three devices. One uses affixes to mark voice in the verb: subjective voice is usually indicated by an infix *-um-* and the prefixes *ma-, mag-, maka-* (the actual choice varies with the different verb classes); objective voice is shown by the suffix *-in* or *-en;* referential voice by *-an;* and instrumental voice by the prefix *i-.* The other methods of distinguishing passives include the marking of case with nominal (noun) particles and transposing phrases and changing stress. An example from Tagalog is: "The child bought the mango," *b-umili, ang bata, n-ang mangga; bili* is the verb "buy" with the subjective infix *-um-*, *bata* is "child," and *mangga* is "mango"; *ang* marks the focus common-noun phrase ("the child"), and *n-ang* indicates the nonfocus common-noun phrase ("the mango"). To say "The mango was bought by the child," the sentence construction is rearranged to *b-in-ili nang bata ang mangga,* which in actual Tagalog order is "buy-objective voice, the child, the mango." A number of other arrangements of the sentence elements are possible to cover the instrumental and referential voices.

Members of the Western Indonesian and Formosan groups share a great deal of morphological and syntactic structure with the Philippine languages, although they are typologically less homogeneous. Most of the Indonesian languages have developed more elaborate vowel systems than that of Proto-Austronesian, with five (Havunese), six (Malay, Balinese, Buginese), seven (Sundanese), eight (Javanese, in one analysis), or nine (Cham) contrasting vowels. Maanyan and the Malagasy dialects retain the four original vowels. Most Western Indonesian languages preserve the distinction between four voiced and four voiceless stops (the voiced *b, d, j, g,* and the voiceless *p, t, c, k*) and four nasals (*m, n, ñ,* ŋ), plus *l, r, w, y, s,* and *h.* Only Javanese, however, also distinguishes retroflex stops and alveolar stops (made with the tip of the tongue touching the ridge behind the upper teeth). This distinction seems likely to have arisen from the later influence of Sanskrit, which possessed retroflex consonants, on Javanese. As in all branches of Austronesian, most word bases may be used in the function of nouns or verbs. There are numerous transformative and formative affixes in Western Indonesian that derive adverbs, adjectives, nouns, comparative

Possible influence of Sanskrit on Javanese forms, ordinal forms, and various verb constructions, in addition to several kinds of verb reduplication; these features also appear in the Philippine languages.

The large Hesperonesian group, in general, appears to be the most conservative branch of Austronesian. A high proportion of the 2,207 words attributed to Proto-Austronesian, for example, are retained by Tagalog (1,125), Toba–Batak (1,299), Javanese (1,446), Malay (1,627), and Ngaju (1,170). Much lower proportions are present in Oceanic languages (Bauan Fijian 461, Tongan 328, Samoan 385).

Eastern Indonesian languages in general appear to be morphologically simpler than members of the Hesperonesian group, in which regard they resemble Oceanic languages. They have, however, retained several distinctions in sound that have been lost in the Oceanic group.

Although books and articles on Western Austronesian languages run into several thousands, documentation of the speech traditions of many of the smaller, minority communities is still poor, and many gaps remain in the understanding of the history of the group as a whole. Indonesian and Filipino linguists are now working extensively on their own languages, while several Dutch and American universities, and the Summer Institute of Linguistics in the Philippines (an association of Protestant missionary linguists that specializes in studying unrecorded languages) are active in both descriptive and comparative studies.

European languages contain a number of words borrowed from the better known languages of the Indonesian region. From Malay derive such English words as sarong, (to run) amuck, mandarin (through Portugese), the kris (or creese) dagger, and the names of several animals (orangutan, pangolin, dugong, kalong) and plants or plant products (kapok, paddy, nipa). Javanese contributions include junk (sailing vessel) and batik. The American English word boondocks is from Tagalog *bundok* "mountain."

EASTERN AUSTRONESIAN (OCEANIC)

Roughly 300 Austronesian languages are spoken east of Sorera Bay in New Guinea and on the islands of Melanesia, Micronesia, and Polynesia.

Classification. The classification of these languages remains somewhat controversial. Dempwolff found evidence to indicate that all known Austronesian languages in the Oceanic region, apart from Chamorro and Palauan, belong to a subgroup that excludes Western Austronesian languages. The names Oceanic and, less commonly, Eastern Austronesian are now applied to this grouping. The evidence consists of a number of common innovations in the treatment of Proto-Austronesian sounds. All Oceanic languages agree in having lost the original final consonants of most word bases in most contexts, and in simplifying the initial and medial consonants as shown in Table 56. In addition, Oceanic languages reflect a five-vowel system in which Proto-Austronesian *a, i, u* were retained, *ay* became *e,* and *e* and *aw* became *o* in Proto-Oceanic.

Table 4: Correspondences Between Proto-Austronesian (PAN) and Proto-Oceanic (POC) Consonants

PAN	p b	mp mb	t nt	d D	nd nD	l r		(n)s (n)z (n)j (n)Z
POC	p	mp	t nt	d	nd	l r		s,z
PAN	k g	ŋk ng	m	n ñ	ŋ w	q R	h	y
POC	k	ŋk	m	n	n w	q R	Ø (zero)	y

Though the evidence cited above is not challenged, the conclusions drawn from it have been. Scholars remain puzzled by the great differences among Oceanic languages, especially in vocabulary items. Comparisons of basic vocabulary show that several New Guinea–West Melanesian groups share less than 15 percent of a common basic vocabulary with all other members of Austronesian, indicating a divergence from the parent language at least 5,000 years ago. It has been suggested that the Oceanic languages may not constitute a single subgroup but may divide into several primary divisions within the Austronesian family, and indeed that West Melanesia may be the primary dispersal centre for Austronesian. A quite different theory, with a few modern adherents, explains the diversity among the languages of Melanesia by deriving them from relatively recent mixtures of Indonesian and Papuan languages, which results in disparate varieties of pidginized Indonesian with Papuan substrata in each island group.

Puzzling differences among Oceanic languages

There are increasing grounds for accepting Dempwolff's theory of the Western Austronesian–Oceanic division as correct, while still allowing for a more dramatic time depth. It is also evident that, after a period of development as a single language, most of the descendants of Proto-Oceanic that dispersed in the New Guinea–West Melanesian area were influenced by nearby Papuan languages; this, together with other factors, such as the small size of the speech communities, word taboos, and several millennia of separate development, probably account for the extreme diversity found in this part of Oceania. It is unlikely that the New Guinea–West Melanesia region was the primary dispersal centre for Austronesian languages, but it is certain that it was a very early one. From there Oceanic-speaking groups moved fairly rapidly into the unoccupied islands of East Melanesia, Micronesia, and Polynesia. As has been noted, both Fiji and Tonga were inhabited by 1000 BC, almost certainly by speakers of the branch of Oceanic ancestral to Fijian and Polynesian, while the Solomons, New Hebrides–Banks Islands, and New Caledonia were probably settled by 2000 BC or earlier. In New Guinea and the western Melanesian islands, Oceanic-speaking peoples encountered Papuan populations and in many places intermarried with them and borrowed from their languages. One result of this contact is that many Austronesian languages in that region have changed faster than those situated further east.

Austronesian encounter with Papuan languages

Characteristics of Proto-Oceanic. Among the additional evidence for Dempwolff's hypothesis that has come to light in recent years is the sharing by the Oceanic languages of many words and grammatical features that are not characteristic of Western Austronesian languages. More than 30 percent of Proto-Oceanic words from a standard list of 215 "basic vocabulary" meanings are not represented at all in Western Austronesian. Oceanic languages also agree in having peculiar forms for certain common Austronesian words; *e.g.,* the Proto-Oceanic forms are **au* "I," **mai* "come," **suRi* "bone horn," **pati* "four," **Moze* "sleep," **katoluR* "egg," where Western Austronesian languages reflect **aku, *maRi, *duRi* or **DuRi, *e(m)pat, *peZem, *teluR.* In grammar, a number of common simplifications and elaborations are observable. The general features of the following discussion of Proto-Oceanic grammar apply generally to present-day Oceanic languages, although some languages retain the parent system more completely than others. Most languages of the central and eastern Solomons, some languages of the central and northern New Hebrides and Banks Island, the Fijian dialects, the Polynesian group, and certain languages in New Guinea–West Melanesia appear to have preserved the Proto-Oceanic system more completely than others.

Some morphological simplifications occurred in Proto-Oceanic: many verb and noun affixes were lost, some prefixes lost their prefix quality and were assimilated as nonproductive elements into the word base, and some affixes normally appended to a word became reinterpreted as free-form particles. On the other hand, Proto-Oceanic developed several new prefixes and changed the shape or function of certain old ones.

Whereas tense, aspect, mode, and probably person and number of the verb were marked in Proto-Western Austronesian by affixes (*e.g.,* as in English "walk," "walked"), in Proto-Oceanic they were indicated by separate particles. Many verb bases also functioned as postverbal particles indicating direction or tense-aspect—*e.g.,* Proto-Oceanic **zake* "ascend," **zipo* "descend," **mai* "come," **nopo* "stay" also occurred as modifiers meaning "upward," "downward," "toward," and "progressive aspect," respectively.

Nominal (noun) phrases were of at least four types. Common nouns were marked by the "common article" **na,* personal pronouns and possibly personal names were indicated by personal articles (**i,* probably **a* or zero—unmarked—in certain contexts), locative nouns were shown by **(q)i* immediately preceding the base word, and temporal nouns were marked by a temporal prefix or zero (the lack of a prefix). One subclass of locatives, called relationals, used personal suffixes, as *(*q*)*i lalo-na* "at inside-his" (*i.e.,* "inside him").

Pronouns in Oceanic

As in Proto-Austronesian, personal pronouns fell into two sets of variants: subjective and possessive forms, and focus and objective forms. Examples from the Wayan dialect of Fijian that preserve the Proto-Oceanic forms are: "I go," *qu laka* (*qu* is the subjective pronoun "I," *laka* is "go"); "my arm," *qu-lima* (*qu* is the possessive pronoun "my," *lima* is "arm"); "It is I," *o au* (*o* is the personal article, *au* is the focal or emphatic pronoun "I"); "give (it to) me," *vaganī au* (*vaganī* is "give," *au* is the objective pronoun "me"). Several new distinctions appeared in Proto-Oceanic, however. One was the development of dual (two-person) and trial (three-or-more-person) suffixes. The forms **ru*(*a*) "two" and **tolu* "three" were suffixed when two or a few people, rather than an unlimited plurality, were referred to. Plurals were indicated by the simple nonsingular base, as Fijian *keda* "we plural (including hearer)" contrasting with *keda-ru* "we two (including hearer)" and *keda-tou* "we three (including hearer)."

It was in possessive constructions that Proto-Oceanic underwent some of the most distinctive elaborations. Common nouns in such structures fell into at least three "genders": edible (foods, drinks, etc.), familiar or inalienable (kinship terms, parts of a whole), and neutral. The possessor was linked to the possessed noun by a particle of variable form. When the possessed noun was of edible gender, the consonant of the possessive particle was **k*-; in instances of neutral gender it was **n*-; and when of familiar gender, the consonant was omitted. The vowel or vowels of the possessive particle varied according to whether the possessor was a personal pronoun, a common noun, or a proper noun. Similarly, word order depended on the class of the possessed and possessor nouns: the particle plus possessor normally followed the head noun, as in Fijian *na uvi kei Manu* "Manu's yam," *na vosa nei Manu* "Manu's speech," *na tina i Manu* "Manu's mother," but, when the possessor was a pronoun and the possessed noun was of other than familiar gender, then the particle and pronoun preceded, as Fijian *na ke-mu uvi* "your yam," *na no-mu vosa* "your speech" (but *na tina-mu* "your mother," *na mata-mu* "your eye").

Proto-Oceanic interrogative pronouns included several retained from Proto-Austronesian: **zapa* "what?", **zai* "who?", **piza* "how much? how many?", and **kuya* "how? in what state?" together with **pai* "where?", and **ŋ*(*a*)*iza* "when?". Demonstratives distinguished three positions or persons: *[*i,e*]*ni* "this, here, near me, now," **ina* "that, there (by hearer), then," and a form marking remote position in space and time.

As in Proto-Austronesian, most word bases, other than proper nouns, functioned as both verbs and nouns, and adjectives, numerals, and negatives played the role of verbs in sentences.

Subgroups of Oceanic languages. Many aspects of the subgrouping of Oceanic languages are still obscure. In the west, many small divisions are evident, but no large subgroup has been attested. In the east, there is some evidence for a wider subgroup—Eastern Oceanic—comprising the languages of the southeast Solomons, much of the New Hebrides–Banks Islands, Rotuma, Fiji, and Polynesia, and possibly also most of the Micronesian languages. Within Eastern Oceanic, there is fairly clear support for a subgroup consisting of Fijian and Polynesian, whose closest immediate relatives may lie in the central New Hebrides. Table 57 presents a classification that has some degree of general acceptance.

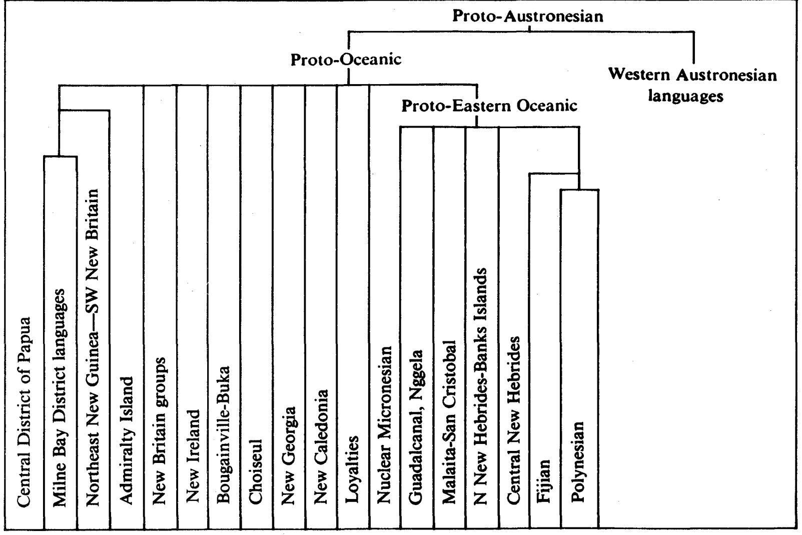
LANGUAGES OF MELANESIA

At a conservative estimate, 250 different Austronesian languages (not counting dialectal variants) are found in Melanesia. Most of them are spoken by communities of a few hundred to a few thousand people; with several exceptions they remain documented only by brief word lists and sketches. The term Melanesian is often applied to the Austronesian languages of Melanesia but has validity only as a geographical label, for these languages are not a linguistic unity comparable to the Polynesian and Nuclear Micronesian groups. Rather, Polynesian and Micronesian are each branches of particular Oceanic subgroups whose other members are in Melanesia (see Table 57).

Papuan and Austronesian languages in New Guinea

In New Guinea, Austronesian languages occur on the Doberai (Vogelkop) Peninsula, extending southwest as far as Kaimana, and in patches along the north coast, and along the southeast coast of Papua New Guinea as far west as Cape Possession (100 miles northwest of Port Moresby). Only in the Markham Valley and parts of the Central Province of Papua New Guinea are Austronesian languages found any distance inland. The interior and large stretches of the coast of New Guinea are occupied by Papuan. Apart from those tongues situated around Sorera Bay and on the Doberai (Vogelkop) Peninsula, all the New Guinea Austronesian languages appear to belong to the Oceanic division. The best known of them is Motu, spoken by about 13,600 people who occupy the coastal strip around Port Moresby. Police Motu, a simplified form of Motu, is used as a lingua franca throughout Papua. Motu is quite closely related to the eight other languages of Papua's Central Province. This group, in turn, appears to form a larger grouping with many of the languages spoken around the southeast tip of Papua New Guinea and the offshore islands, including Suau (Suau Island and adjacent coastal area) and Dobu (Dobu and Normanby Islands). These two languages, together with Kiriwina from

Table 57 : Classification of the Oceanic Languages



the Trobriand Islands and Wedau spoken around Wedan, have some currency as lingua francas in their respective regions of the Milne Bay Province.

Austronesian languages are found in coastal pockets in the Northern Province of Papua New Guinea, in the Morobe, Madang, East Sepik, and West Sepik provinces of Papua New Guinea, and on many offshore islands in these regions. Most of them have been strongly influenced by Papuan languages; *e.g.,* Yabêm and several closely related languages in the Morobe Province have developed tone, like their Papuan neighbours. Yabêm and Graged (Gedaged) of Kranket Island, in the Madang Province, have gained wider currency as the lingua francas of the Lutheran Mission in their regions. Some of the Morobe and Madang languages appear to fall into a subgroup with Kove (Kombe) and certain other languages of southwest New Britain.

Austronesian tongues of New Britain

The large island of New Britain contains around 25 extremely diverse Austronesian languages, the most important being Tolai (Kuanua, Tuna, Raluana), spoken around Rabaul on the Gazelle Peninsula and used extensively by missions in New Britain and New Ireland. New Ireland's languages are poorly known but appear to be of only moderate diversity. On the small Admiralty Islands group at least 20 (in some estimates as many as 50) different languages are found.

The Solomon Islands contain more than 60 languages that belong to several divergent groups. Bougainville is predominantly Papuan speaking, but several Austronesian languages are present on the north coast, and a few in the south and on offshore islands. The adjoining island of Buka contains several more, most of them closely related to the Bougainville languages.

On Choiseul Island are found an indeterminate number of languages and dialects closely related to each other but not to any outside group. Babatana is the literary language of a Methodist Mission there. Santa Isabel (Ysabel Island) is sparsely peopled by communities speaking perhaps 10 languages; Kia in the northwest and Bugotu in the southeast are the best known. Bugotu, the lingua franca of the Melanesian mission on Santa Isabel, is closely related to the Guadalcanal and Nggela languages. The New Georgia group contains more than a dozen languages—*e.g.,* Roviana (southwest New Georgia) and Marovo (east New Georgia and Vangunu Island)—most of which belong to a single division. Roviana is used by the Methodist Mission everywhere in the Solomons except on Vella Lavella. The languages of the southeast Solomons divide into two subgroups. Almost all those on Malaita and San Cristóbal belong to a single group whose best documented members are Sa'a, Kwara'ae (Fiu), Lau (all of Malaita), and Arosi (of San Cristóbal). The other group includes most of the Guadalcanal languages, together with Bugotu and Nggela, of Florida Island.

Polynesian languages spoken on several islands near the main Solomons group include: Nukuria, Nukumanu, Taku, Luangiua (Ontong Java), and Sikaiana to the north; Rennell–Bellona to the south; and Tikopia, Anuta, and Pileni–Taumako (Duff Islands) to the east. The latter is spoken in the Santa Cruz group, which also contains one Papuan and several little-known Austronesian languages.

Languages of the New Hebrides–Banks–Torres archipelago

Approximately 60 different languages, many of them dialectally very diverse, are spoken in the New Hebrides–Banks–Torres archipelago. The best known are Mota (of Mota or Sugarloaf Island, in the Banks group), used in the 19th century throughout the northern islands and in the Solomons as the lingua franca and literary language of the Melanesian Mission, and the Nguna–Tongoa dialects, spoken on north Efate and adjacent islands in the central New Hebride. Many of the central and northern languages are closely related to Fijian and Polynesian. Further south, Eromanga, Tana, and Aneityum harbour three languages that show no very close relationship to other members of the Oceanic group. Polynesian languages are spoken on Emae, Futuna, and Aniwa and on Mele and Vila islands in the New Hebrides.

Both New Caledonia and the Loyalty Islands are internally fairly diverse. New Caledonian languages appear to form a subgroup characterized by far-reaching sound changes. Languages of the central and far southern regions of the island are tonal. A Polynesian language, West Uvean, is spoken on, Beautemps-Beanpré (Heo) Island in the Loyalties.

Easily the best described, and politically the most important, language in Melanesia is Fijian, or Bauan Fijian, spoken as a first or second language by about 260,000 Fijians and also by other ethnic groups in the Fiji group. The Fijian dialects are sharply differentiated into Western and Eastern Divisions. Bauan, an Eastern dialect, became Fiji's lingua franca in the 19th century and is now widely used in the popular press, broadcasting, and administration. Rotuman, spoken on Rotuma Island, 200 miles north of Fiji, has borrowed heavily from Polynesian, but its classification as a language closely related to Fijian and Polynesian has been disputed.

LANGUAGES OF POLYNESIA

A triangle-shaped area, the apexes of which are Hawaii, Easter Island, and New Zealand, embraces almost all the islands of the central and eastern Pacific. Early European explorers found the inhabited islands in this region occupied by culturally homogeneous peoples speaking some 16 closely related languages. The name Polynesian was first applied to them by 19th-century scholars, who also observed that another 12 to 14 languages belonging to the Polynesian group are situated to the west of the Polynesian Triangle, on islands in Melanesia, and on the southern fringes of Micronesia; these are the Polynesian "Outliers."

Origins of the Polynesians

The question of Polynesian origins has been the subject of much romantic speculation. Various writers have pictured the ancestral Polynesians as migrating from such far-flung homelands as India, Egypt, Mesopotamia, and the Americas. It is now quite clear that, linguistically at least, the Polynesians developed their distinctive characteristics within Polynesia itself. Their languages form a subgroup of the Oceanic branch of Austronesian, whose immediate relatives lie in eastern Melanesia (see Table 57). Fijian, Polynesian, and certain languages of the New Hebrides were a single language until approximately 1000 BC, when some speakers of this early eastern Oceanic language appear to have moved from the New Hebrides into Fiji. There a dialect developed that subsequently split into Fijian and Polynesian branches. During several centuries of isolation somewhere in west Polynesia, probably in Tonga, the Polynesian branch underwent those further modifications that characterize all present-day Polynesian languages.

Characteristics of Proto-Polynesian. The general features of the following sketch of Proto-Polynesian are applicable to all its descendants. Each has changed some of the details for the Polynesian languages show an internal diversity comparable to the Romance or Germanic subfamilies of the Indo-European language family.

Phonology. Proto-Polynesian retained the five-vowel system of Proto-Oceanic but added a contrast between long, or geminate (double), vowels and short vowels. Several simplifications occurred in the consonant system, with the loss of the nasal element in stops (Proto-Oceanic *p* and *mp* both became *p; t* and *nt* became *t; k* and *ŋk* became *k*) and the loss of retroflex *R.* Proto-Polynesian consonants thus comprised four stops, *p, t, k,* and ʔ (the glottal stop); three nasal sounds, *m, n,* ŋ; three fricatives, *f, s, h*; plus *l, r,* and *w.* No consonant clusters or final consonants were permitted, but vowel sequences were common. Most word bases had two syllables, but longer forms were not uncommon, while some prebasic particles were monosyllabic—*e.g.,* **fale* "house," **waka* "canoe," **wai* "water," **uaua* "vein," **fafine* "woman," **tuakana* "older sibling of same sex," **maanifinifi* "thin," **ki* "to," **ma* "and," and **o* "of."

Significance of the phrase in Polynesian

Grammar. Proto-Polynesian retained most features of the Proto-Oceanic grammar, the main changes being in possessive constructions and in the development of distinctive passive and nominal (*i.e.,* verbless) sentence structures. The smallest natural phonological or pause unit was the phrase rather than the word, a phrase consisting of a head word flanked by affixes and particles. The verb phrase contained several classes of preposed particles: (1) tense-aspect particles (marking past, prospective, subjunctive, imperative, negative subjunctive, hypothetical, nonpast, and progressive); (2) subject-marking personal pronouns occurring between tense-aspect particle and verb; (3) manner particles, indicating the "manner" of the action; and (4) negatives. Verbal prefixes included **fe-* reciprocal, **faka-* causative, **ma-* abilitative (*i.e.,* "able to," "capable of"), **tua-* ordinal, **taki-* distributive, **toko-* human number. Verbal suffixes included a passive **-(C)ia,* instrumental **-aki,* and derivative or transformative **-(C)i, *-(C)aki.*

Noun phrases divided into common, personal, locative and temporal, each class of noun being marked by a distinctive article or other syntactic marker. Proto-Polynesian probably distinguished only two common articles, a definite **(t)e* and an indefinite **ha,* but several of its descendants have more elaborate series. Most interrogative pronouns were retained from Proto-Oceanic, but manner and temporal interrogatives were innovations. In personal pronouns the distinction between trial (three) and plural was lost, the old trial forms becoming the new plurals. Demonstratives included **eni* "this, these, here," **ena* "that, those, there (near hearer)," and **ia* "that, those (particular ones)."

In possessive constructions the three-way gender distinction of Proto-Oceanic was replaced by a two-way distinction between nouns subordinate to the possessor and nouns not subordinate to the possessor.

Case relations between noun phrases and the verb were indicated by particles placed at the beginning of the phrase. Active sentence structures normally consisted of verb phrase plus subject noun phrase plus complement, and contrasted with passive, nominal, and topicalized structures.

Samoan, Tahitian, and Hawai'ian

Languages and subgroups. The best known Polynesian languages are Samoan, spoken by about 200,000 people in Western and American Samoa and by sizable migrant communities in New Zealand and the United States; Tongan, with about 96,000 speakers; Tahitian, spoken as the native language in the Society Islands and as a second language by Marquesans, Tuamotuans, Magarevans, Austral Islanders (Îles Tubuai), and other ethnic groups in French Polynesia; New Zealand Maori, with about 100,000 speakers; and Hawai'ian, once the language of perhaps 100,000 people, but now used as a daily medium of communication by only a few hundred Hawai'ians. Samoan and Tongan are the national tongues of Western Samoa and the Kingdom of Tonga. Hawai'ian, Maori, and Tahitian, together with the languages of the Marquesas, Tubuai (Austral), Tuamotuan, and Cook island groups, and Magareva and Easter Island, form a well-marked subgroup known as Eastern Polynesian. Samoan with Futunan, Uvean, Tokelauan, Ellice, Pukapukan (in the northern Cooks), and all the Outlier languages spoken in Melanesia and Micronesia form a second division known as the Samoic-Outlier group. A third subgroup consists of Tongan and Niuean. Eastern Polynesian and Samoic-Outlier languages possess certain innovations that suggest they shared a period of common development after separating from the language ancestral to Tongan and Niuean. This wider grouping is known as Nuclear Polynesian.

The numerous subgroupings and language divisions in Polynesia are the result of the geography of this region. Around 2,000 years ago, following the breakup of Proto-Polynesian and the build-up of populations on the main islands in the west Polynesia area, there was a fairly rapid settlement of all the main uninhabited islands to the immediate east and west. Distances between most of the island groups were such that regular contact between them was impossible, and a distinct language evolved for each group. The last islands to be settled appear to have been the marginal east Polynesian islands, such as Hawaii, New Zealand, Mangareva (all settled about AD 1000), and some of the Outliers. Surprisingly, Easter Island was one of the first eastern islands to be settled, a fact attested by archaeological remains and by the considerable differences between the Easter Island language and those of other members of the Eastern Polynesian group.

The oral literature of Polynesia

The oral literature of Polynesia is extremely rich. Much of it has been recorded by Polynesian and European scholars and by missionaries, who in the 19th century developed excellent orthographies—using the roman alphabet—for Polynesian languages. In the precontact period, Easter Island was the only community to have developed a script of its own; it was an ideographic form of writing with very restricted uses. Missionary scholars also produced many excellent grammars and dictionaries, and intensive linguistic research continues, chiefly at universities in Hawaii and New Zealand.

European languages have borrowed many words from Polynesian. Among the English borrowings are taboo (Polynesian *tapu*), tattoo (Tahitian *tatau* "to mark"), mana, ukulele, hula, Kanaka (South Sea Islander), tapa cloth, kava (drink and plant), and the names of many plants and animals native to Pacific regions; *e.g.,* the New Zealand birds kiwi, moa, tui, huia, kaka, and takahe, the tuatara lizard, the kauri family of trees, and many others.

LANGUAGES OF MICRONESIA

The widely dispersed groups of small islands comprising Micronesia lie to the north of Melanesia, between the Philippines and the Polynesian Triangle. Micronesia has received linguistic infusions from each of the surrounding regions. Of the 13 or so languages native to Micronesia, two are intruders from the Philippines–Indonesia region, and two are Polynesian. The rest have probably been there longer but almost certainly derive ultimately from some part of Melanesia. Most or all of this last group fall into a single subgroup of Oceanic that has been called Nuclear Micronesian.

The two Indonesian-type languages are Chamorro, spoken in the Mariana Islands, and Palauan, spoken on Palau in the western Carolines. Chamorro, with about 63,000 speakers, most of them living on Guam, shows close resemblances to Philippine languages. The Palauan-speaking community numbers about 13,000. Palauan is Hesperonesian, but its immediate affiliations are uncertain. The two Polynesian languages in Micronesia are Nukuoro and Kapingamarangi; they are spoken on atolls lying south of Truk and Ponape by about 500 and 1,300 people respectively.

Nuclear Micronesian languages

Nuclear Micronesian languages include Marshallese (Marshall Islands, 30,000 speakers), Gilbertese (Gilbert Islands, 56,000 speakers), Trukese (Hall Islands, Truk, Mortlock Islands, 31,000 speakers), Ponapean (20,000 speakers on Ponape, Ngatik, Mokil, and Pingelap), Kosraean (Kosrae Island, 4,700 speakers), Carolinean (4,000 speakers, with a migrant community on Saipan in the Marianas and a source community in the eastern Carolines speaking a number of fairly diversified dialects), and Ulithian (originally applied to the dialect of Ulithi Island, but now used for the language or dialect continuum that extends from Tobi and Sonsoral through Ulithi and Woleai to Lamotrek, in the western Carolines). In addition, there are two languages that are possibly Nuclear Micronesian but their divergent structure and vocabulary make their membership doubtful at present. Yapese, spoken on Yap Island by about 5,800 speakers, is one of them, and Nauruan, with about 6,000 speakers on Nauru Island, is the other.

Nuclear Micronesian languages show a variability in vocabulary and structure exceeding that of the Polynesian group, thus indicating that their immediate common ancestor began to divide into several languages at least 3,000 years ago. Gilbertese and Marshallese, for instance, share only about 21 percent of related words in their basic vocabularies, while Kosraean and its neighbouring languages, Ponapean and Marshallese, share only about 26 percent and 24 percent, respectively. The figures for Yapese and Nauruan and other languages are still lower.

The greatest diversity lies in east Micronesia, suggesting that the homeland of Proto-Nuclear Micronesian may have been in that region. Further west, the languages are more alike; a continuum of mutually intelligible dialects or very closely related languages extends from Truk through the main body of the Caroline Islands southwest to Tobi Island, and includes the Trukese, Carolinean, and Ulithian languages.

The following features appear to be characteristic of Nuclear Micronesian languages (with the partial exception of

Gilbertese): velarized consonants (made with movement of the tongue toward the soft palate), long or geminate consonants, the assimilation of vowel pronunciations to neighbouring sounds, and a number of other fairly complex phenomena that modify the pronunciation of particular sounds in particular contexts. Grammatical features include verb phrases introduced by subject prefixes marking person and number, attachable to the verb or to one of the many preverbal tense-aspect particles; many subclasses of nouns distinguished in numerical and possessive constructions; elaborate sets of demonstrative pronouns paralleling many of the meaning distinctions found in personal pronouns; complex numeral and classifier systems; and several kinds of reduplication of base words carrying several grammatical–semantic functions.

In general, the Nuclear Micronesian languages have been among the most innovative of the Oceanic languages. They have considerably modified the original Oceanic sound system and lost a number of grammatical distinctions made in Proto-Oceanic, while elaborating others and developing several new features. Grammatical sketches exist for most of the languages, but as yet few dictionaries have been published. Micronesian communities are almost 100 percent literate. Because of the phonological complexity of many of the languages, however, few of the orthographic systems, all of which are in roman script, are as uniform or satisfactory as those for the Polynesian or Indonesian languages.

(A.K.Pa.)

PAPUAN LANGUAGES

The Papuan languages are those languages spoken in an area centred upon New Guinea and extending from the islands of Alor, Halmahera, and Timor in the west to the Santa Cruz Archipelago in the east. The group includes approximately 740 languages, used by about 3,000,000 speakers. The term Papuan was originally employed merely to distinguish these languages from the Austronesian (Malayo-Polynesian) and Australian languages, and, until recently, most Papuan languages were believed to be unrelated to each other. Intensive research by teams of linguists since the late 1950s, however, has resulted in a revolutionary change in the Papuan linguistic picture, and it is now known that about 350, and perhaps even as many as 450, of the approximately 740 identified Papuan languages are related. More than 500 of these are spoken by 2,900,000 speakers, who occupy almost three-quarters of the New Guinea mainland. Belonging to 76 to 105 language families (depending on the classification used), these languages together form the Central New Guinea macrophylum. (A macrophylum is a group of languages related less closely than those of a language family or stock.)

Recent extensive research into the Papuan languages has resulted in the preliminary classification of most of them. Numerous new languages have been discovered, though large areas, mainly in Irian Jaya, Indonesia, remain unknown. Despite the concentration of research on discovery and classification, a number of grammatical and lexical studies have been prepared, and folklore has been collected. The Summer Institute of Linguistics, an association of Protestant missionaries specializing in studying primitive languages and involved in literacy training and Bible translation, has carried out an extensive native language literacy program in New Guinea with a measure of success.

Classification and distribution. Interrelated language families are grouped in phyla, groups of languages more distantly related than those of a family or stock but more certainly or closely related than those of a macrophylum. Apart from the isolates, or unrelated languages, the number of known Papuan phyla is 21. Of these, 8 are small and consist of only 2 to 6 languages each. Of the remaining 13, 6 constitute the central New Guinea macrophylum, and 3 more can tentatively be included in it.

Central New Guinea macrophylum

The central New Guinea macrophylum comprises: (1) the East New Guinea Highlands phylum, found predominantly in the Highlands and Chimbu provinces of Papua New Guinea; (2) the Finisterre-Huon phylum in the Morobe and Madang provinces, Papua New Guinea; (3) the Central and South New Guinea phylum, located mainly in the Gulf and Western provinces of Papua New Guinea and in southern and northeastern Irian Jaya; (4) the West New Guinea Highlands phylum of the highlands area of Irian Jaya; (5) the South-East New Guinea phylum, in the Central, Northern, Morobe, and Milne Bay provinces of Papua New Guinea; (6) the Madang phylum in the Madang provinces of Papua New Guinea.

Tentative member phyla of the Central New Guinea macrophylum are: (1) the Adelbert Range phylum in the Madang province of Papua New Guinea; (2) the Middle Sepik-Upper Sepik-Sepik Hill phylum, which could also be classified as constituting three separate phyla; (3) the Anga stock (stocks are intermediate between phyla and families) spoken in the Eastern Highlands, Morobe, and Gulf provinces of Papua New Guinea.

Large Papuan phyla that are not members of the Central New Guinea macrophylum are the Ramu phylum in the Madang and East Sepik provinces of Papua New Guinea; the Torricelli phylum in the same provinces; the West Papuan phylum on the Doberai (Vogelkop) Peninsula of Irian Jaya, and in northern Halmahera, eastern Timor, and Alor; and the Bougainville phylum on the island of Bougainville. Small isolated phyla are located mainly in the West Sepik and Gulf provinces of Papua New Guinea and in the Santa Cruz Archipelago. The highest concentration of known isolates is also in these provinces and in the East Sepik Province, the New Britain-New Ireland area, and the Solomon Islands.

Some Papuan languages are difficult to classify because of strong Austronesian influence upon them. Most Papuan languages are of only regional importance, but a few have achieved some cultural significance outside their immediate area because of their use as missionary languages.

Of the known Papuan languages that cannot yet be linked with the macrophylum, about 130 belong to large groups, and 26 to small individual groups. Approximately 50 languages, called isolates, seem to be unrelated either to each other or to the established groups. Further research may well show that additional Papuan languages belong to the macrophylum, especially after hitherto linguistically unknown areas in New Guinea have been studied.

Enga, the numerically largest language

Each of the individual Papuan languages is, for the most part, spoken by only a few hundred to a few thousand people, though the numerically largest one, Enga, in Enga Province of Papua New Guinea, has more than 150,000 speakers, and several other languages are each spoken by tens of thousands. Even if related, the languages generally show considerable diversity, especially in vocabulary. A few basic grammatical characteristics are, however, shared by many languages. In numerous instances it is difficult to determine the border line between the languages and dialects, despite the presence of marked differences between two forms of speech.

Linguistic characteristics. Most Papuan languages show extreme grammatical complexity. Their verbs vary to reflect a wide range of numbers and other features of the subject as well as of the direct and indirect objects and the beneficiary. For example, in Kiwai, a language of the Central and South New Guinea phylum, the verb *ai-ni-mi-bi-du-mo-iauri-ama-ri-go* means "they three will certainly see us two." Similarly, in the Monumbo language of the Torricelli phylum *mbepe*$_1$*-nge*$_2$ *tsi*$_3$*-p*$_4$*-ing*$_5$*-em*$_6$ can be translated as "you gave him a taro." Literally, the parts of this utterance are: taro$_1$–singular$_2$, 2nd person singular subject (you)$_3$–give, past form$_4$–masculine class 3rd person singular indirect object (to him)$_5$–plant class singular direct object$_6$. Verbs also indicate tense, aspect, mood, and the direction and circumstances in which the action they designate is performed; *e.g.,* in Gadsup, of the East New Guinea Highlands phylum, *kùmù-`ə-nk-`əd ´əd-òn-ték-´əp-*

ón-ì-nó-ké, "had he indeed wanted to go down for him?", has several prefixes and suffixes indicating emphasis, tense, direction, and the question status. (The accent marks indicate tones; the ə is a sound pronounced like *a* in "sofa." The dashes separate the various components of the utterances, but do not normally appear in nonlinguistic texts).

Verb forms for one or more actions

There are, basically, two major types of verb forms in many Papuan languages. One, which can be referred to as the normal type, is found in sentences in which only one action is referred to; the other, which may be referred to as the special verb form, occurs in sentences in which more than one action is mentioned. In the latter type of sentence, which may have one or more special verb forms, the normal verb form also appears with the last verb in the sentence.

Numerous Papuan languages have gender and noun class systems, some with up to ten or more classes. Bewilderingly complex variations of adjectives, numerals, demonstratives, and subject and object markers often result, since these words have special forms for each of the various classes of nouns. An example is the sentence *ame akwum kuvambakwum sumupar amenakwum salikəmba,* "I saw my two big women," from Angoram of one of the small Sepik phyla. If "women," *akwum,* is replaced by "arrows," "gardens," or "frogs," etc., the entire sentence, rather than a single word, is changed: when "arrows" is substituted, the phrase becomes *ame pwanggli kəpanggli klupar amenakanggli salikəanggliya;* and when "gardens" is used, it changes to *ame konggəmbər kəvambər pələpar amenkəmbər salikəmbəra.*

Many languages are tonal, with changes of pitch in words and syllables that affect the meaning of the words. The interaction of Papuan tonal systems with patterns of stress and syllable length can be extremely intricate. (S.A.W.)

AUSTRALIAN ABORIGINAL LANGUAGES

There are approximately 260 Australian Aboriginal languages. This group of genetically interrelated tongues embraces the entire Australian continent as well as the western islands of Torres Strait, but apparently excludes Tasmania. The languages are characterized by great similarities in their sound systems and considerable agreement in grammar but often by markedly few similarities in vocabulary. Intelligibility between neighbouring forms of speech is common, and dialect chains stretching over amazing distances occur, though the two extremes of such a chain seem to be quite distinct languages.

Secret languages and special vocabularies

Every tribe speaks at least one distinct dialect, but bilingualism and multilingualism are common in many areas. Many individual languages have parallel forms, characterized by special vocabularies and sometimes by special sounds that are used in cultural avoidance situations (*e.g.,* to mothers-in-law) or as secret languages among initiated men on certain occasions.

No genetic link is known to exist between the Australian languages and any outside language. It is believed that languages ancestral to the present-day ones were introduced into Australia by peoples that crossed Arnhem Land in northern Australia many millennia ago. With the apparent exception of the influence of Papuan languages on the languages of the Cape York Peninsula, the Australian languages remained free from outside influence until the arrival of European settlers late in the 18th century.

The great majority of the Australian languages were nearing extinction by the third quarter of the 20th century, with about 50 or more extinct, predominantly in the east, south, and west of the continent. Speakers of languages believed extinct for decades are, however, occasionally discovered. Most languages have very few surviving speakers; still-vigorous languages have, for the most part, only a few hundred speakers each, though Mabuiag (the language of the western Torres Strait islands) and the Western Desert language have 8,000 and 4,000 speakers, respectively. About 45,000 Aborigines may still have some knowledge of an Australian language, but accurate figures of the speakers of individual languages are almost impossible to obtain.

Extensive research on the Australian languages has been carried out since 1960, largely through the Australian Institute of Aboriginal Studies in Canberra. The results of this and earlier research have shown the Australian languages to be interrelated and have made it possible to explain their structural differences in terms of a typological development from a simple to a complex structure. In addition, a considerable amount of detailed information on the grammar of numerous languages has been recorded.

Genetic relationship of all Australian languages

Classification. Earlier classifications regarded the northern and northwestern languages, which are structurally rather different from the southern languages, as genetically distinct from the latter. (The interrelationship of the southern languages was recognized quite early.) Later, various classifications based on language type were established, and these demonstrated the ultimate unity of all Australian languages. The most recent classification is based on the degree of lexical (vocabulary) interrelationship between the languages; it subdivides the languages into 28 families, of which 27 are located in the north and northwest, covering about one-eighth of the continent, and a single family is found occupying the remaining seven-eighths of Australia. This skewed picture may be the result of the thorough spreading of a language form referred to as Common Australian (dated at about 5,000–6,000 years ago) from somewhere in northwestern Australia through most of the continent except the north and northwest regions. This diffusion appears to have coincided with that of an archaeologically recognized cultural revolution. The spreading of this Common Australian language may have brought about greater linguistic uniformity in much of Australia. Most of the 28 families are subdividable into groups and often into subgroups of individual languages. In the following list of language families these abbreviations are used: G = group or groups; SG = subgroups; L = language or languages.

Family
Tiwian (1L)
Yiwadjan (4G, 2SG, 5L)
Kakadjuan (1L)
Mangerian (2G, 2L)
Gunavidjian (1L)
Nagaran (1L)
Gunwingguan
(6G, 3SG, 11L)
Bureran (2G, 2L)
Nunggubuyuan (1L)
Andilyaugwan (1L)
Maran (2G, 2SG, 3L)
Mangaraian (1L)
Ngewinan (1L)
Yanyulan (1L)
Djingili-Wambayan
(2G, 3L)
Karawan (2G, 2L)
Minkinan (1L)
Larakian (2G, 2L)
Kungarakanyan (1L)
Warraian (1L)
Daly (3G, 4SG, 10L)
Murinbatan (1L)
Djamindjungan (4L)
Djeragan (2G, 5L)
Bunaban (2G, 2L)
Wororan (3G, 12L)
Nyul-Nyulan (4L)
Pama-Nyungan
(41G, 50SG, 177L)

Some aspects of this classification are only tentative. The locations of the families are shown on the accompanying map.

Bidjandjara

In most areas in which Australian languages are still in daily use, individual languages have gained prominence over others and have become lingua francas (common languages), as a result either of their use as mission languages or of social and cultural factors active among the Aborigines themselves. One of the dialects of the Western Desert language, Bidjandjara (Pitjantjatjara), has become the Aboriginal lingua franca over a sizable portion of the western half of the continent.

Grammar. The Australian languages generally show considerable grammatical complexity. Affixes (word parts added initially, internally, or terminally) play an important role; prefixes used initially and suffixes used terminally are found in northern and northwestern Australia, and suffixes, for the most part, are used elsewhere. A peculiar feature of many languages is the suffixing of markers indicating the subject and object of the verb to

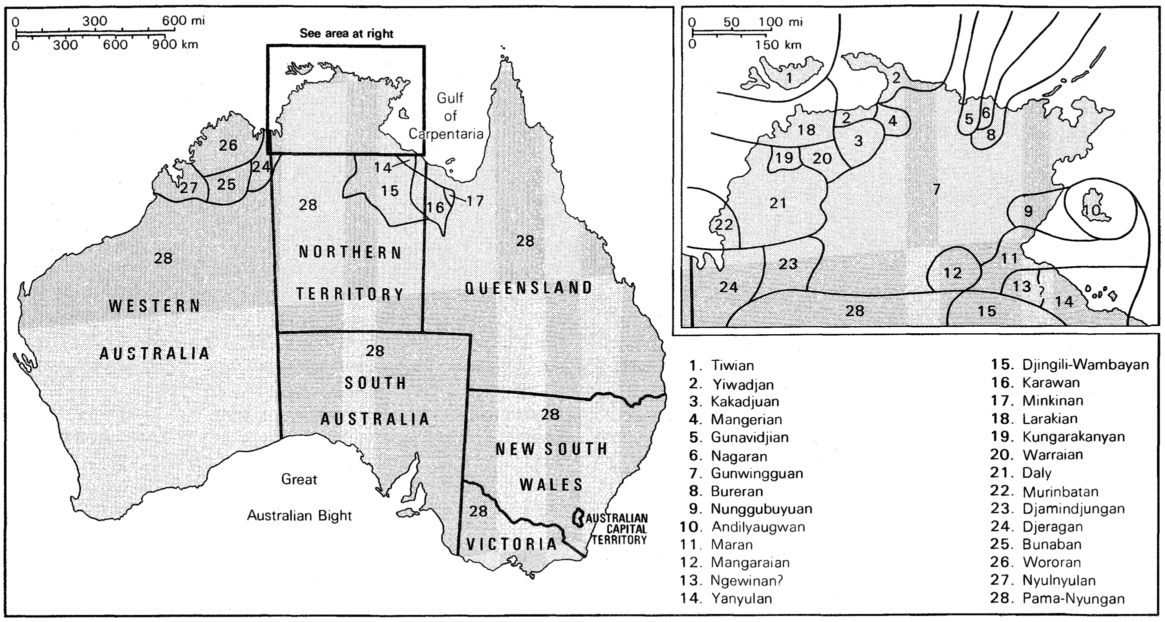


Figure 29: Distribution of the Australian Aboriginal languages.
Adapted from S.A. Warm, *Languages of Australia and Tasmania* (1971); Mounton & Co.

Common Australian words

the first word of the sentence, irrespective of what it is, or to special particles not connected with the verb. An example from Wanman of the Pama-Nyungan family is the use of the suffixes -ŋ*a* and -ŋ*ku* in *paral*^{y}*i-*ŋ*a-*ŋ*ku t*^{y}*i*ŋ*ka-*ŋ*a,* literally, "boomerang-I-you make-past," or "I made a boomerang for you." Another widespread feature is an ergative or agentive suffix, attached to nouns and pronouns, that indicates the actor of an action referred to by a transitive verb. In the Dungidjau language of the Pama-Nyungan family, *t*^{y}*a:n-tu pukin*^{y}*-n*^{y}*a pumi* is directly translated as "*man*-ergative *dog*-object *hit*-past." Here *-tu* is the ergative suffix; attached to *t*^{y}*a:n* ("man"), *-tu* signifies that "man" is the actor of the verb, thus rendering the meaning "the man hit the dog." Great freedom of word order is a feature of many Australian languages. A number of languages, mainly northern ones, have gender and noun class systems, with adjectives, numerals, and demonstratives showing special forms for each of the classes of nouns and often, also, for their number. For example, in Andilyaugwa of the Andilyaugwan family, "where is (located)" is expressed as ŋ*i-*ŋ*ampa na-mpil*^{y}*a* when referring to one man, but as *wunala-*ŋ*ampa wu-pil*^{y}*a* when referring to two men, *wura-*ŋ*ampa na-mpil*^{y}*a* in regard to three or more men, *ta-*ŋ*ampa i*ŋ*a mpil*^{ys}*a* to one woman, and *ma-*ŋ*ampa numa-mpil*^{y}*a* to a ship.

Phonology. The sound systems of the Australian languages are extremely similar. Most of them share from four to six different points of articulation for stop consonants (made with complete stoppage of the breath from the lungs) and nasal consonants (made with the airstream passing through the nose), with many languages having such consonants produced with the tip of the tongue placed between the teeth (interdental consonants), or behind the teeth (alveolar consonants), indicated as *t, n,* or curled up against the hard palate (retroflexed consonants), written as *ṭ, ṇ*. The series of consonants may thus include labial consonants (*p, m*), interdental consonants, alveolar consonants (*t, n*), retroflexed consonants, palatalized consonants (*t*^{y}, *n*^{y}), and velars (*k,* and ŋ as the *ng* in "sing"). In addition, most Australian languages show no distinction between voiced and voiceless stops (such as voiced *b* and voiceless *p* in English), no fricative consonants (*e.g., f, v, s, z*), only three vowels (*a, i, u*), but two or three distinct *r* sounds.

Vocabulary. In spite of the great vocabulary differences among the Australian languages, a number of common words are encountered in a great many languages all over the continent. These are believed to constitute a Common Australian element. In the vocabulary of a given language, various classes of words, such as nouns, verbs, and others, are clearly distinguishable and definable, and word formation is through the use of affixes. Australian languages have contributed to Australian English mainly animal and plant names and objects in nature—kangaroo, wallaby, kookaburra, budgerigar, galah, coolibah tree, billabong.

Collections of mythological and other text materials have been and still are being made in a number of languages. Native literacy in Australian languages is limited but is on the increase as a result of the work of members of the Summer Institute of Linguistics (an association of Protestant missionaries that specializes in studying primitive languages), who, like the people from the Methodist, Anglican, and Australian Inland missions, write the languages in specially adapted versions of the English alphabet.

(S.A.W.)

AFRICAN LANGUAGES

Although no definitive count exists, the number of separate and distinct African languages, as estimated by various authorities, ranges from 800 to more than 1,000. The linguistically most homogeneous area of Africa is the northern part, in which Arabic predominates from Egypt to Mauritania, albeit in a number of sharply distinguished dialects; the most important dialect cleavage is between the Egyptian–Sudanese varieties and those of the Maghrib (from Libya westward). Intermingled with the Maghrib dialects are the Berber languages, concentrated principally in Algeria and Morocco. They range as far east as the Siwa Oasis in western Egypt, as far west as the Senegalese–Mauritanian border area, and as far south as the southern rim of the Sahara. The nomadic Tuaregs speak a Berber language. Nubian, a totally different language, is spoken as far north as southern Egypt along the Nile, but its links are clearly with the languages of sub-Saharan Africa.

Sub-Saharan languages

In sub-Saharan Africa the linguistic picture is far more complex. Except for the Khoisan (Bushman and Hottentot) languages of the extreme south, approximately the entire southern third of Africa is occupied by the relatively closely interrelated Bantu languages. They extend eastward from roughly the Nigerian–Cameroon border to the Indian Ocean. Bantu and non-Bantu languages are considerably interspersed around the northern Bantu border, running through Cameroon, slightly south of the northern boundary of Zaire, and through Uganda and Kenya. There are Bantu enclaves as far north as Somalia and non-Bantu enclaves in northern Tanzania. It is the area north of the Bantu and south of the Sahara that is linguistically the most diverse; there the languages are most numerous and their interrelationships most remote and difficult to establish with certainty.

General considerations

African lingua francas

Because probably not more than 40 African languages have more than 2,000,000 native speakers, a number of lingua francas have developed in various areas as a means of coping with this enormous linguistic diversity. Arabic, in addition to having the largest number of native speakers of any language of the continent, is sometimes used as a lingua franca, in its literary form by educated non-Arab Muslims, and in a number of colloquial varieties by much of the non-Arab population of The Sudan and Chad. Swahili, a Bantu language heavily influenced by Arabic, is official in Tanzania and Kenya and is used as a lingua franca throughout most of East Africa including eastern Congo. Other important lingua francas include Lingala, also a Bantu language, in western Congo; Fanagalo (vulgarly known as "kitchen Kaffir" or "mine Kaffir"), a pidginized form of Zulu with many English and Afrikaans loanwords, in South Africa, particularly in the mines; and Sango, a pidginized form of Ngbandi (included in the Adamawa-Eastern subdivision of Niger-Congo by some scholars; see below) with many French loanwords, spoken in the Central African Republic.

In addition, there are Bambara-Maninka, a Mande language used in Mali, Guinea, and Ivory Coast; Hausa, spoken in northern Nigeria and neighbouring areas; Amharic, used throughout Ethiopia, where it is the official language; Wolof, the language of Dakar, widely spoken throughout Senegal and The Gambia; and Kongo (Kikongo), a Bantu language that is used in the area of the mouth of the Congo River and is also spoken in a pidginized variety known as Kituba.

Standard European languages, particularly English and French and, to a lesser extent, Portuguese and Italian, also serve as lingua francas, generally among the better educated, in areas in which they are or were the colonial languages and where they enjoy or have enjoyed official status. Localized varieties of European languages are also spoken in numerous parts of Africa, such as Pidgin English, a widely used lingua franca of Cameroon and West Africa; Krio, a creolized form of English, differing slightly from Pidgin English and used by the local population of Freetown, Sierra Leone; various forms of Creole Portuguese, spoken in the Cape Verde Islands, the islands of São Tomé and Príncipe, and Guinea-Bissau; and Afrikaans, an outgrowth of Dutch, which serves along with English as an official language of South Africa.

LANGUAGE CLASSIFICATION

Early attempts at classification

Although word lists of a number of sub-Saharan languages were recorded by Europeans as early as the 15th and 16th centuries, and the first grammar of a sub-Saharan language (Kongo) was published in 1659, extensive study of African languages and systematic attempts at grouping and classifying them did not begin until the 19th century. In 1854 Sigismund Koelle, a missionary living in Freetown, Sierra Leone, was able to collect substantial word lists of roughly 150 languages, principally of West Africa but also including some belonging to the Bantu family. He established 11 groups, leaving approximately 40 languages

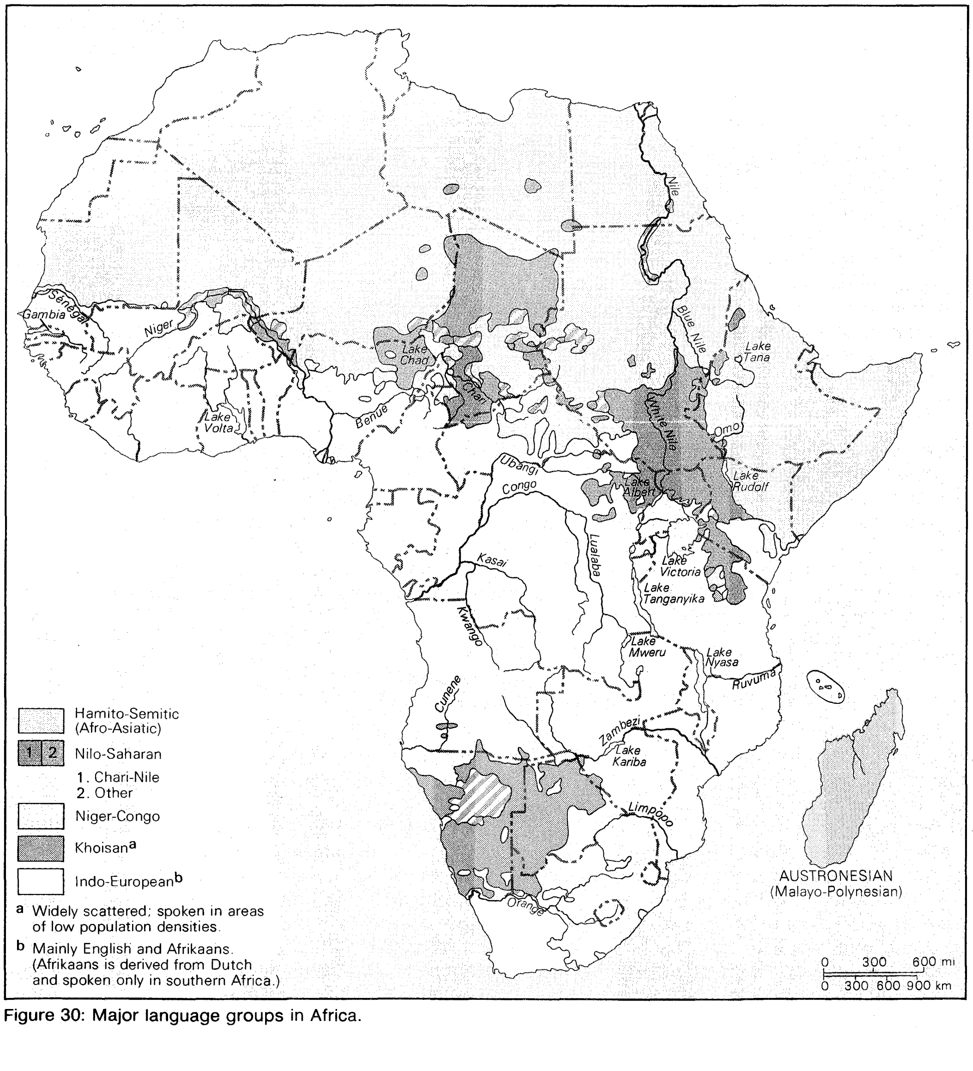


Figure 30: Major language groups in Africa.

unclassified. His groups are generally small, rarely exceeding 15 members, and he made no systematic attempt to explore wider relationships. He laid the groundwork for establishing the Mande family, which has been universally accepted ever since. (The term "family" is used loosely in this article, sometimes referring to large groupings of languages and sometimes referring to subgroups or branches of these groupings. Niger-Congo, for example, is usually called a language family, but its subgroups [*e.g.,* Mande] and its sub-subgroups [*e.g.,* Bantu] are often referred to as families as well.)

The unity of the Bantu family, although suggested by a number of earlier writers, was first established in 1862 by the German scholar Wilhelm Bleek, who gave it its name, the word for "people" in most languages of the family. During approximately the same period, the work of several scholars suggested that Ancient Egyptian, the Berber languages, certain languages of northeastern Africa called Cushitic (*e.g.,* Galla, Somali, Beja, and Afar [Danakil]), and perhaps others (*e.g.,* Hausa) were remotely related to one another and to the Semitic family. These languages were called Hamitic at the suggestion of a French Semitist, Ernest Renan, who was also the first to propose the name Cushitic, because the peoples speaking these languages coincided roughly with the descendants of Ham (one of whom was Cush), according to the biblical account in Genesis, just as the people speaking the Semitic languages coincided with the descendants of Shem (Sem).

The first attempts at an overall classification of African languages were those made by the Austrian linguist Friedrich Müller in 1876–88 and by the German Egyptologist Karl Richard Lepsius in 1880. Müller, who classified all the world's languages on the basis of a correlation of linguistic and racial groupings (the latter determined largely by hair type), set up the following divisions within Africa: (1) Bushman-Hottentot, as a subdivision of the languages of the tufted-haired peoples (the Papuan languages of New Guinea formed the other subdivision); (2) Bantu and (3) Negro, both subdivisions of the languages of the fleece-haired (*vliesshaarig*) peoples; (4) Hamitic, (5) Semitic, and (6) Nuba-Fulah, all three subdivisions of the languages of the wavy-haired (*lockenhaarig*) peoples. Linguistically, Nuba-Fulah was completely heterogeneous and was abandoned by later scholars.

Lepsius' classification

Lepsius proposed an alternative classification of the languages and peoples of Africa, of which the three major divisions were (1) Semitic; (2) Hamitic, including Hausa, which he grouped with the Berber or Libyan branch, and Hottentot, neither of which had been classed as Hamitic by Müller; and (3) the Negro languages, consisting of two subdivisions, Bantu and Mixed Negro, composed of many subgroups, and containing all the indigenous languages of Africa not included in the other divisions.

During the early part of the 20th century, a revision of this classification emerged from the work of the two leading German Africanists, Carl Meinhof and Dietrich Westermann. Meinhof added a number of languages to Lepsius' Hamitic grouping, namely Fulani, a major West African language, and a group of East African languages, including Masai and Bari. Westermann attempted to show the unity of virtually all the non-Bantu, non-Hamitic, and non-Semitic languages of Africa; he designated them Sudanic. Thus, a threefold division was established: Bantu, Hamitic, and Sudanic. The Bushman languages of southern Africa were sometimes treated as a separate group and sometimes linked with Sudanic, and Semitic was generally linked to Hamitic, though not included within it. Not everyone accepted this classification. French scholars, in particular Maurice Delafosse and Lilias Homburger, restricted Hamitic to Ancient Egyptian, the Berber languages of North Africa, and the more heterogeneous Cushitic group of languages of northeastern Africa. They connected Hottentot with the Bushman languages while grouping the other languages classed by Meinhof as Hamitic in the very general Negro-African or Sudano-Guinean family, to which they also believed the Bantu languages to be affiliated.

Westermann himself revised his views of the earlier classification scheme and came to regard the great majority of the languages of West Africa (*i.e.,* the Western Sudanic languages) as more closely related to Bantu than to the languages farther east, previously regarded as Sudanic. Sir Harry Johnston, a British Africanist, likewise pointed out similarities between Bantu and many West African languages that he termed Semi-Bantu.

Joseph Greenberg's classifications

No single scheme was generally accepted in 1955, when Joseph Greenberg, a United States linguist, presented the earlier version of his major classification of the languages of Africa. In 1963, Greenberg made public a substantially revised version of his work that consolidated previously unclassified languages and smaller language families into four large linguistic stocks—Niger-Kordofanian, Nilo-Saharan, Khoisan, and Afro-Asiatic (Hamito-Semitic). They included all the indigenous languages of Africa except Malagasy, the language of Madagascar, which belongs to the Austronesian (Malayo-Polynesian) family and is thus unrelated to the languages of the continent. The Niger-Kordofanian group consists of two branches, Niger-Congo, which is very extensive, and the much smaller Kordofanian group, so named because all of its members are located in Northern and Southern Kurdufān provinces, The Sudan. Kordofanian contains a number of subgroups, of which Kadugli-Korongo is most distantly related to the others. In fact, its inclusion within the group is open to serious question. Niger-Congo consists of Westermann's Western Sudanic group plus Bantu, whose interrelationship had earlier been advocated by Westermann. Greenberg differed from Westermann, however, in that he treated Bantu as a subgroup of one of the branches of Western Sudanic rather than an independent unit related to Western Sudanic as a whole. Greenberg also argued strongly for the inclusion of Fulani in this family. (Fulani had been regarded as Hamitic by Meinhof, but some French scholars had already advocated its relationship to other West African languages, such as Serer of Senegal.) Furthermore, Greenberg extended Niger-Congo eastward to include languages of Central Africa, such as Ngbandi of the Central African Republic and Azande (Zande) of Zaire, with which Westermann had not dealt.

The Nilo-Saharan group consists of six branches, of which Chari-Nile is by far the most heterogeneous. This family is the least adequately substantiated of the four. The Khoisan languages comprise the Bushman and Hottentot languages of southern Africa and two languages of Tanzania, Sandawe and Hadza (Hatsa). Afro-Asiatic, which is Greenberg's name for Hamito-Semitic, consists of five branches: (1) Semitic, represented in Africa by Arabic and the Ethiopic languages, of which the most important are Amharic, the official language of Ethiopia; Tigrinya, the predominant language of Eritrea; and Geʿez, the liturgical language of the Ethiopian Church, which is no longer spoken; (2) Ancient Egyptian, now extinct, and its daughter language, Coptic, the liturgical language of the Coptic Church of Egypt, (3) the Berber languages of northern Africa, including Kabyle of Algeria, Riffian and Shluh (Shilha) of Morocco, and Tuareg, or Tamashek, of the Sahara; (4) the extensive and diversified Cushitic group of northeastern Africa, whose most important members are Somali and Galla; (5) Chadic, spoken in Chad, Cameroon, and northern Nigeria, although Hausa, its best known and most widely spoken member, is used throughout much of West Africa.

Branches of Afro-Asiatic (Hamito-Semitic)

Hamito-Semitic languages also are spoken in the Middle East. For this reason, they are treated in a separate section (see *Hamito-Semitic languages*).

Because of the inexact nature of the procedures for establishing genetic classification and also because of incomplete data on the majority of African tongues, differences of opinion exist on various aspects of Greenberg's classification. Some scholars, such as the Hungarian linguist István Fodor and the English Bantuist Malcolm Guthrie, insist that genetic interrelationship of a group of languages can be established only by demonstrating regular phonological correspondences among them, as have been shown among the Bantu languages, for example. Greenberg and his supporters, however, maintain that similarities in both sound and meaning between two or more languages, which are too numerous to be dismissed as mere coincidence and are found in that part of language least susceptible to borrowing (*e.g.,* grammatical elements, body parts and functions, common natural objects and phenomena, and lower numerals), must be attributed to genetic relationship. Nevertheless, there are no rigorous procedures for determining how similar two items must be in sound and meaning to be considered evidence for genetic relationship or how many items are needed to rule out other explanations.

SIMILARITIES AND CORRELATIONS

Common features of the African languages. Although African languages are extremely diverse in structure, certain phonological and grammatical features are widespread throughout much of the continent or at least within extensive and well-defined areas.

Tones and clicks

Phonological features. Tone—*i.e.,* the use of pitch to distinguish words and grammatical forms that otherwise would be phonetically identical—occurs in the overwhelming majority of sub-Saharan languages. The famous click sounds are largely restricted to two language groups of southern Africa: the Bushman and Hottentot (Khoisan) languages and the southern Bantu languages (*e.g.,* Zulu, Xhosa, and Sotho), which borrowed them from the Bushman- and Hottentot-speaking peoples. Clicks are also common to three East African languages, Sandawe, Hadza, and Sanye, but nowhere else on the continent or outside it. The consonants *kp* and *gb,* called coarticulated labiovelar consonants, are, in Greenberg's words,

> distributed over a wide belt of languages of diverse genetic affiliations from the Atlantic Ocean almost to the Nile Valley. Outside of Africa these sounds are only known from a restricted area of Melanesia. . . .

Other phonological features common in Africa but rare in the rest of the world are word-initial combinations of a nasal sound plus another consonant (*e.g., mb* or *nd*), and implosive consonants.

Grammatical features. In some African languages, words generally consist of a root only, with grammatical affixes rare and even nonexistent in a few languages. These are often referred to as isolating languages. Other African languages have words that are composed of many elements. Those languages are subdivided into the agglutinative languages (in which each element of a word has a distinct and separate meaning and form) and inflectional languages (in which the various elements of meaning may be fused into such forms as prefixes or suffixes or the entire word may change internally to indicate grammatical relationships).

In some of these languages, a single word might be equivalent to a whole sentence in English or other European languages. Thus, in Swahili, a Bantu language, "we will not hit him" is expressed in a single word, *hatutampiga,* in which *ha* is the negative marker, *tu* the verb subject "we," *ta* the future-tense marker, *m* the verb object "him," and *piga* the verb "hit." The verb likewise may undergo modifications in meaning through the addition of suffixes; thus *pigwa* means "(to) be hit," and *pigana* "(to) hit one another."

Noun classes and gender

A widespread feature among African languages employing words composed of more than one element is a system of noun classes (see below *Benue-Congo subgroup*), particularly marked in the Bantu family but also occurring in a wide variety of West African languages, as well as in certain languages of the Northern and Southern Kurdufān provinces, The Sudan. All of these languages have been grouped by Greenberg in his Niger-Kordofanian family, although others in this family lack this feature altogether.

Grammatical sex gender is not nearly as widespread as noun classes. It is characteristic of the Semitic and Hamitic languages (although lacking in some members of the Chadic branch) and of other languages such as Hottentot and the closely related Naron Bushman language of South Africa and of the much more distantly related Sandawe and Hadza languages of Tanzania. Gender is likewise found in East Nilotic but not South Nilotic (these two comprise what was earlier known as "Nilo-Hamitic") or West Nilotic (previously known simply as Nilotic). Gender plays an even more restricted role in other African languages, such as Bongo, a language of the southwestern Sudan in which different pronouns are used for masculine and feminine 3rd person singular. The great majority of African languages, however, do not make this distinction, but many of them do distinguish between animate and inanimate personal pronouns.

Linguistic, racial, and cultural interrelations. It is difficult to find strict correlations between linguistic, cultural, and racial groupings, either in Africa or elsewhere in the world. Groups speaking identical or closely related languages may be strikingly different racially, culturally, or both; while groups that are very similar racially and culturally may speak different and even unrelated languages. Thus the Damara (or Bergdama) of South West Africa/Namibia, who speak the language of the Nama Hottentot, are sharply distinguished from the latter both racially and culturally. The Damara have much darker pigmentation and other more Negroid physical features and are predominantly nonpastoralists. Likewise, the Arusha of northern Tanzania, who speak the language of the pastoralist Masai, are horticultural (*i.e.,* agricultural without use of the plow). On the other hand, the forest-dwelling Pygmies, who are spread over much of central Africa, form a distinct racial and cultural group and are exclusively hunters and food-gatherers, yet they have no distinctive language. Instead, they speak the languages of their various neighbours, such as Bantu, Moru-Mangbetu (grouped by Greenberg in the Central Sudanic branch of Chari-Nile), and reportedly Sere-Mundu (grouped by Greenberg in the Adamawa-Eastern branch of Niger-Congo).

Broad correlations between race, language, culture

Some broad correlations can nevertheless be made between language groupings and the variables of race and culture, as long as these are not interpreted too rigidly. The Bushmen of southern Africa form a distinct racial, cultural, and linguistic group, and, though the Hottentot are akin to them linguistically and racially, they show important cultural differences, most notably that the Hottentot are pastoralists, not hunters and gatherers. The Berber-speaking peoples of North Africa likewise form a relatively homogeneous racial, cultural, and linguistic group; they are Caucasoid and pursue a mixed agricultural and pastoral way of life. Only the Sahara-dwelling Tuareg are exclusively pastoral nomads and caravaners. The Cushitic peoples of northeastern Africa, who like the Berbers speak languages of the Hamito-Semitic family, represent varying degrees of Caucasoid and Negroid elements. Some of the more prominent among them, such as the Beja or Bedawiye of The Sudan and the Saho-Afar and Somali of the Horn, are nomadic pastoralists. The Galla of southern Ethiopia, the most numerous of all Cushitic peoples, include some pure pastoralists, but the great majority practice agriculture and husbandry. Their neighbours to the north, the Agau of central Ethiopia, do likewise, though they have largely assimilated culturally to their Semitic-speaking neighbours, while retaining their Cushitic form of speech.

The unusually tall and slender Nilotes of The Sudan, such as the Dinka and Nuer, are sometimes considered a subrace of the Negroid; but it has also been suggested that their physical peculiarities are at least to some extent the result of differences in diet rather than heredity. Herding of cattle and other livestock is predominant among the majority of Nilotic peoples (including the socalled "Nilo-Hamites"), but nearly all practice some horticulture, with the striking exception of certain groups of Masai. Certain Nilotic peoples, however, are predominantly sedentary horticulturalists, such as the Anuak of The Sudan and the Luo, principally in Kenya, who are the most numerous of the Nilotic peoples.

The overwhelming majority of the Bantu and the peoples of western and central Africa are sedentary horticulturalists. Likewise, most of these peoples belong linguistically to Greenberg's Niger-Congo family, although some fall within Nilo-Saharan, and others within the Chadic branch of Hamito-Semitic; furthermore, they comprise the overwhelming majority of the Negro race, which, however, is no more homogeneous a physical group than the peoples

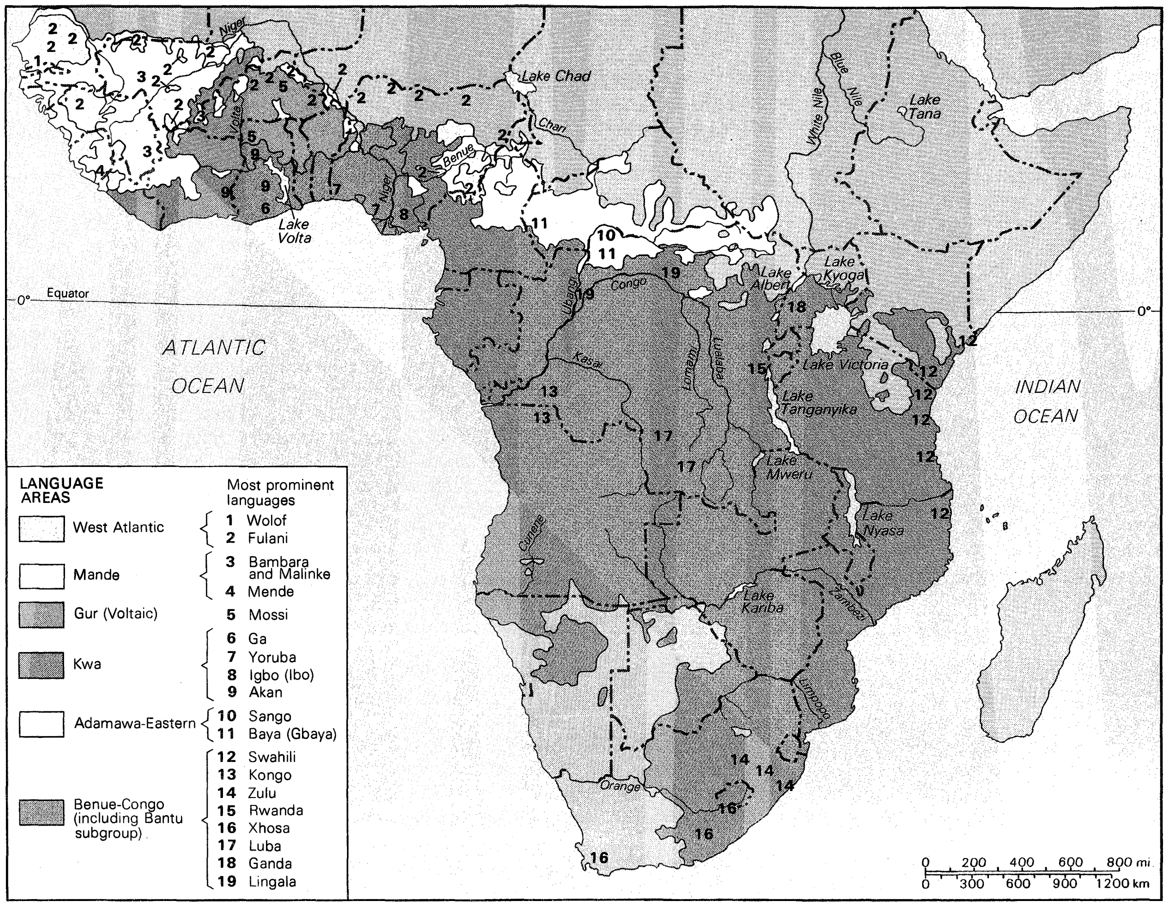


Figure 31: Distribution of the Niger-Congo languages.

West Africa are the cattle-herding Fulani, who are widely distributed over the savanna belt just south of the Sahara. They are physically as well as culturally distinct from their neighbours, but linguistically their closest affinities are with the sedentary horticultural Negroes of Senegal. Cattle raising likewise plays an important role among certain Bantu peoples of eastern and southern Africa, such as the Ganda, Nkole (Ankole), Sotho, and Nguni peoples. Among some of the East African Bantu (*e.g.,* the Nkole, Rwanda, and Rundi) cattle ownership has been the prerogative of a privileged group probably of Nilotic origin and known by various names, such as Hima (Panda) and Tutsi; these people are physically and culturally distinct from their agricultural neighbours, although totally assimilated to them linguistically. (M.F.Go.)

Niger-Congo languages

The Niger-Congo family is the largest family of languages in sub-Saharan Africa, with 890 known member languages. All of them are considered to be distinct languages and not simply dialects. The named dialects of these languages number in the many thousands; if the variant names for the same language or for the same dialect are to be counted, the figure must be further expanded. For example, for Swahili, there are 17 named separate dialects, 15 additional variant names for some of the dialects, and at least four "nicknames" for popularized or debased forms of the language.

The classification of such an enormous and diverse family was first accomplished in 1949 by Joseph Greenberg, who demonstrated the genetic unity of this group on the basis of an examination of documentation of all the languages of Africa. The primary sources for such a massive comparison consist of published descriptions or vocabularies of tribal languages made by missionaries and colonial officers; these were written in European languages because practically none of the African languages had any written tradition of their own. Usable documents stretch back for only about 125 years, and there are still a number of indigenous languages that either have not been reported at all or lack sufficient detail or accessibility to be included in any final count of Niger-Congo languages. As a consequence, the figure given here may have to be revised upward, especially as there is no evidence that the indigenous languages of sub-Saharan Africa are dying out. Quite to the contrary, of a list of such languages collected during the mid-19th century, only one is known to have become extinct.

Location of the Niger-Congo languages

The area in which the Niger-Congo languages are spoken is fairly easily demarcated (see map) and includes the entire sweep of sub-Saharan Africa with the exception of the Horn of Ethiopia. All indigenous peoples of sub-Saharan Africa west of the Nile are speakers of Niger-Congo languages with the exception of a few to the extreme south (Bushman-Hottentot peoples) and a more substantial number along the northern rim of the area (speakers of Chadic or Nilo-Saharan languages).

A conservative estimate of the total number of speakers of Niger-Congo languages is 300,000,000. Such a figure is obtained by adding the populations (as of the early 1980s) of the 32 countries of tropical and southern Africa and then subtracting the number of European and Asian speakers reported for South Africa. (It is assumed that the number of speakers of Niger-Congo languages in the countries to the north of the tropical belt omitted by this method—especially in Mali and Upper Volta—would be balanced out by the number of indigenous non-Niger-Congo speakers enumerated in the included populations—especially in Nigeria and Uganda.)

CLASSIFICATION

The Niger-Congo family is divided into six genetic subgroups. Viewed on a geographical continuum from northwest to southeast, these are:

1. West Atlantic, 43 languages
2. Mande, 26 languages

3. Voltaic (Gur), 79 languages
4. Kwa, 73 languages
5. Adamawa-Eastern, 112 languages
6. Benue-Congo, 557 languages

Niger-Kordofanian

This subclassification results from very fundamental and very old divisions in the language family, and each subgroup represents the results of separate developments that can be estimated to be about 5,000 to 10,000 years old at the least. Obviously, such a deep division has yielded widely separated languages—separate not only in geographic distribution but in elements of the sound systems, grammars, and vocabularies as well. It is surely safe to say that there is no one feature common to the entire group and that the recognition of their relationship depends upon a truly global view of the evolution of a few separate ancestral languages into the thousands of descendant languages spoken throughout the world today. Because this linguistic evolution is neither random nor chaotic, it is possible by means of a very careful plotting of common clusters of features to establish the approximate degrees of relationship even when neither the intermediate nor earliest stages of the protolanguages concerned have been reconstructed. Greenberg used this method to include the nearest relatives of the Niger-Congo languages with them in a superfamily that he designates as "Niger-Kordofanian." This grouping contains two branches: all the languages here considered, namely the Niger-Congo family, and a coordinate branch of the Kordofanian family of languages, spoken in the Nuba Hills (Jibāl an-Nūbah) of Southern Kurdufān province, The Sudan.

Table 58: Niger-Congo Languages with More Than Five Million Speakers

	subgroup	estimated number of speakers*
Fulani (Fulbe)	West Atlantic	11,500,000
Yoruba	Kwa	18,100,000
Igbo (Ibo)	Kwa	14,700,000
Akan group	Kwa	8,600,000
Kongo	Benue–Congo	6,300,000
Zulu	Benue–Congo	5,500,000
Rwanda (Banyaruanda)	Benue–Congo	5,200,000
Xhosa (Xosa)	Benue–Congo	5,300,000
Luba	Benue–Congo	5,100,000
Shona	Benue–Congo	5,400,000

*Figures rounded off to the nearest 100,000.

West Atlantic subgroup. The centre of gravity of the 43 West Atlantic languages would seem to lie in the Senegal–Guinea area, in which are found the two most important languages of the subgroup—Wolof and Fulani (the latter of which is also known as Fulbe, Fula, and Peul). The area is also the locus of the great majority of the other languages in the subgroup. Fulani, the language of a highly mobile pastoralist people numbering as many as 11,500,000, is also found in a scattered distribution as far east as Chad. The immense extension of this language, together with its association with cattle and Islām, led earlier researchers to give it a special place apart from the other West Atlantic languages—in fact, outside what is now recognized as Niger-Congo. It was first isolated in 1883 by an Englishman, Robert Needham Cust, who based his work on that of Friedrich Müller, an Austrian linguist. The German Karl Richard Lepsius classified it as one of the "mixed Negro" languages not fully sub-Saharan or Bantu (1880). Subsequently, it was included as a "Hamitic" language in the German Africanist Carl Meinhof's influential classification (1912), which was adopted by many scholars writing in English.

A feature of Fulani that is shared systematically with some of the other West Atlantic languages is an "alternation" whereby both the beginnings and the endings of words go through parallel changes according to grammatical considerations. This feature is found in its greatest elaboration only in Fulani; it is represented in either vestigial or undeveloped form in most of the other West Atlantic languages. This fact made Fulani appear more different from the others than it actually is and contributed to its misclassification. It is of interest to speculate why this most "vigorous" of the West Atlantic languages should be the most difficult or complex in its grammatical system. ("Vigorous" is used here in the sense that it has several millions of speakers, and also in the sense that it is associated with a dynamic expansionism). By comparison, Latin indicated grammatical case, gender, and number only by the endings of words, and English has virtually eliminated case and gender and has largely simplified number to a single suffix.

Bambara and Mende languages

Mande subgroup. The Mande subgroup includes 26 languages, of which the most prominent are Bambara of the Mali–Guinea area and Mende of Sierra Leone. In contrast to the classification of the West Atlantic subgroup, there has been little doubt as to the identification of the languages in the Mande subgroup. This certainty results from the subgrouping conforming closely to the "classic" expectation of language differentiation, in which a small number of contemporary languages are clearly seen to share a number of striking features not found elsewhere. Such features presumably provide evidence of early segregation of the ancestor language of the subgroup and of a continued effective isolation. Examples of such key features in the Mande languages are the systematic distinction between free and dependent nouns (see below) and the form and use of numerals.

As would be expected, however, the results of a longer period of independent development have also deflected the attention of many researchers to the differences of Mande from the Niger-Congo group—*i.e.,* to the common negative evidence of relationship with the rest of the Niger-Congo family. This negative evidence led Maurice Delafosse, in 1924, to propose two criteria for the identification of this subgroup: (1) a complete absence of the noun-class system that characterizes other West African languages and (2) a complete absence of tonal contrasts. Investigators who view Mande languages in terms appropriate to the group itself have shown that this is a prejudiced view and that a more positive formulation of these criteria is possible. Such a restatement would run somewhat as follows: (1) Instead of the particular distinctions among nouns reflected in the grammatical classes of the other subgroups of Niger-Congo, the Mande subgroup has elaborated a grammatical distinction between free and dependent nouns—nouns referring to "alienable" or "transferable" possessions, as opposed to "inalienable" possessions such as kinsmen or parts of the body. (2) The role of tone is tied to features other than those of semantic reference (*i.e.,* the meaning of the word taken in isolation).

Role of tone in the Mande languages

Whereas Chinese and other tone languages use changes of pitch to distinguish meaning in words with the same configuration of consonants and vowels (when pronounced by themselves), the Mande languages reveal a different use of tone, one that also differs from practically all the other Niger-Congo languages. In the Mande languages, a difference in meaning based on tone seems to demand a contrast of complete utterances, which are usually made up of more than one word. This contrast of entire utterances places Mande at one end of a continuum of distinctions based on tone, because linguists are still finding many variations in the situation of "meaningful semantic tone" or "contrastive tone patterns in isolation" that seem to demand some sort of context in order to be operative. In this respect, it is of interest that most of the differences in meaning between complete utterances that are identical except for tone in the Mande languages are more of a grammatical nature than of a semantic one. It must be stressed, however, that the Mande languages *are* tone languages and that the contrasts so produced are discrete and absolute and thus not of the "sliding" nature of speech melody. In form, they are closer to the question intonation of English, but in function they perform grammatical work internal to the sentence rather than coextensive with it.

It is striking that several independent writing systems based on the syllable (which is the unit that bears the tone) have been reported for Mande languages. the best known of these is the one devised for the Vai language, although Mende, Toma (Loma), and Kpelle also have indigenous syllabic writing systems. This type of writing is somewhat reminiscent of the writing system of Chinese (in which the syllable is often both the tone-bearing unit and the meaning-bearing unit) and of Cherokee, an American Indian language.

Voltaic (Gur) subgroup. The Voltaic (Gur) subgroup includes 79 languages spoken in Upper Volta and the upper halves of Ghana and Ivory Coast. Mossi is the most prominent of them. The term Gur was used as early as 1886 to designate the languages centred around Ouagadougou in what is now Upper Volta and northern Ghana. Languages with such names as Gurma or Grunshi (Gurunsi) were later related, and the first syllable of the names was taken as a name for the group. In 1911 Delafosse used the designation "Voltaic" for the same group, probably because many of the Gur languages are spoken in the Volta River Basin.

Typological significance of the Voltaic languages

The study of the Voltaic languages as a group is important because they do not partake strongly of any entrenched typological or grammatical "manifestations" (such as the "manifestations" of sex gender distinctions in Indo-European, triconsonantal roots in Semitic, and so on). A wide range of possibilities is present, and, if any "manifestation" is at all apparent, it is a tendency toward simplification and serial constructions. This means that there are a very small number of closed grammatical classes and a greater reliance on constructional markers. Nevertheless, it is important, when considering any of the 1,000 or more languages surveyed in this article, to emphasize that each language within the subgroup is also important for its own sake. Each language reveals a particular way of communicating about the world, a way that is fully developed and elaborated in its own terms. These spoken languages are certainly not "underdeveloped" or primitive. In fact, the very complexity of analyzing the Voltaic languages (or any other group of natural languages) in terms and categories sufficiently "universal" to be operational both for those who already speak the languages and those who want to learn something about them reflects their developed nature.

The most widely known characteristic of the Voltaic (Gur) languages is the realization of the Niger-Congo noun classes in the shape of parallel prefixes and suffixes, in which the suffixes appear to be dominant. In languages like Mõõre and Dagbane, for instance, nouns for a "person" end with the suffix *-a* (*-u* in Kasem), which is replaced by *-ba* in the plural. This feature is somewhat reminiscent of Fulani of the West Atlantic subgroup but differs in that it typically appears in the form of separable syllables in the Voltaic languages, rather than as an alternation or mutation of the beginning and ending consonants of the root itself, as in Fulani. The partial classification of human beings, animals, and liquids into separate grammatical classes provides evidence for the original nature of such classes in the ancestral Niger-Congo language. So does the fact that this classification governs the participation of variant forms of the detachable suffixes (or suffix–prefix frames) in sentence constructions calling for pronoun-like behaviour—referential, possessive, demonstrative, relative, and so on. Thus, Bargu has the suffix *-a* for the "person" class, a suffix that always "governs" the occurrence of *yé* as independent pronoun, *yù* as object marker, and *yè* as one demonstrative.

Ga, Yoruba, Igbo: prominent Kwa tongues

Kwa subgroup. The Kwa subgroup includes 73 languages, many of which are prominent (*e.g.*, Ga, Yoruba, Igbo). This grouping includes some of the most well known of the Niger-Congo languages and some with the greatest number of speakers. The Kwa languages are spoken in the belt of tropical forest through the southern section of the West African bulge. They are the languages of the former great kingdoms of tropical West Africa, such as Ashanti (Akan or Twi language), Dahomey (Ewe), Oyo (Yoruba), and Benin (Edo, or Bini language). In addition, the large decentralized group of Ibo (Igbo) also speak a Kwa language. Each of these languages has millions of speakers and a growing written literature, and one of them, Yoruba, boasts the earliest published dictionary and grammar written by a native speaker. It is symptomatic of the position of these languages in modern life, however, that these Yoruba works were written in English by an African Anglican churchman, Bishop Samuel Crowther. English or French remains the national language of the countries in which Kwa languages are spoken, and interest in employing the local languages is largely restricted to religious or academic institutions based on European models.

Like the Voltaic subgroup, the Kwa languages are diverse in their structure and characteristic features, and examples of this diversity are found in the roles of tone and vowel harmony. The continuum from clear tones in isolated dictionary citations to modification in sentence context referred to in connection with the Mande languages is particularly striking in the Kwa languages. Thus, some investigators heard consistent pitch levels that could be noted on a musical staff for any one speaker (in at least one Kwa language, Grebo of Liberia, there are four such levels). Others, such as those investigating Akan, heard just as clearly that there is a complicated system of pitches that are varied according to transitions between syllables or interrelations among syllables. It is as yet unclear just which, if either, of these systems is historically earlier: whether the clear tone levels are a crystallization of the distinctions imposed by contextual reactions with other tones or whether the languages that have almost lost all independent lexical tone are a newer reinterpretation of a previous system of fixed pitches.

Vowel harmony in the Kwa tongues

Vowel harmony involves a system of mutual selection or restriction among the vowels that can go into the makeup of a word. The vowels in a word are said to be in "harmony," with high vowels usually occurring with high vowels, front vowels with other front vowels, and so forth, in such a way that a constant feature seems to be realized throughout the word. Vowel harmony of a type peculiar to sub-Saharan languages is called "cross-height" vowel harmony by the British scholar John M. Stewart. In this situation, the higher of two varieties of a vowel selects the matching higher variety of another vowel and vice versa. Thus, *u* (as in "rule") would occur with *o* (as in "bone"), and the lower sound ʊ (as in "pull") with the correspondingly lower ɔ (as in "law"). Stewart also summarized evidence for contrast in the formation of the sounds in terms of the advancement of the root of the tongue. Under conditions of change of a vowel system, this leads to reinterpretations that cloud the nature of earlier contrasts and add to the variety found in the Kwa languages.

It is very difficult to provide a generalized "profile" of the Kwa languages that would give an idea of what is typical of a language of this subgroup. A stereotype would be that the words are mostly short and the nouns begin with vowels. Proper names often appear to be quite long, however, because they are compounds or phrases made up of several words. Many words of Kwa origin are found in the Americas, as is the pantheon of Yoruba gods in South America and the Caribbean.

Adamawa-Eastern subgroup. The Adamawa-Eastern subgroup comprises 112 languages spoken in the Central African Republic and northern parts of Cameroon and Zaire. These geographically remote languages are the least known of the Niger-Congo family. The subgroup's most prominent members are Sango and Baya (Gbaya). The name Adamawa-Eastern implies a combination of two recognized sub-subgroups: Adamawa and Eastern. None of the language names of this subgroup is well-known to outsiders, with the possible exceptions of the ethnographically noted Azande (Zande) and the trade language Sango.

Sango is clearly simplified from a dialect of one of the Adamawa-Eastern languages. It shows the result of contact with French as well as the usage of many river tribes in trade along the Ubangi River. It is the lingua franca of the Central African Republic and is extremely widespread as a second language in Central Africa. Many other pidginized trade languages based on an African vernacular are also spoken in this central area of the great rivers; *e.g.*, a pidgin dialect of Swahili, Kituba (a pidgin form of Kongo [Kikongo]), and Lingala.

Benue-Congo subgroup. The Benue-Congo subgroup comprises 557 languages, spoken from Nigeria to South Africa. Its most prominent members include Swahili, Kongo (Kikongo), and Zulu. In terms of numbers of speak-

Bantu, largest Benue-Congo subgroup

ers and geographical extension, Benue-Congo and Bantu are practically the same entity. That is, compared with the many millions of speakers of Bantu, the speakers of non-Bantu languages are only a tiny minority of Benue-Congo speakers. Furthermore, the Bantu speakers are spread out over most of middle and southern Africa, whereas the remainder of Benue-Congo speakers occupy only a corner of the northwest extension, mostly in Nigeria. The only two major non-Bantu Benue-Congo languages are Efik and Tiv, found in eastern Nigeria.

The following description sacrifices a consideration of the numerous but less important non-Bantu languages of the subgroup for a still superficial view of the Bantu languages. The most notable feature of the Bantu languages is their system of noun classes (concord). This has also been called "alliterative" concord because it often entails a repetition of the same syllable as prefixes at the head of successive words in the sentence. An example, drawn from Swahili, provides an illustration: *wa-tu wa-le wa-mefika* (noun-demonstrative-verb) "those people have arrived." Such prefixes as *wa* represent an intersection of two systems: one syntactic (function in the sentence) and the other paradigmatic (mutually exclusive replacements in the same position in the sentence). Syntactically, the shape of the prefix can vary according to the part of speech to which it is attached—a prefixed syllable may be different in the noun, adjective, possessive, locative, or verb. That there is agreement (*i.e.,* a concord element is present) with virtually all parts of speech creates an aesthetically pleasing system by means of an exhaustive display of marked relationships in the sentence.

Noun concord in Swahili

Paradigmatically, the system of concord is most simply viewed as being based upon the governing noun in its singular or plural form. That is, the noun belongs to one of a number of noun classes that determine the proper set of concording elements to be used throughout the part of the sentence dominated by that noun. These sets (which can be thought of as replacement sets) come in pairs according to whether the governing noun is singular or plural; *e.g.,* in Swahili, the singular form *m-tu yu-le a-mefika* "that person has arrived" contrasts throughout with the plural form *wa-tu wa-le wa-mefika* "those people have arrived." This alternation of singularity and plurality is, of course, more closely related to the semantics of the noun than to its purely grammatical function. The paradigmatic system goes even further in this direction by making a rather loose use of other possible semantic contrasts (especially: person–nonperson and countable–noncountable) in assigning nouns to noun classes. In no language, however, is there a strict relation between meaning and noun class. Nevertheless, the tendency in this direction again stimulates a sense of completeness of aesthetic form and has added to the fame of Bantu languages as abstract figures.

Correlates of these noun classes (in both form and function) are found in all branches of Niger-Congo except Mande and only slightly in Kwa. This led some early researchers to combine all of them as one subgroup and later led to a typological distinction between noun-class and non-noun-class languages. This distinction is not very useful, however, because there are a number of languages with rudimentary or partial noun-class systems, and it even seems possible for a language of entirely different ancestors to adopt noun prefix alternation.

Area of tonal reversal

Swahili remains the best known Bantu language. It is the national language of Tanzania and is used as a lingua franca throughout East Africa. The use of Swahili for this purpose is facilitated because it no longer makes use of tone—it is an isolated non-tone language in the Bantu area. There is another phenomenon related to tone in the Bantu languages besides its disappearance in Swahili that should be mentioned: there exists an area of "tonal reversal," a geographic area centred in the western Congo in which the tonal realizations in words and sentences are the mirror images of what they are elsewhere in Bantu languages; *i.e.,* high tones become low tones, and vice versa.

Another distributional characteristic is the presence of a prefix form made up of two syllables as opposed to the form with only one. Following out the implications of the geographic distribution as well as the function of this feature is of help in mapping the prehistory of the Bantu languages. As in all other areas of the world, however, it is the distribution of certain items of vocabulary, or of vocabulary replacements, combined with the application of laws of sound change, that produce the most valuable historical conjectures. An impressive amount of work has been done on the reconstruction of ancestral Bantu word roots and the sound system used to realize them. As early as the 1890s, Carl Meinhof postulated an original sound system and started working out the shifts that it had been subjected to in order to produce the various Bantu languages as they are spoken today. Meinhof called his systematic reconstruction Ur-Bantu, and most subsequent historical research (and much actual description of contemporary languages) has been based on his surprisingly trustworthy descriptions. *Comparative Bantu* (4 vol., 1967–70), by Malcolm Guthrie of the University of London, is a much fuller and more painstakingly detailed compendium of the entire geographical distribution of the Bantu languages.

Although textbooks and lessons have been published for many Niger-Congo languages, those of the Bantu area are by far the most numerous. In part, this availability reflects the influence of the South African linguist Clement Doke and the department of Bantu studies at the University of the Witwatersrand, which was in direct touch with a huge area of Bantu-speaking peoples.

ADOPTION OF FOREIGN WORDS AND SOUNDS

Borrowing of click sounds

All the subgroups of Niger-Congo contain languages that show evidence of intensive contact with each other or with languages outside the group to such an extent that special mechanisms are in operation for the inclusion of borrowed words (such as the assignment of nouns to a particular noun class). This does not make them especially unusual from the standpoint of the world's languages, but the borrowing of the very difficult click consonants from the neighbouring Bushman-Hottentot languages into the Bantu languages of South Africa is definitely remarkable. Clicks are especially widespread in Zulu and Xhosa—the first consonant of even the name Xhosa itself stands for a click. This dramatic exception to the theory of language change that states that languages alter in order to simplify or economize the effort of sound production seems to require a special explanation. Possibly in-marrying Bushman or Hottentot wives used their native click sounds as disguised forms of words that would be taboo in Zulu or Xhosa. (D.W.C.)

Chari-Nile and Nilo-Saharan languages

The term Chari-Nile refers to a group of languages presumed to be genetically interrelated, or descended from a common ancestral language. Joseph Greenberg proposed the Chari-Nile family as part of his overall classification of African languages. At first, he designated the group by the name Macro-Sudanic (1955), but in 1963 adopted the term Chari-Nile. The name Chari-Nile is geographically descriptive in that virtually all of the languages in the family are located in the watersheds of the Nile and the Chari rivers or in the areas in between them.

CLASSIFICATION

The larger Nilo-Saharan group. Greenberg, as of 1963, regarded Chari-Nile as only a branch of a larger family, Nilo-Saharan, of which it is by far the most heterogeneous. The other branches of Greenberg's Nilo-Saharan family are: (1) Songhai, an important language with no close relatives, which consists of a number of dialectal varieties (*e.g.,* Zerma) spoken along the Niger River in Mali and Niger; (2) Saharan, a language group including Kanuri, the major language of northeastern Nigeria, as well as its relative Teda and its more distant relative Zaghawa, both of which are spoken to the east of Kanuri, principally in Chad, but also in The Sudan; (3) Maba, a group of interrelated languages of Chad; (4) Koma (also written Coman), a group of interrelated languages of the Ethiopia–Sudan border area; and (5) Fur, a language of Northern and Southern Darfur provinces, The Sudan.

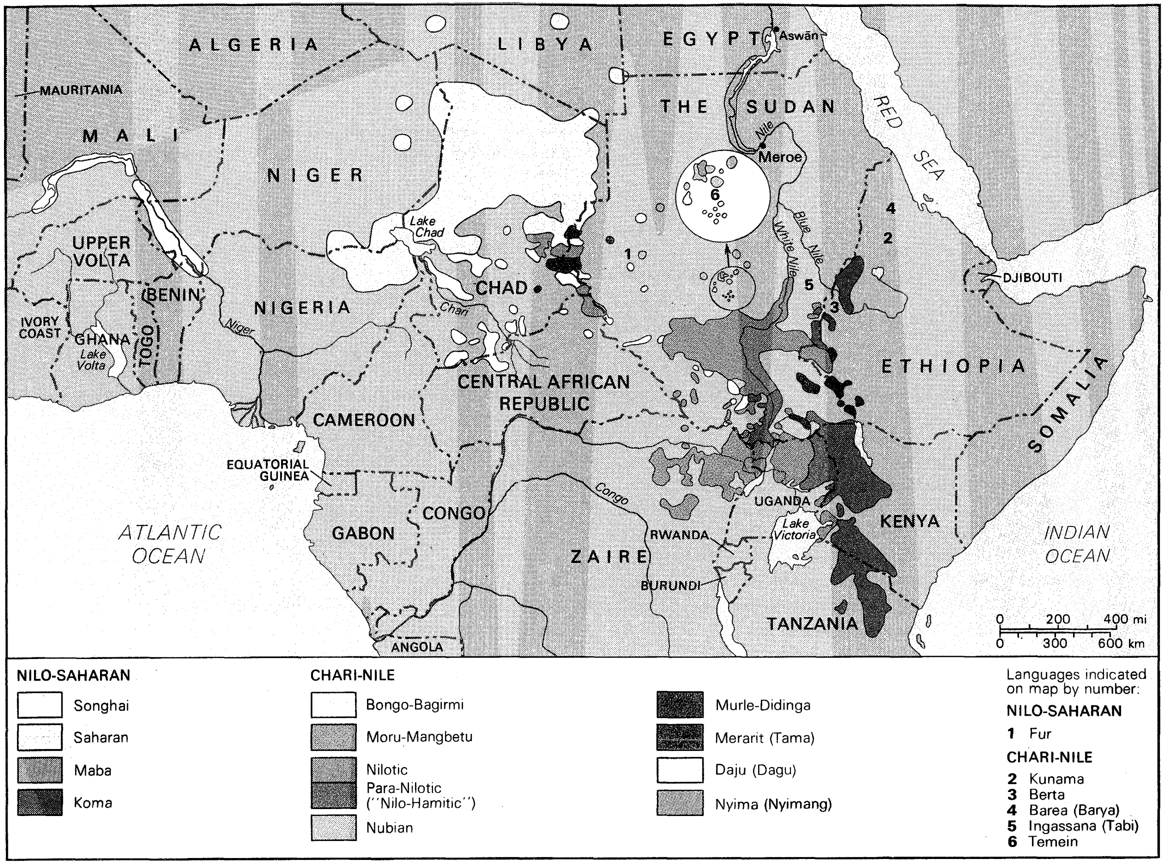


Figure 32: Distribution of the Nilo-Saharan and Chari-Nile languages.

Adapted from A.N. Tucker and M.A. Bryan, *Linguistic Analyses: The Non-Bantu Languages of North-Eastern Africa* (1966); Oxford University Press (for International African Institute), London

Languages of the Chari-Nile group. Greenberg considered Chari-Nile to be composed of four branches. The first, Kunama, is a language of northern Ethiopia (Eritrea), contiguous to Barea (Barya), an Eastern Sudanic language (see below); Berta, the second branch, is spoken in the Ethiopia–Sudan border area. Central Sudanic, a large and heterogeneous group of languages, extends over northwestern Uganda, the southern Sudan, northeastern Zaire, Chad, and the Central African Republic, although it is interspersed with languages of other groups. The British Africanists Archibald N. Tucker and Margaret A. Bryan (1956) regarded Central Sudanic as two distinct unities, which they hesitated to group together; one group is Bongo-Bagirmi, of which Bagirmi, Sara, and a number of other languages are found in Chad and the Central African Republic, while Bongo and its closest relatives are concentrated in the extreme south of The Sudan near the Zaire and Uganda boundaries. The other division is Moru-Mangbetu (like Bongo and Bagirmi, these are the names of individual languages), which is concentrated in northeastern Zaire but overlaps into Uganda. In addition, it includes Efe, Lendu, and Madi, which is linked to Moru. The fourth and most heterogeneous branch of Chari-Nile is Eastern Sudanic.

The 10 subgroups of Eastern Sudanic

Eastern Sudanic consists of 10 subgroups: (1) Nilotic extends from southern Sudan through Uganda, Kenya, and northern Tanzania, although it is interspersed with languages of other groups, and includes the so-called "Nilo-Hamitic" languages. (This name is controversial; see below *Nilotic and "Nilo-Hamitic."*) The Nilotic subgroup includes Shilluk, Dinka, and Nuer in the southern Sudan, Acholi in northern Uganda, and Luo, in western Kenya along Lake Victoria but overlapping into Tanzania. The Nilo-Hamitic languages include Bari and Lotuko (Lotuho) in the extreme south of The Sudan; Karamojong in northern Uganda (just to the west of Acholi) and its neighbour and close relative Turkana across the border in Kenya; Nandi and Suk, just to the south of Karamojong; and Masai, extending from southern Kenya into northern Tanzania.

(2) Nubian is spoken along the Nile from Aswān, Egypt to Meroe, The Sudan (*i.e.,* Nile Nubian), and a number of dialect enclaves in the Nuba Hills of Southern Kurdufān province, The Sudan (*i.e.,* Hill Nubian), as well as in two separate enclaves, Midobi and Birked, in Northern and Southern Darfur provinces, The Sudan. The relationship of Midobi and Birked to one another and to the other varieties is not close. Currently, Birked is giving way to Arabic. Nile Nubian consists of three languages, of which the northernmost, Kenuzi (Kenzi), and the southernmost, Dongolese, are closely related, almost to the point of mutual intelligibility; Mahas, located between the two, is a more distant relative and includes a subdialect, Fadidja, spoken to the north.

(3) Murle-Didinga designates a group of closely interrelated languages spoken on both sides of the Sudan–Ethiopia boundary; Murle and Didinga are individual languages within the group. (4) Barea (or Barya), a language of northern Ethiopia, is spoken just to the north of Kunama. (5) Merarit (also written Mararit), spoken on both sides of the Chad–Sudan border, comprises a group of closely interrelated languages; actually, Merarit is only one language of the division, which also includes Tama, for which the group is sometimes named. (6) Ingassana, also known as Tabi, is found in the eastern Sudan just north of Berta. (7) Daju, or Dagu, is the name of a group of closely related languages spoken in widely separated enclaves in Chad and The Sudan; most of them are called Daju or some variant of the name.

Recent additions to the Eastern Sudanic branch

Three subgroups of Eastern Sudanic were added by Greenberg in 1963; the first two had previously (1955) been treated as entirely independent language families, while the third had not been mentioned. These are: (8) Nyangiya, also known as Teuso, a group of dialects or closely related languages of northern Uganda, whose inclusion in Eastern Sudanic Greenberg regards as only tentative (this is omitted on the map); (9) Temein, a language of Northern and Southern Kurdufān provinces, The Sudan; and (10) Nyima, or Nyimang, also found in the

Kurdufān region. Another language, Afitti, spoken about 100 miles (160 kilometres) to the north, evidently should be grouped with Nyima.

Controversies concerning classification. There has been much disagreement about the classification of various groups of African languages and considerable confusion in terminology.

Objections to grouping Nilotic with "Nilo-Hamitic"

Nilotic and "Nilo-Hamitic." Some scholars have objected to grouping the so-called "Nilo-Hamitic" languages together with Nilotic. This question has been at issue since the earliest attempts at African linguistic classification. One of the first proposed systematic classifications (1877–78) established the very heterogeneous Nuba-Fula group, named for two of its members, Nubian (an Eastern Sudanic language, according to Greenberg) and Fula, or Fulani (a West African language). The group also included Masai, subsequently classified as "Nilo-Hamitic," but not the other Nilotic or "Nilo-Hamitic" languages. This classification, made in the 19th century by Friedrich Müller, was based mainly on a simplistic correlation with racial categories determined largely by hair type. Müller recognized a close affiliation between Bari (subsequently classed as "Nilo-Hamitic") and Dinka, a Nilotic language. Dinka, Shilluk, Nuer, and Bari were grouped under the heading Nile languages (*Nil-Sprachen*) within the Negro group, as distinct from the Nuba-Fula group. (Later Müller admitted the close affinities between Masai and Bari, previously pointed out by Karl Richard Lepsius, but attributed them to a mixture of two distinct elements in Bari.)

In 1880, Lepsius proposed another of the early classifications of African tongues. As a result of his views that Masai and Bari were related, the genetic affiliation of Masai, Bari, Nuer, Dinka, and Shilluk was generally accepted, and languages showing clear affinities to them were included as they became known (*e.g.,* Lotuko, Nandi-Suk, Luo).

In 1902 Sir Harry Johnston divided these languages into two groups, one of which he called Nilotic, including Dinka, Shilluk, Acholi, Alur, Lango, and Luo. He proposed no specific name for the second group, which later came to be known as "Nilo-Hamitic," but included in it Masai, Turkana, Lataka (*i.e.,* Lotuko), Nandi, Suk, and Karamojong. He posited a distant relationship between the two groups but also noted some striking lexical similarities between Somali (a Hamitic language) and the second, or "Nilo-Hamitic," group that led him to conclude that the latter resulted from a mixture of one group of people related racially and linguistically to the Somali and another related to the Nilotes.

Correspondence of Sir Charles Eliot's classification with Nilotic and "Nilo-Hamitic"

Most investigators of this period grouped these languages in much the same way as Johnston; for example, the English diplomat and scholar Sir Charles Eliot stated:

> East Equatorial Africa is the home of a group of tribes which cannot be called Bantu or Hamitic, and have received no satisfactory designation. The group contains at least two subgroups. One of these, including the Shilluk, Dinka, Bari, Acholi, and Jaluo [*i.e.,* Luo], inhabits the banks of the Nile and the shores of Lake Victoria; the other has its headquarters in the highlands of . . . East Africa. . . . Its chief members are the Masai, Nandi, Turkana, and Suk. . . . To these should probably be added Latuka [*i.e.,* Lotoho] but our information about it is slight. . . . Karamojo is not well known but is apparently closely allied to Turkana.

Except for Bari, which was put in the first group for reasons more geographic than linguistic, but which is actually closely related to Masai, Eliot's two divisions correspond exactly to Johnson's—namely, Nilotic and what later became known as "Nilo-Hamitic."

A sharply different view was introduced in 1912 by Carl Meinhof, who denied the linguistic unity of the two groups suggested by Johnston and Eliot. Meinhof regarded the term Nile (or Nilotic) merely as a geographic concept. He clearly considered Bari and Masai to be Hamitic languages (*i.e.,* related to Somali and Galla, among others) and also, presumably, Lotuko, Turkana, Nandi, and Suk. (He was not, however, explicit about these.) Dinka and its close relatives—Shilluk, Nuer, Luo, and others—were classified as Sudanic; the resemblances between the two groups (in particular those between Dinka and Bari, which had been pointed out by Müller) he attributed to borrowing. Meinhof's most important reason for allying Masai and its close relatives to the Hamitic languages was that in both groups nouns have grammatical gender and other inflectional complexities, which are absent in Nilotic languages such as Shilluk, Luo, and their close relatives.

Much earlier, Lepsius also had noted the presence of grammatical gender in Masai, a trait to which he attributed great importance for genetic classification, but he specifically ruled out any Hamitic affiliation because gender is expressed very differently in the Hamitic languages. Because of Meinhof's great prestige and preeminence among Africanist linguists, his views were widely accepted. The scholar Dietrich Westermann (in 1912), however, regarded all of the above as Nilotic languages, dividing them into the Niloto-Hamitic (Nilo-Hamitic) and Niloto-Sudanic subdivisions, which correspond to the groups delineated by Johnston. On the other hand, contrary to general practice, he also included among the Niloto-Sudanic languages another group, Moru-Madi, of western Uganda and eastern Congo. Greenberg subsequently included Moru-Madi within the Central Sudanic family, which he considered a branch of Chari-Nile and, thus, ultimately but distantly related to Nilotic.

Doubtful affiliation between "Nilo-Hamitic" and Hamitic

In 1935, Westermann claimed that the relation between Niloto-Sudanic and the Sudanic languages of West Africa was not as close as the relation of the latter to Bantu, whose interrelationship he had sought to demonstrate in 1927. Greenberg was to designate this family, Bantu included, as Niger-Congo. Westermann likewise expressed doubt as to the affiliation of "Nilo-Hamitic" to the Hamitic languages and reaffirmed the unity of Nilotic and "Nilo-Hamitic." Greenberg also adopted this position and argued persuasively against the relationship of "Nilo-Hamitic" to the Hamitic languages and, consequently, against the misleading nature of its name, which he changed to Great Lakes and which, together with Nilotic, formed the southern branch of his Eastern Sudanic family.

The German Africanist Oswin Köhler proposed, in 1955, the name Nilotic for the entire southern branch and divided it into three subdivisions: West Nilotic (including Nuer, Dinka, Shilluk, Luo), East Nilotic (including Bari, Masai, Lotuko, Karamojong), and South Nilotic (including Nandi, Suk, Tatoga). (The last two had usually been grouped together as Nilo-Hamitic, while the first had simply been designated Nilotic.) In 1963 Greenberg subsequently accepted Köhler's grouping and nomenclature. Tucker and Bryan, while still maintaining the unity of "Nilo-Hamitic" and hesitating to group it unequivocally with Nilotic (in the older restricted sense, equivalent to Köhler's West Nilotic), proposed a new name for the group, Para-Nilotic, thus avoiding any implication of Hamitic affiliation.

The wider relationships of the Nilotic family. The question of the relationship of Nilotic (in Köhler's extended use of the term) to other African languages and language families has been raised at various times. In 1920 the English anthropologist G.W. Murray noted similarities between Bari and Nubian and, in fact, between Nubian and the Nilotic and "Nilo-Hamitic" languages as a whole. He also linked these languages to Kunama, Barea, and the recently discovered Tabi (also known as Ingassana). Westermann suggested in 1935 that Didinga might be a distant branch of Nilotic and had earlier (1912) made a similar claim for the Moru-Madi group. All these languages were subsequently classed by Greenberg either as Eastern Sudanic or within the wider grouping of Chari-Nile.

Apart from Eastern Sudanic, the most heterogeneous subgroup of Chari-Nile is Central Sudanic. Its unity was first suggested in 1940 by Tucker, who included it in a group of languages that he called Eastern Sudanic, a term that he admitted was essentially geographic because it included another group of different and distinct languages, among them Ndogo-Sere and Azande. (Greenberg subsequently incorporated these into Niger-Congo.) Greenberg introduced a terminological confusion here by calling part of Tucker's Eastern Sudanic group Central Sudanic and using the name Eastern Sudanic for an entirely different group of languages. Although Tucker hesitated to unite the Moru-Madi and Bongo-Bagirmi languages—which together comprise Greenberg's Central Sudanic—into a single family, the evidence that he himself presented appeared to support such a grouping.

Opposition to Greenberg's groupings

Some scholars doubt that adequate evidence has been presented for regarding the whole of Eastern Sudanic as a valid subdivision of Chari-Nile or even for regarding the whole of Chari-Nile as a valid subdivision of Nilo-Saharan. Although Nilo-Saharan and the more heterogeneous subdivisions within it are far from conclusively established, no convincing alternatives have been suggested.

LINGUISTIC CHARACTERISTICS

The Chari-Nile languages are a very heterogeneous group whose structural features, both phonological and grammatical, are in no way uniform or easily characterized. Furthermore, the great majority of the languages are not adequately described, and their comparative study is still in its infancy.

Phonology. Nearly all Chari-Nile languages seem to be tonal (*i.e.,* words and grammatical forms are differentiated by pitch), as are the overwhelming majority of sub-Saharan languages. Some varieties of Nubian may be exceptions, but the question has yet to be fully investigated. Central Sudanic languages generally have only syllables that are open (*i.e.,* end in vowels); some final vowels in the Bongo-Bagirmi languages of this division, however, are semi-mute or whispered. Other Chari-Nile languages have syllables that are open as well as closed (*i.e.,* end in consonants). Central Sudanic languages likewise differ from other Chari-Nile languages in that they are the only ones with the labiovelar consonants *kp* and *gb* (articulated with both the lips and the soft palate), which are common in much of West and Central, but not East, Africa. As in other African language families, a large number of Chari-Nile languages distinguish dental from alveolar or retroflex stops, among them most Central Sudanic languages, Nyima, Temein, Murle-Didinga, Hill Nubian, and West Nilotic. (The dental stops, *t* and *d,* are pronounced with the tongue tip against the back of the upper teeth; the alveolar and retroflex stops are pronounced with the tongue tip farther back along the roof of the mouth.) Some Chari-Nile languages—most Central Sudanic languages, Daju, the Murle-Didinga group, and Bari (East Nilotic)—distinguish implosive and explosive voiced stops; the former are pronounced by drawing the air into the mouth, the latter by expelling it. The vowel systems of most Chari-Nile languages are fairly rich, the great majority seeming to distinguish seven or more vowel qualities. Apparent exceptions are Daju, Nile Nubian, and possibly Kunama, which have only five vowels. Many languages make distinctions in the length of vowels (*e.g.,* Nile Nubian, Nilotic, Daju, Kunama), but not all (*e.g.,* Central Sudanic, Nyima). In Nilotic languages a contrast exists between so-called breathy versus hard vowels, sometimes analyzed as close versus open. Nasal vowels occur in Sara-Mbai of the Bongo-Bagirmi group within Central Sudanic but in no other Chari-Nile language.

Vowel systems of the Chari-Nile languages

Grammar. No Chari-Nile language has concordial noun classes like those of Bantu (see above *Benue-Congo subgroup*) and many of its more distant relatives in West Africa, but grammatical gender is distinguished in three very different Chari-Nile groups. It is prominent in East Nilotic (*i.e.,* "Nilo-Hamitic" minus the Nandi-Suk-Tatoga group) but absent in the other Nilotic languages. There are two genders of nouns, normally indicated by a determiner (in Masai, these are *l* for masculine, *n* for feminine) with various grammatical functions. On the other hand, gender is not distinguished in the personal pronouns, but in Daju (grouped by Greenberg in Eastern Sudanic) there are three genders (masculine, feminine, neuter) distinguished in the 3rd person singular pronoun ("he," "she," "it") but not in the noun. The same usage occurs in Bongo but in no other Central Sudanic language.

A few Chari-Nile languages distinguish exclusive forms from inclusive forms of the 1st person plural pronoun. The exclusive pronoun excludes the hearer (*e.g.,* "he and I" = "we"); the inclusive includes him (*e.g.,* "You and I" = "we"). Among such languages are Lendu (alone among Central Sudanic languages), Daju, most West Nilotic languages, and Teso (alone among East Nilotic languages). Virtually all Chari-Nile languages except Berta and Nyima formally distinguish plural from singular nouns, some by means of prefixes (*e.g.,* the Moru-Mangbetu languages within Central Sudanic, and Temein) but most with suffixes (*e.g.,* the Bongo-Bagirmi languages within Central Sudanic, Kunama, Barya, Daju, Ingassana, Murle-Didinga, Nyangiya, and East and South Nilotic, which together comprise "Nilo-Hamitic"). West Nilotic utilizes suffixes sometimes, as well as change of tone, vowel length, or vowel quality. These processes can also be found to some extent in other branches of Nilotic. Affixes are sounds or groups of sounds that are added to the beginning (prefixes) or end (suffixes) of a word or inserted in the middle (infixes). In Chari-Nile, it has been noted that two pairs of singular-plural affixes are widespread, though not universal; these are also used to some extent outside this family. The affixes are *t* and *k,* confined principally to nouns, and *n* and *k,* principally used with pronominal elements. The *t* and *k* forms occur in fewer languages than the *n* and *k* ones, and only in languages in which the *n* and *k* are also found. The significance of these grammatical features for the historical and comparative study of the Chari-Nile languages remains to be investigated.

Widespread use of specific affixes in Chari-Nile

Writing. Only one Chari-Nile language, Nubian, has a written tradition of any antiquity. Dating from the 8th to the 11th centuries, the script was employed by Christianized Nubians in much the same area as that occupied by modern Nile Nubian speakers. Generally called Old Nubian, this written language most closely resembles Mahas among the modern dialects. It was abandoned after the Nubians adopted Islām and survives only in ancient manuscripts and inscriptions that were not deciphered until 1906. (Since the adoption of Islām, Nubian has occasionally been written in Arabic script.) Old Nubian utilized a script derived from that of Coptic, which in turn was adapted from the Greek alphabet. Coptic added seven non-Greek letters of Egyptian demotic origin to represent non-Greek sounds. Nubian retained three of these letters, in addition to adding three new ones from cursive forms of Meroitic letters.

Meroitic, the extinct language of an even more ancient Sudanese civilization (*c.* 300 BC to AD 100), may be related to the Chari-Nile languages, but too little is known to establish any conclusive tie. It is not closely allied to any existing language. Surviving only in inscriptions, Meroitic was written in a consonantal alphabet derived from Egyptian hieroglyphic writing and in a cursive form partly of Egyptian demotic and partly of indigenous origin. In the first decade of the 20th century the script was deciphered by the English Egyptologist Francis Griffith.

Languages using Latin-based alphabets

Since the colonial period a number of Chari-Nile languages, particularly in Kenya, Uganda, and The Sudan, have been written in derivatives of the Latin alphabet and are used for religious and educational purposes. Nearly all of the languages are Nilotic. Of these, Luo, a West Nilotic language and one of the major indigenous languages of Kenya, has the largest vernacular literature. It also has the largest number of speakers of any Nilotic tongue (and, in fact, of any Chari-Nile language). Other West Nilotic languages, however, such as Dinka and Nuer of The Sudan and Acholi of Uganda, have some vernacular literature, as does Kalenjin (Nandi-Kipsigi), a South Nilotic language of Kenya. (M.F.Go.)

Khoisan languages

The Khoisan languages are click languages spoken in southern Africa. The term Khoisan was created to refer to the related peoples known as Bushmen and Hottentots (*i.e.,* the Khoisanid peoples) under a common name and has become increasingly accepted since its creation in 1928. The word is derived from Khoikhoi and San, the names of the peoples called, respectively and pejoratively, Hottentots and Bushmen.

The languages of the group are now usually divided into three groups—North, Central, and South Khoisan. They are spoken by remnants of a pre-Bantu population, the so-called Bushmen, who are hunters and collectors liv-

Table 59: Notation of Clicks

	characters used in text	International Phonetic Alphabet	Zulu conversion
Dental	/	ʇ	c
Alveolar	≠	o	
Alveopalatal (or retroflex)	!	ʗ	q
Lateral	⫽	ʖ	x
Bilabial	⊙		

ing in and around the Kalahari and who number about 78,000. A Khoisan language is also used by the so-called Hottentots, pastoralists surviving with about 50,000 Nama speakers in South West Africa/Namibia, where the Nama language has also been adopted by the Bergdama, a non-Khoisanid people (about 77,000), and by the Hai-⫽'om Bushmen (more than 3,000 speakers). The Khoisan family comprises about a dozen languages and dialect clusters such as Nama, !Kora, Naro, /Kham, !Khung, Kxoe, and others. (The unusual symbols in the names of the languages stand for clicks; they are listed in Table 59 and are explained below in the section on phonology.)

Problems of Khoisan genetic relationship

The assumption of a genetic relationship between all the languages of the Khoisanids and the application of the term Khoisan to this language "family" are problematic. Criteria brought forth against the point of view of a genetic relationship are the cleavage in morphology and vocabulary between sex-gender languages (Central Khoisan) and nongender languages (North Khoisan and South Khoisan), on the one hand, and a similar cleavage in morphology and vocabulary between North Khoisan and South Khoisan, on the other. These arguments are countered by the fact that there are a number of common words and a few particles, traceable by comparison of sound and meaning, in two or more of the three groups of Khoisan. These permit the establishment of a strong hypothesis for a genetic, though remote, relationship. Regular sound correspondence, however, has not been found in all common items; this is the most difficult part of the Khoisan problem.

The Khoisan hypothesis is built on: (1) common special features of the phonological system—*e.g.,* clicks are regarded as inherent to the languages; (2) widespread and common patterns of root formation, combined with special patterns of consonant distribution; (3) the occurrence of some probably related particles in more than one group; and (4) the occurrence of related words (sound-meaning units) in two of the groups, but more rarely in all three groups—*e.g.,* the terms for "chin," "lungs," "throat," and "wound."

In addition to Khoisan proper, there are two click languages, Sandawe and Hadza (Hatsa), spoken by peoples in Tanzania; these languages possess a few words, affixes, and particles that justify the assumption of a distant relationship with Khoisan. The apparent link between Central Khoisan and Sandawe is supported by the racial affinities of the Sandawe and Khoikhoi (Hottentot) peoples. The relations between Khoisan and Hadza are more remote. All studies on this subject are still in the initial stages. The U.S. linguist William E. Welmers proposed the tentative extension of the Khoisan family into Macro-Khoisan, including Sandawe and Hadza. In 1962 Joseph Greenberg subsumed Khoisan proper and Sandawe and Hadza under the title Khoisan.

Some Central Khoisan gender affixes were at one time considered as possibly related to the Hamitic languages. Today, the basis for the Hottentot-Hamitic hypothesis is

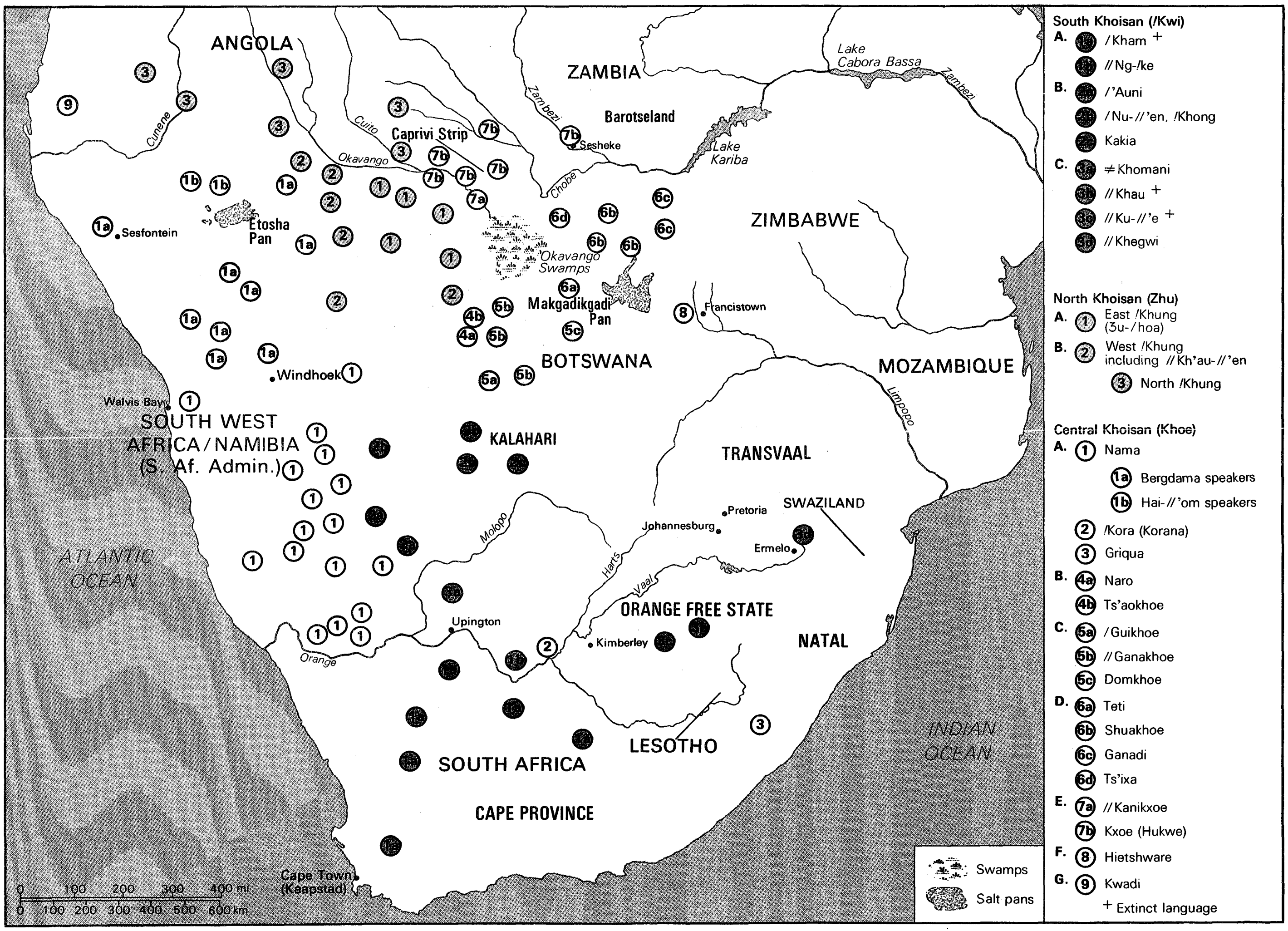


Figure 33: Tentative distribution of the Khoisan languages.

generally regarded as too narrow for conclusiveness and the hypothesis has been given up for lack of evidence.

CLASSIFICATION

All classifications of Khoisan are based on a general comparison of morphology and vocabulary. Lack of sufficient language material limits systematic comparative studies. While the division into groups is fairly consistent, the classification within the groups must be regarded as preliminary. The first classification of these languages, made by the German linguist Wilhelm Bleek in 1858, distinguished between the Hottentot and Bushman languages according to the gender-nongender dichotomy. This remained the traditional model until 1927, when Bushman was subdivided into Northern, Central, and Southern groups. These groups were maintained in D.F. Bleek's *A Bushman Dictionary* (1956), in which the Hadza language was included in the Central Group.

Modern theories of classification

Since 1950, several attempts at a reclassification of the languages of the Khoisanids have been proposed. The contribution of Greenberg to the Khoisan hypothesis was to include Hottentot in the Central group and, as mentioned above, to extend Khoisan to include the more closely related Sandawe and the more distantly related Hadza. Greenberg's main aim was to trace far-spread common elements in Khoisan as evidence for the genetic relationship of the groups. The method is based on sound-meaning comparison. The South African scholar E.O.J. Westphal separated the languages of the Bushmen from the gender languages that he interpreted as of Hottentot origin. Later, he divided the languages of the Khoisanids into five genetic groups, and in 1963 he further revised his classification of the non-Hottentot tongues by establishing four genetically unrelated groups, to which he gave the status of language families (1965); in 1971 he reduced these to three families. A new classification of Central Khoisan was put forth by Oswin Köhler in 1962, which showed the special position of the Nama and !Kora dialects of the Hottentots within the Central Khoisan group (see below).

There is no general agreement on the division of the South Khoisan languages. Some special phonological and morphological features characterize South Khoisan as a genetically related group of languages whose divergences must go back to a much older separation than that in Central and North Khoisan. South Khoisan speakers were probably the first Khoisanid migrants to reach South Africa.

The following is a classification, proposed by Köhler, of the languages subsumed in the three major Khoisan groups (a comma between languages indicates a close relationship; a semicolon designates a more distant relationship):

South Khoisan: /Kham, ⫽Ng-!ke; /'Auni, Xatia, /Nu-⫽'en, !Khong, /Namani, Kakia; ≠Khomani, /Nhuki, ⫽Khau, ⫽Ku-⫽'e, Seroa, ⫽Khegwi (Batwa), !Gãne.

North Khoisan (dialects): East !Khung (ʒu-/hoa), West !Khung (called ⫽Kh'au-⫽'en or Auen in its southern area), North !Khung (!'O-!Khung in Angola), closely related to West !Khung.

Central Khoisan: Nama, spoken by the Bergdama and by the Hai-⫽'om; !Kora (Korana); Griqua; Naro, Ts'aokhoe; /Guikhoe, ⫽Ganakhoe, ≠Heβa-khoe, Domkhoe; Teti, Danisa, Hura, Shuakhoe, Ts'ixa; ⫽Kanikxoe, Kxoe (Hukwe, Kwengo, Bumakxoe, including Bugakxoe); Hietshware, /Haitshuari, Mohisa; Kwadi (tentatively classified as Central Khoisan by Köhler in 1968 and Westphal in 1971).

Sandawe has similarities to Central Khoisan and may be regarded as distantly related. Hadza shows some relationships to Khoisan and relatively few connections to Sandawe. Its inclusion in Khoisan is problematic and assumes a very high, abstract level of classification, presupposing that the languages separated a long time ago and that a great number of cognates and common features that normally attest to a common origin were lost over an extended period. This high level of classification is also characteristic of all three Khoisan groups.

LINGUISTIC CHARACTERISTICS

Phonology. Most characteristic of the Khoisan languages are the click sounds, which form a subsystem of the consonants. Clicks are produced by simultaneous closure of the tongue with two areas in the mouth, one at the soft palate (the velum), and the other at such prevelar locations as the teeth, alveolar ridge, or palate. After the closures are made, the body of the tongue is moved down and back, so that the enclosed part of the mouth cavity becomes larger. This causes a lowering of the air pressure in the cavity. The click noise occurs when the tip of the tongue is lowered, breaking the partial vacuum and allowing air to rush into the cavity.

Click sounds as a subsystem of the consonants

There are two types of release of clicks. In one the bilabial, dental, and lateral closures are released by an imploding affricate (a sound beginning as a stop and ending with more or less distinctly audible friction); the alveolar and alveopalatal (or retroflex) closures are released by a hard implosive sound. Both releases are called ingressive, or influx, releases, meaning that the air rushes into the mouth. The second type is called egressive, or efflux, release; it results from the velar closure. In this case the air rushes out of the mouth. A click plus efflux are considered as a single sound rather than as a consonant cluster. In the South Khoisan languages there is also a bilabial click with labial closure; this is like the sound of a kiss. The Khoisan click mechanisms also combine with other phonetic features—*e.g.,* voicing (vibration of the vocal cords), nasalization—to produce additional sounds. The clicks occur only in initial position in word roots (also in compound roots), rarely in affixes and particles. Some of the Khoisan click sounds were adopted into Bantu languages, notably the Nguni group that includes Zulu and Xhosa (see Table 59 for click notations).

Unless followed by a vowel, the effluxive part of a click is represented by a separate symbol. In this text, the symbols are follows: /a = dental click plus vowel; /h = dental click plus aspiration; /x = dental click plus velarization; /x' = dental click plus velar ejected efflux; /' = dental click plus glottalization; and /n = dental click plus nasalization. According to this system, as well as in the general orthography of personal and place names, *k* or *g* is used to indicate the velar closure of the click to make the words readable for the general reader who omits the initial click sound. In this case *k* is used to indicate an unvoiced click and *g* a voiced click. The examples given above thus read as follows: /a = /ka or /ga; /h = /kh or /gh; /x = /kh or /gh; /x' = /kh' or /gh'; and /' = /k' or /g'. In the case of nasalization (/n), *k* and *g* are not usually written.

In the Khoisan languages, there are intricate systems of tones. The main vowels are *a, e, i, o,* and *u;* ɛ as in "bed" and ɔ as in "all" do occur, but not in all languages, and the *ə* sound, as the *a* in sofa, is relatively rare. Nasalized *ã, ĩ,* and *ũ* are most common. Pharyngealized (or pressed) vowels and vowels with intermittent glottalization occur in North Khoisan and, to a much lesser degree, in South Khoisan but seldom in Central Khoisan. (A pharyngealized vowel is one produced with constriction of the pharynx, the area at the back of the mouth before the esophagus.) A characteristic feature of the consonants is their use in clusters with final velarization (articulation of the back of the tongue with the velum, or soft palate), aspiration (an accompanying puff of breath), or glottalization (accompanying closure of the glottis). This occurs in all three Khoisan groups.

Grammar. Because the grammatical features of the various Khoisan languages are so divergent and lack overall uniformity, they will be discussed in three groups—South Khoisan, North Khoisan, and Central Khoisan.

South Khoisan. The languages of the South Khoisan group are characterized by a great number of different grammatical elements and processes. For example, the plurals are formed not only by the addition of suffixes but also by reduplication (*i.e.,* the repetition of an element to indicate plural, as "bird-bird" for "birds"), by partial change of the stems, and by completely different stems, called suppletive forms. In the formation of tenses and moods, a variety of particles occurs before the verb. Significantly, there is no class or gender distinction in the South Khoisan languages, except for a few traces in the /Kham language. The normal sentence order is subject-predicate-object.

Types of plural formation

North Khoisan. The North Khoisan dialects are characterized by an almost complete absence of affixes (endings, prefixes, and so on). The relations of the words to one another and their function in particular utterances are designated by particles (similar to English prepositions such as "to," "for," or "with") and by word order. The tense of roots functioning as verbs is shown by an optional preceding particle for present and future time, by nothing for the past tense, or by an expression describing time ("now," "tomorrow," etc.) if used in a sentence without context. As in the South Khoisan languages, there are suppletive (different) stems for some plurals; *e.g., !hũ* means "to kill one," but *!'oa* means "to kill many." The North Khoisan dialects, also in common with the South Khoisan tongues, lack a passive construction and have the basic sentence order of subject–predicate–object.

Central Khoisan. In the treatment of the Central Khoisan languages, the Kxoe language of the west Caprivi Strip (South West Africa/Namibia) will serve as representative. The Central languages have a highly developed morphology. Kxoe adds suffixes that indicate gender—masculine, feminine, and common in the dual and plural forms, and masculine, feminine, and neuter (*i.e.*, undetermined, with affix *a* or no affix) in the singular—to roots that take noun (nominal) endings. the same gender suffixes are added to the 3rd person pronouns (originally demonstratives). In grammatical gender, masculine implies "long, narrow, strong," and feminine suggests "short, round, broad, weak." For example, */u* or */uʘma* "canoe" is masculine, while */uʘhɛ* "ferry" is feminine. Gender distinction is also found in the 1st persons (except singular) and 2nd persons.

Masculine and feminine gender

There are six tenses for the verb, with the present tense also expressing whether the subject is standing, sitting, or lying. Unlike the other Khoisan groups, the Central Khoisan languages have a passive form, which is frequently used. In sentence order, subject–object–predicate is the most usual, but subject–predicate–object is also common, and object–subject–predicate occurs if the subject is a pronoun or a nominal.

Vocabulary. Khoisan vocabulary is highly adapted to the needs of life in a poor natural environment, especially to the hunt, to the gathering of food, and to all skills serving the preservation of life. As a result of contact with other cultures, loanwords have been adopted—*e.g.*, from European languages into Nama, from Bantu into Kxoe. Native Khoisan creations that reflect non-native innovations also occur; *e.g.*, Kxoe */am-mũ-'o-xo* "clock, watch" is literally "sun-see-on-thing." The use of words in a figurative sense to label new objects is another process for word-formation. Examples include Kxoe *yɛ-≠'am* "bridge," which is literally "hole upper side," and *kuru* "to drive [a car]," which derives from "to press the bellows." Gender symbolism also plays a role in the formation of words for objects in the contact cultures; *e.g.*, the Nama masculine term */nū-b* means "leg," while the feminine counterpart */nū-s* means "wheel."

Basic numbers exist for 1 and 2 and, in some languages, for 3; higher numbers, however, are circumscribed. For example, the number 4 in Kxoe is literally "lick hand bone quantity"—that is, "the finger with which one licks out a pot." The numbers in the languages of the Bushmen seldom go beyond 10; this contrasts greatly with the practice in Nama, and in Hottentot in general, in which there is a decimal system with noncompound numerals from 1 to 10.

In general, the unit of the vocabulary equivalent to the word in European languages is a root whose category as a noun or verb is determined by the context and by the use of noun and verb affixes or particles. Compound words are often used to extend meaning, as English derived words do (*e.g.,* in Kxoe, *kx'â-xò,* literally "drink [verb]-thing" means "beverage").

Writing and texts. About 10 Khoisan languages have been recorded with various degrees of intensity by missionaries and linguists. The best known are Nama, Kxoe, !Khung, and /Kham, followed, to a lesser degree, by !Kora, ≠Khomani, ⫽Khegwi, Naro(n), and /'Auni. The old individual methods for recording clicks were superseded by the Lepsius system (1854), which is, with some modifications, still in use. Studies in Kxoe, !Khung, !Khong, / Guikhoe, and Sandawe have been carried out in recent times, with current research concentrating on !Khung. In view of the difficulties of research among vanishing bands of peoples, the prospects of gathering sufficient material in the remnant Khoisan languages are not good.

Folklore collections

The largest collection of folkloristic texts was taken down in the now extinct /Kham language of the Cape of Good Hope. Next to /Kham, the most comprehensive collection of texts on history, folklore, and traditional and modern life was recorded in the Kxoe tongue. Texts on folklore and history also exist in Nama and Korana and to some extent in !Khung. Literacy is found only in the Nama language, among the Nama and Bergdama. In !Khung, the first two primers were printed in 1969. (O.R.A.K.)

LANGUAGES OF THE AMERICAS

Eskimo-Aleut languages

Eskimo and Aleut, once neighbour languages on the Alaska Peninsula, are related but quite distinct; together they form the Eskimo-Aleut language family.

Eskimo is spoken in Greenland by a native population of about 45,000, in Arctic Canada by 17,000 or more, in Alaska by some 37,000 native inhabitants, and on the coast of the Chukchi Peninsula in northern Siberia by about 900 of the Siberian Eskimos. In Greenland, Eskimo is used in the schools, in the church, and on the radio; in Canada and in Alaska the language is employed mainly as a means of religious education. It is the language of instruction in the first grade of primary school in the Soviet Eskimo area. There are numerous Eskimo dialects; between them there is a more or less clear-cut break at the Bering Strait and at Norton Sound in Alaska. Within each of the dialect chains the neighbouring dialects are mutually intelligible, but the cumulative differences impede or prevent understanding between the dialects at the geographical extremes.

Intelligibility of dialects

Aleut, now greatly reduced in number of speakers, is the smallest branch of the family. It is spoken in the Aleutian Islands and in the Pribilof Islands in the Bering Sea, settled in the 1820s, by fewer than 1,000 Aleuts; on the Soviet Commander Islands, settled in 1826, there are fewer than 100 Aleut speakers (1970). Three mutually intelligible Aleut dialects are left, although one is nearly extinct. The Eskimo language is unintelligible to the speakers of Aleut, and vice versa.

Classification. The Eskimo languages are usually classified into two main branches—an Inuk, or Inupik, branch, which spreads eastward from the Bering Strait and the Norton Sound, and a twofold Yuk, or Yupik, branch, consisting of Asian (Siberian) and Alaskan varieties. This grouping is based on the dialectal forms of the term for people or real people that most Eskimos use to refer to themselves. The exact interrelationship of the dialects remains to be clarified.

The term Aleut, introduced by Russian traders from Kamchatka in northeastern Siberia after 1741, is commonly used both for the Aleuts proper, who call themselves Unangan (or, on the island of Atka, Unangas), and for the Eskimos of former Russian Alaska. The linguistic relationship between Eskimo and Aleut, sensed as early as the 1820s by the Danish linguist Rasmus Rask, could be proved only through a fairly advanced analysis of both languages. The proposed relationship of Eskimo-Aleut with other language families, such as Chukchi-Kamchadal, Uralic, and/or Indo-European, remains speculative.

Relationship between Eskimo and Aleut

Alphabets and orthography. The first book in Eskimo was published in 1742 by Hans Egede, a Dano-Norwegian missionary to Greenland. It was printed in the Roman alphabet, with the letter *r* used to represent the Eskimo uvular fricative (a sound involving friction in the airstream, made at the back of the mouth with the appendage known as the uvula). The Greenlandic orthography was systematized in 1851 by Samuel Kleinschmidt, a German of the Moravian Brethren. He introduced a small capital ᴋ to distinguish the uvular stop, made with stoppage of the airstream by contact of the uvula and the back of the tongue (also written *q*), from the velar sound *k*, made with the back of the tongue touching the velum, or soft palate (like the English *k*). Three accents were introduced to differentiate long sounds from short ones; *e.g., anâna* (equivalent to *anaana*) "mother," *mána* (equivalent to *manna*) "this," *mãna* (equivalent to *maanna*) "now." That is, the circumflex,ˆ, marks a long vowel, the acute accent, ´, indicates a long consonant following the marked vowel, and the tilde,˜, notes a long vowel and a long succeeding consonant. The Kleinschmidt orthography is still used today, and similar orthographies have been employed by the Moravian Mission in Labrador since 1800.

In Russian Alaska, an adequate Cyrillic alphabet was designed for the Aleut language by the Orthodox missionary Ivan Veniaminov in about 1830. In 1848 it also became used for southern Eskimo. Books in both languages were published by the Russian Orthodox Church in America until 1903.

Adaptation of the Cree syllabary

In 1878 the syllabic characters originally designed for the Cree Indians were introduced to the Eskimos of central Canada by the Rev. Edmund J. Peck of the Church Missionary Society. Since 1953 this form of writing has also been used in an Eskimo periodical sponsored by the Canadian government, alongside a new, tentative orthography in the Roman alphabet.

The ordinary Roman alphabet was used for Alaskan Yupik by Moravian missionaries from Pennsylvania in the 1920s. From 1948 on, Protestant missionaries in northern Alaska developed an alphabet with seven additional letters. In 1961 a program was started at the University of Alaska, with the active participation of Yupik Eskimos, for working out a systematic Eskimo orthography in the Roman alphabet to be used in Alaskan public schools from 1971 on. In the Soviet schools for Asian Eskimos a Roman alphabet with two additional letters was introduced in 1932, but it was replaced by the Russian Cyrillic alphabet in 1937.

Phonological characteristics. Eskimo and Aleut have relatively simple systems of distinctive sounds. The accent (stress) depends upon the length of the syllables and never has independent value as in English.

In Eskimo there are four distinctive vowels. Three of them, usually written as *a, i, u,* vary considerably in pronunciation according to the consonants that follow or precede. These three vowels occur both short and simple, and combined into long vowels and diphthongs. The vowel combinations result primarily from the loss of former intervocalic consonants; *e.g.,* Greenlandic *ûvoq,* the equivalent of *uu-vuq* "is burnt," is related to the Asian Yupik form *ugu-,* which has a fricative *g* sound between the two vowels. The fourth Eskimo vowel is represented in Yupik by a short *e* and is pronounced somewhat like the vowels in "but" or "bird." In the Inupik area it has become identical with the vowel written *i*. The Aleut language has only *a, i,* and *u* in both short and long varieties. The fourth vowel has become identical with one of the first three or has been dropped from the language.

The consonantal system

Of consonants, Eskimo has from 13 to 21, depending on the dialect. The stop sounds include the uvular *q,* the velar *k,* the dental *t,* made with the tip of the tongue touching the back of the upper teeth, and the labial *p,* made with the lips; in Alaskan Yupik there is also a palatal *c* (like English *ch*), to which an *s* corresponds in the other dialects. In parts of Canada this has changed to *h*. The nasal sounds, made with the breath passing through the nose, include ŋ (as in "sing"), *n, m,* and, in Asian Yupik, also a voiceless *n,* made without vocal cord vibration. Voiced and voiceless varieties of the continuant consonants *r, g, l,* and *v* are distinctive sounds in the western dialects but in eastern Inupik they are only variants. In addition to *j,* pronounced as *y* in English "year," some dialects have sounds similar to English *r* or *z* or to *sh* (in Greenlandic written *ss*). Corresponding to *ss* in Aleut there is a fricative *d* (pronounced as the *th* in "that"); *e.g.,* Aleut *da-* "eye" is related to Eskimo *izi* and *ii* and to Greenlandic *isse.* Aleut shares with Eskimo most of the consonants articulated with the tongue, including *c* (the English *ch* sound) and *s,* but has *p* and labial fricatives (*f* and *v*) only in loanwords from Russian or English. Aleut *m* corresponds with both Eskimo *m* and *v;* to Eskimo *p* corresponds the Aleut *h* (in initial position) and the Aleut aspirated nasal sound *hm* (in noninitial positions)—*e.g.,* Aleut *hum-* "to swell" corresponds to Eskimo *puvi-;* Aleut *ahmat-* "to ask" is cognate with Yupik *ap(e)t-*. (An aspirated sound is one pronounced with an accompanying puff of air.)

In initial position, Inupik uses only a single stopped consonant (*i.e., p, t, k,* or *q*), or *s, n,* or *m* (rarely *l*); between vowels it employs at most two consonants, including double ones. In contrast, Yupik and Aleut have initial consonant clusters, resulting from the loss of a vowel in the first syllable from an older historical form; *e.g.,* Yupik and Aleut *sla* "weather," Inupik *sila.*

Grammatical characteristics. Eskimo has a great number of suffixes but practically no prefixes or compounds. In Aleut the word forms are simpler, but the syntactic constructions (the way the words are arranged in sentences) are more complicated. Suffixes are often accompanied by changes in the stem, especially in Eskimo; *e.g., nanuq* "polar bear," dual *nannuk* "two polar bears," plural *nannut* "several polar bears"; *inuk* "person," dual *innuk,* plural *inuit; umialik* "owner of boat (*umiaq*), chief," dual *umiallak,* plural *umialgit* (North Alaska dialect). Note the doubled consonants in the dual and plural forms of some terms, and the other consonant and vowel alternations.

Grammatical numbers—singular, dual, plural—combine with suffixes for person—*e.g., ulu-ga* "my knife," and *ulu-t-ka* "my knives," in which *-t-* means "several" and *-ga* or *-ka* "my." The possessor of someone or something occurs in the so-called relative, or subordinative, case—*e.g., umialgum pania* "the chief's (his) daughter," in which *pani-a* means "his or her daughter"; this is distinguished from the reflexive *panni* "his or her own daughter" and the stem *panik.*

Personal pronouns

The personal pronouns consist of a stem and person suffixes—*e.g.,* Aleut *ti-*ŋ "I, me," with the ending -ŋ as in *ada-*ŋ "my father"; Eskimo *uva-*ŋ*a* "I, me," with a demonstrative stem (*uva*) as in *uva-ni* "here." Corresponding to the separate personal pronoun forms in Aleut are the Eskimo suffixes for the 1st, 2nd, and reflexive 3rd persons (himself, etc.). These indicate the subject of an intransitive verb or the object of a transitive verb—*e.g.,* -ŋ*a* means "I" in *uqaqtu-*ŋ*a* "I said," and "me" in *uqautigaa-*ŋ*a* "he told me." A noun that would be the subject of "said" (the intransitive verb) or the object of "told" (the transitive verb) is in the absolute case (it is not marked with a suffix), whereas the subject of a transitive verb (*e.g.,* "tell") occurs in the subordinative case. For example, compare the unmarked form for "chief," *umialik,* in *umialik uqaqtuq* "the chief said" with its corresponding form *umialgum* in the subordinative case in *nukatpiarzuk umialgum uqautiga-a* "the young man (object), the chief told him" (*-a* "he - him"). With a 3rd person object ("him, them"), a transitive verb has largely the same suffixes as a noun—*e.g.,* the *-t-* and *-ka* in *ulu-t-ka* "my knives" and *tigu-gi-t-ka* "I took them" (*-gi-* is indicative) and the corresponding Aleut forms, *-ni-* and -ŋ in *ukina-ni-*ŋ and *su-ku-ni-*ŋ.

Eskimo nouns and pronouns have six adverbial cases, expressing relations such as "in," "to," "from," "along," "with," and "like"—*e.g., iglu-mi* "in the house," *iglu-pti*ŋ*-ni,* "in our house." In the Yupik dialects there are only five such adverbial cases. These are reduced to two (in/to, from/along) in Aleut and are limited to pronouns and relation words—*e.g., ula-m nag-a-n* "of house (*ula-m*) in its (*-a-*) interior"; this corresponds to Eskimo *iglu-m ilu-a-ni* "in the house."

Verbal modes include indicative ("he goes"), interrogative ("did he go?"), imperative ("go!"), optative ("may he go"), participles ("going, gone"), and other forms corresponding to subordinate clauses in English—*i.e.,* clauses beginning with "if," "when," etc. Other modal relations and tenses are specified by derivational suffixes, and in Aleut also by auxiliary verbs; *e.g., haqa-l sara-nar* "coming he slept" is equivalent to "he came yesterday." An Eskimo derivative form may also correspond to an English complex sentence—*e.g., tikit*$_1$*-qaar*$_2$*-mina*$_3$*-it*$_4$*-ni*$_5$*-ga*$_6$*-a*$_{7-8}$ is "he (A)$_8$ said$_{5-6}$ that he (B)$_7$ would not$_4$ be able$_3$ to arrive$_1$ first$_2$," or, in exact Eskimo order, "to arrive first be able would not said him he."

Vocabulary. A remarkable feature of the vocabulary is the great number of demonstratives, about 30 in western Eskimo and in Aleut. For example, in Aleut there is *hakan* "that one high up there" (as a bird in the air), *qakun* "that one in there" (as in another room), and *uman* "this one unseen" (heard, smelled, felt).

The vocabulary naturally has its local particularities, the various groups having lived under very different conditions. The Eskimo word that means "meat" from Greenland to Siberia means "fish" south of Norton Sound and also in the Aleut language. Word taboo has also played its part, as in East Greenlandic, in which the general Eskimo-Aleut word for "eye" (*da-, izi,* see above) has been replaced by *uitsatai (ui-sa + uta-i)* "those by which he keeps gazing."

Word formation by derivation

Eskimo-Aleut derived words (*e.g.,* similar to English "winter-ize" or "anti-dis-establish-ment-ari-an-ism") correspond quite often to simple, nonderived English words. The possibility of derivation is virtually unlimited in the languages, and the number of word stems is comparatively small; *e.g.,* there are fewer than 2,000 in well-known West Greenlandic. Examples of derivatives are: *nalu-voq* "is ignorant," *nalu-vâ* "does not know it," and *nalu-na-er-* "make not (*-er-*) to be (*-na-, -nar-*) ignored," which is equivalent to "communicate." *Nalunaer-asuar-ta-ut* "that by which (*-ut*) one communicates habitually (*-ta-, -tar-*) in a hurry (*-asuar-*)" is the form for "telegraph," a term coined in the 1880s.

In Greenlandic there are four loanwords from medieval Norse; from the colonial period after 1721 there have been surprisingly few borrowings until recently. In Aleut and in the Eskimo of former Russian Alaska there are many borrowings from Russian, and there are several in Asian Eskimo from English, and many from Chukchi, a Paleo-Siberian language. Notable Eskimo contributions to the vocabulary of English and other European languages are "igloo" (or *iglu*) and "kayak" (*qayaq*). (K.B.)

North American Indian languages

The term North American Indian languages usually refers to those languages that are indigenous to the United States and sub-Arctic Canada, and that are spoken north of the Mexican border. A number of language groups within this area, however, extend as far south as Central America. The present article will concentrate on the languages of Canada and the United States. (For further information on languages of Mexico and Central America, see *Meso-American Indian languages*; for languages of Arctic America, see above *Eskimo-Aleut languages*).

The Indian languages of North America are both numerous and diverse. Their original number has been estimated at 300; these tongues were spoken by a native population of approximately 1,500,000. The number of languages still used was estimated at around 200 by the American linguist Wallace Chafe in 1962. Some of these had only one or two elderly speakers. The numbers continue to drop, but with some notable exceptions—*e.g.,* Navajo is steadily increasing in number of speakers. As a consequence of the growing trend toward extinction in the American Indian languages, the field of study is becoming more concerned with the past than the future. Even so, the rich diversity of these languages provides a valuable laboratory for linguistic theory; certainly the discipline of linguistics could not have developed as it has, especially in the United States, without the native American languages. In this article, the present tense will be used in referring to both extinct and surviving languages.

Within the diversity of the North American Indian languages, no general characterization is possible; various features of structure are common to them, but there is no feature or complex of features shared by all. At the same time, there is nothing primitive about these languages. They draw upon the same linguistic resources and display the same regularities and complexities as do the languages of Europe. If historical connections are sought among the Indian tongues, some languages clearly show numerous and systematic resemblances comparable to those between Spanish, French, and Italian. These similarities strongly suggest classification as a linguistic family. North American languages can then be grouped into some 57 families. On this level, too, the diversity of some areas is notable. Thirty-seven families lie west of the Rockies and 20 in California alone; California thus shows more linguistic variety than all of Europe. Some families seem to be related to each other in more remote historical groupings, often called phyla. Such classifications border on speculation, however, partly because data are lacking on many languages (because they are extinct or still unstudied), and partly because of the difficulty in distinguishing, at the deeper historical levels, between resemblances caused by common origin and those resulting from linguistic borrowing.

Language families and phyla

In any case, no theory of common origin for the North American languages has become established. Although most anthropologists believe that North America was populated mainly by people who migrated across the Bering Strait from Asia, attempts to relate native American languages to Asian languages have not gained general acceptance. (There is one possible exception—the relationship of Eskimo-Aleut to certain Siberian languages.) The linguistic diversity of North America suggests, indeed, that the area was populated as a result of several waves of migration by peoples of distinct linguistic stocks of Asia; these stocks may have no modern survivors.

Classification. The first comprehensive classification into families of the North American Indian languages was made in 1891 by the American John Wesley Powell, who based his study on impressionistic resemblances in vocabulary. A principle of nomenclature adopted by Powell has been widely used ever since: families are named by adding *-an* to the name of one prominent member; *e.g.,* Caddoan is the family including Caddo and other languages. For this most obvious level of relationship, the Powell classification remains essentially unchallenged. Various scholars, however, have attempted to group the families into larger units that reflect deeper levels of historical relationship. Of these efforts, one of the most ambitious and best known is that of Edward Sapir, which was first published in the *Encyclopædia Britannica* in 1929. In Sapir's classification, all the languages are grouped into six phyla—Eskimo-Aleut, Algonkian-Wakashan, Na-Dené, Penutian, Hokan-Siouan, and Aztec-Tanoan—established on the basis of very general grammatical resemblances. In 1958, research of the American linguist Mary R. Haas revealed precise sound correspondences between the Algonkian languages and a "Gulf" group in the southeastern United States that Sapir had assigned to the Hokan-Siouan phylum. Since that time, various reconsiderations of Sapir's groupings have been proposed. A classificatory map published by Charles F. and Florence M. Voegelin in 1966 offers one such classification, and it is likely to serve as a standard reference point for some time. Although preserving Sapir's Eskimo-Aleut, Na-Dené, Penutian, and Aztec-Tanoan groups, it also proposes reconstituted Macro-Algonkian, Macro-Siouan, and Hokan phyla, and allows nine families to remain unclassified, pending further research.

Sapir's six phyla

Table 60, based on the Voegelin map, gives approximate indications of the aboriginal home territories and of the number of speakers estimated from published data in the early 1980s.

Language contact. The Indian languages of North America, like all languages in the world, have always existed in contact with other tongues. From this situation bilingualism, or multilingualism, has resulted; the extent is determined by sociological factors. The Indian languages show varying degrees of linguistic acculturation; *i.e.,* there may be borrowing between languages not only of vocabulary items, but also of phonological, grammatical, and semantic features. In aboriginal times, in areas where bilingualism was most important (*e.g.,* the Northwest), there tended to be well-defined linguistic areas in which languages of diverse genetic affiliations came to share numerous structural characteristics through the process of borrowing. As noted above, such phenomena create difficulties for attempts at genetic classifications. In a few cases, situations of language contact have given rise to a pidgin or compromise language that is composed of elements from various sources and is used as a second language, especially in trading. An example is the Chinook Jargon of the Northwest; this came to be used by many whites and absorbed many loanwords from French and English before its eventual obsolescence.

Pidgin languages

In more recent times, contact of Indian languages with European languages—French, English, Spanish, and Russian—has again resulted in bilingualism. With the Indian languages generally relegated to a socially subordinate position (and with many of them headed for extinction), borrowing, however, has involved the relatively superficial level of vocabulary more often than the deeper levels of language structure, such as the sound system or grammar. The effects on European languages are apparent mainly in place names like Massachusetts and Seattle and in names like squash and abalone for native American plants and animals. Among the Indians, the type and degree of linguistic adaptation to European culture has varied greatly, depending on sociocultural factors. For example, among the Karok of northwestern California, a tribe that suffered harsh treatment at the hands of whites, there are only a few loanwords from English (*e.g., ápus* "apples"), a few calques or loan translations (the "pear" is called *vírusur* "bear," because English "pear" and "bear" are merged in Karok pronunciation), but a large number of new formations from native materials; *e.g.,* a hotel is called *amnaam* "eating place."

Grammar. The term grammatical structure as used here refers to both the traditional categories of morphology—how words are made up—and syntax—how words are combined into sentences. It should again be emphasized that in grammar, as well as in phonological or semantic structure, neither the American Indian languages nor any other languages in the world display anything that could be called primitive in the sense of undeveloped or rudimentary. Every language has a structure as complex, as subtle, and as efficiently adaptable to cultural needs as that of Latin or English, for example.

The North American Indian languages display great diversity, so that it is not possible to characterize them as a group by the presence or absence of any particular grammatical peculiarities. At the same time, there are some characteristics that, though not unknown elsewhere in the world, are sufficiently widespread to be considered typical of the continent or of particular linguistic areas within North America. The phenomenon of polysynthesis, in which many sentence elements are expressed within the boundaries of a single word by compounding and affixation, is especially characteristic of Eskimo and Algonkian, but is also found elsewhere. An illustration from the Algonkian group is the Menominee form *nekeespestεh-wenah-nεεwaaw* "but I did see him on the way." Incorporation, the compounding of a noun with a verb, is rarely used in English (*e.g.,* "to baby-sit") but is common in some Indian languages; *e.g.,* Mohawk *ke-wẽna-weiẽhõ* "I-language-understand." (The symbols used that are not found in the Latin alphabet have been adopted from phonetic alphabets.)

Polysynthesis and incorporation

Some especially common characteristics of North American languages are the following:

1. In verbs, the person and number of the subject are commonly marked by prefixes; *e.g.,* Karok has *ni-'áhoo* "I walk," *nu-'áhoo* "he walks." In some languages, the prefix simultaneously indicates the object as well as subject; *e.g.,* Karok *ni-mmah* "I see him," *ná-mmah* "he sees me."

2. Tense and aspect of verbs are usually marked by suffixes, as in many languages throughout the world. But in some areas—*e.g.,* among the Athabascan languages—prefixes are used. For example, Chipewyan *hε-tsa*ɣ means "he is crying," ɣ*ĩ-tsa*ɣ is "he cried," and ɣ*wa-tsa*ɣ is "he will cry."

3. In noun forms, the concept of possession is widely expressed by prefixes indicating the person and number of the possessor. Thus Karok has *ávaha* "food," *nani-ávaha* "my food," *mu-ávaha* "his food," etc. When the possessor is a noun, as in "man's food," a construction like *ávansa mu-ávaha* "man his-food" is used. Many languages have inalienable nouns, which cannot occur except in such possessed forms. These generally designate such things as kinsmen or body parts; *e.g.,* Luiseño, a language in Southern California, has *no-yó'* "my mother," *o-yó'* "your mother," but no word for "mother" in isolation.

4. Nouns in many languages have forms with a meaning of location; *e.g.,* Karok *áas* "water," *áas-ak* "in the water." Such a construction is reminiscent of the case forms of Latin, and case systems do indeed occur in California and the southwest. For example, Luiseño has the nominative *kíiča* "house," accusative *kíiš,* dative *kíi-k* "to the house," ablative *kíi-ŋay* "from the house," locative *kíi-ŋa* "in the house," instrumental *kíi-tal* "by means of the house."

The following five grammatical features are less typically North American, but are nevertheless distinctive of many areas. First person pronouns in many languages show a distinction between a form inclusive of the addressee—"we" denoting "you and I"—and an exclusive form—"I and someone other than you." Some languages also have a distinction in number between singular, dual, and plural pronouns. Reduplication, the repetition of all or part of a stem, is widely used to indicate distributed or repeated action of verbs; *e.g.,* in Karok, *imyah* means "breathe," *imyáhyah* means "pant." In Uto-Aztecan languages, reduplication sometimes is associated with plural nouns, as in Pima *gogs* "dog," *go-gogs* "dogs." In many languages, verb stems are distinguished on the basis of the shape or other physical characteristics of the associated noun; thus in Navajo, in referring to motion, *'á*^{n} is used for round objects, *tá*^{n} for long objects, *tí*^{n} for living things, *lá* for ropelike objects, etc. Similar distinctions may refer to dual and plural number. Karok has *ikpuh* "one swims," *iθpuh* "two swim," *ihtak* "several swim."

Inclusive and exclusive pronouns

Verb forms also frequently specify the location or direction of an action by the use of prefixes or suffixes. In Karok, for example, from *paθ* "throw" is derived *páaθ-roov* "throw upriver," *páaθ-raa* "throw uphill," *paaθ-rípaa* "throw across-stream," and as many as 38 other similar forms. Some languages also specify the instrument of an action, generally by prefixation; *e.g.,* Pomo *phi-de-* "to move by batting with a stick," *phu-de-* "to move by blowing," *pha-de-* "to move by pushing with the end of a stick." Lastly, many languages have evidential forms of verbs that indicate the type of validity of the information reported; such distinctions may assume the importance played by tense and aspect in European languages. Thus Hopi distinguishes *wari* "he ran, runs, is running" as a reported event, from *warikŋwe* "he runs (*e.g.,* on the track team)," which is a statement of general truth, and from *warikni* "he will run," which is an anticipated event. In other languages verb forms consistently discriminate hearsay from eye-witness reports. Such a system might be very welcome in other societies; *e.g.,* especially as regards the reliability of news reports.

Phonology. The languages of North America are as diverse in their systems of pronunciation as they are in other ways. In terms of the number of contrasting sounds (phonemes), the Northwest Coast is characterized as a linguistic area by the unusual richness of its systems. A language like Tlingit has approximately 50 consonants and vowels (a comparable count for English would number 35). By contrast, Karok has only 23. The richest sound inventories seem to occur where bilingualism was commonest, and sounds were borrowed between languages.

The large number of consonants that is found in many Indian languages is based on the use of a number of phonetic contrasts that are relatively unfamiliar in European languages. In English, different consonants are produced by vibrating the vocal cords (which results in voiced sounds) or by not vibrating them (which gives unvoiced sounds); by shutting off the air momentarily, thus producing stops, or by letting the airstream pass through the mouth with friction (producing fricatives); and by placing the tongue in a variety of positions. The Indian languages also use

Consonant features

Table 60: North American Indian Languages*

phyla, families, languages	location	speakers remaining†
American Arctic-Paleosiberian		
Eskimo-Aleut		
Chukchi-Kamchatkan (in Siberia)		
Na-Dené		
Athabascan		
Dogrib, Bear Lake Hare	N.W.T.	1,400
Chipewyan, Slave, Yellowknife	N.W.T.	4,400+
Kutchin	Yukon, Alaska	800
Tanana, Koyukon, Han, Tutchone	Alaska	1,450+
Sekani, Beaver, Sarsi	Alberta	450+
Carrier, Chilcotin	B.C.	1,500+
Tahltan, Kaska	N.W.T.	300+
Tanaina, Ingalik, Nabesna, Ahtena	Alaska	1,500+
Eyak	SC Alaska	‡
Chasta Costa, Galice, Tututni	SW Oregon	‡
Hupa	NW Calif.	1,200
Kato, Wailaki	NC Calif.	235
Mattole	NC Calif.	§
Tolowa	NW Calif.	125
Navajo	Ariz., N.M.	137,400+
Western Apache	W Arizona	10,000+
Chiricahua, Mescalero Apache	S New Mexico	2,900+
Jicarilla Apache	N New Mexico	1,000
Lipan Apache	Texas	‡
Kiowa Apache	Oklahoma	‡
Tlingit	SE Alaska	1,500+
Haida	B.C.	700
Macro-Algonkian		
Algonkian		
Cree, Naskapi, Montagnais	E Canada	89,200
Menominee	Great Lakes area	2,200+
Fox-Sauk-Kickapoo	south of Great Lakes	1,000
Shawnee	SC U.S.	2,500+
Potawatomi	Michigan	3,500+
Ojibwa, Ottawa, Algonkin Salteaux	S Ontario	25,000+
Delaware	C Atlantic coast	3,200+
Penobscot, Abnaki	New England	300+
Malecite, Passamaquoddy	New England, Maritime Provinces	1,200+
Micmac	Maritime Provinces	2,100+
Blackfoot	Mont., Alberta	1,000+
Cheyenne	E Wyoming	4,000+
Arapano, Atsina, Nawathinehena	E Colorado	1,000+
Yurok	NW Calif.	1,900+
Wiyot	NW Calif.	1
Muskogean		
Choctaw Chickasaw	N Mississippi	18,000+
Alabama, Koasati	Alabama	700+
Mikasuki, Hitchiti	NW Florida	1,000
Muskogee (Creek), Seminole	Georgia	8,500+
Natchez	N Louisiana	§
Atakapa	SW Louisiana, SE Texas	§
Chitimacha	S Louisiana	§
Tunica	N Louisiana	§
Tonkawa	E Texas	§
Macro-Siouan		
Siouan		
Crow	E Montana	4,600
Hidatsa	North Dakota	1,000+
Winnebago	Wisconsin	1,000+
Omaha, Osage, Ponca, Kansa, Quapaw	central plains	3,400+
Dakota (Sioux)	northern plains	30,000+
Tutelo, Ofo, Biloxi	Gulf Coast	§
Catawba	Carolinas	§
Iroquoian		
Seneca, Cayuga, Onandaga	New York	6,800+
Mohawk	New York	6,700+
Oneida	New York	6,200+
Wyandot (Huron)	SE Ontario	§
Tuscarora	North Carolina	600+
Cherokee	southern Appalachians, Oklahoma	27,000
Caddoan		
Caddo	Arkansas	1,200+
Wichita	Texas, Okla.	500+
Pawnee	Kansas	1,800+
Arikara	Dakotas	300+
Yuchi	southern Appalachians	200+
Hokan		
Yuman		
Walapai, Havasupai, Yavapai	NW Arizona	600+
Mohave, Yuma	lower Colorado River	2,000
Delta Yuman (Cocopa)	delta of Colorado River	300+
Diegueño, Kiliwa	S Calif., Baja Calif.	75+
Seri	Sonora	200
Pomo	NC Calif.	
Northern Pomo		40
Northeast Pomo		1
Central Pomo		40
Southwest Pomo		50
Southeast Pomo		10
Southern Pomo		10
Palaihnihan	NE Calif.	
Achomawi		40+
Atsugewi		‡
Shastan	NE Calif.	‡
Yanan	NC Calif.	§
Chimariko	NW Calif.	§
Washo	EC Calif., Nevada	1,100
Salinan	WC Calif.	360
Karok	NW Calif.	750+
Chumashan	S Calif.	360
Comecrudan	S Texas, NW Mexico	§
Coahuiltecan	S Texas, NW Mexico	§
Esselen	WC Calif	§
Branches in Meso-America		
Penutian		
Yokutsan	SC Calif.	800+
Maiduan	NC Calif.	2,500+
Wintun	NC Calif.	
Patwin		50+
Wintu, Nomlaki		1,200+
Miwok-Costanoan		
Miwok, (Sierra, Coast-Lake)	SC Calif., C Calif.	2,500
Costanoan	WC Calif.	§
Klamath-Modoc	SC Oregon	2,500+
Sahaptian		
Sahaptin (Klikitat, Umatilla, Walla Walla, Warm Springs, Yakima)	NC Oregon	1,400+
Nez Perce	WC Idaho	500+
Cayuse	NE Oregon	§
Molale	NC Oregon	§
Coos	SW Oregon	50
Yakonan	WC Oregon	
Alsea		§
Siuslaw, Lower Umpqua		§
Takelma	SW Oregon	§
Kalapuya	WC Oregon	‡
Chinookan	NW Oregon, SW Wash.	20
Tsimshian	WC B.C.	3,000
Zuni	WC New Mexico	3,000+
Latin American branches		
Aztec-Tanoan		
Kiowa-Tanoan		
Tiwa	NC New Mexico	5,300+
Tewa	NC New Mexico	2,400
Towa	NC New Mexico	2,000
Kiowa	Oklahoma	1,000
Uto-Aztecan		
Mono	EC Calif.	300+
Northern Paiute (Paviotso), Bannock, Snake	NE Calif. SE Ore., N Nev., S Idaho	3,500
Panamint, Gosiute, Shoshone	C Nev., N Utah, SW Wyo.	5,000+
Comanche	N Texas	800
Kawaiisu, Ute, Chemehuevi, Southern Paiute	SE Calif., S Nevada, S Utah, SW Colo.	3,000+
Hopi	N Arizona	7,900+
Tubatulabal	SC Calif.	‡
Luiseño	S Calif.	500+
Cahuilla	S Calif.	800+
Cupeño	S Calif.	‡
Serrano	S Calif.	‡
Pima-Papago	S Arizona	25,400+
Latin American branches		
Unclassified		
Keresan	New Mexico	15,800
Yukian	NC Calif.	
Yuki		‡
Wappo		‡
Beothuk	Newfoundland	§
Kutenai	Mont., Idaho, B.C.	600+
Karankawa	SE Texas	§
Chimakuan	NW Washington	
Quileute		200+
Chemakum		§
Salish		
Lilloet	C B.C.	1,000+
Shuswap	E B.C.	1,000+
Thompson	C B.C.	1,000+
Okanagon, Sanpoil, Lake, Colville	S B.C.	1,700+
Pend d'Oreille, Flathead, Spokan, Kalispel	N Idaho	600+
Coeur d'Alene	N Idaho	400
Middle Columbia, Wenatchee	E Washington	30
Tillamook	NW Oregon	100
Twana	NW Washington	1,000
Upper Chehalis, Cowlitz, Lower Chehalis, Quinault	W Washington	1,700+
Southern Puget Sound Salish	W Washington	50+
Straits Salish	SW B.C.	2,300
Halkomelem	SW B.C.	1,000+
Squamish	SW B.C.	100+
Comox, Sishistl	Vancouver Is.	‡
Bella Coola	WC B.C.	200+
Wakashan		
Nootka	Vancouver Is.	1,000+
Nitinat	Vancouver Is.	10+
Makah	NW Washington	1,500
Kwakiutl	WC B.C.	1,000
Bella Bella, Heiltsuk	WC B.C.	100+
Kitamat, Haisla	WC B.C.	100+
Timucua	Florida	§

*Phyla given in boldface type; families given in italics (including those consisting of single languages); single languages, or dialect groups so closely related that they can be treated as single languages, given in roman type. †1981 estimate. ‡Minimal number of speakers; *i.e.*, under 10. §Extinct.

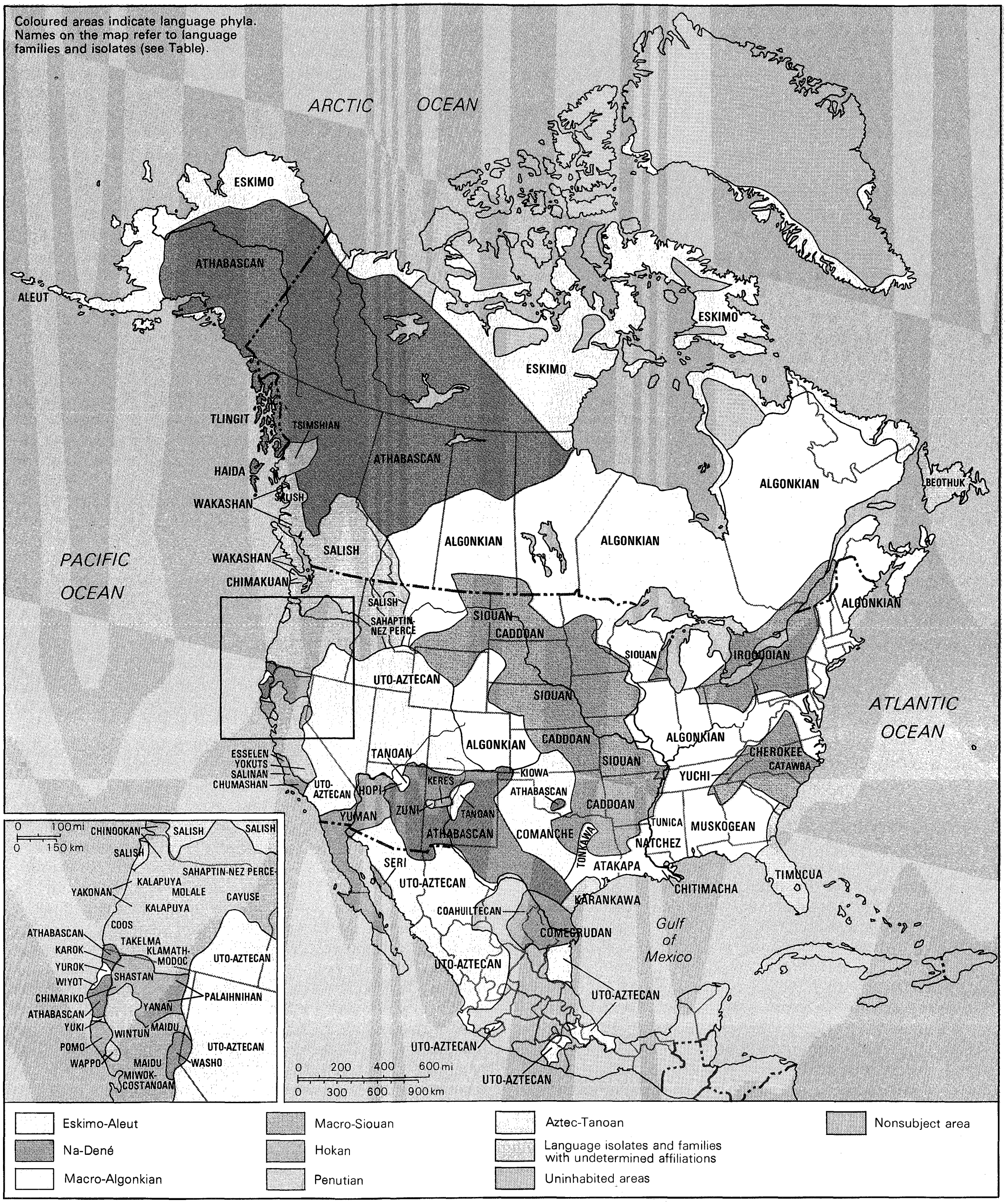

Coloured areas indicate language phyla. Names on the map refer to language families and isolates (see Table).

Figure 34: Distribution of North American Indian languages.

From C.F. and F.M. Voegelin, *Map of North American Indian Languages;* copyright 1966 by University of Washington Press

these mechanisms, but sometimes others as well. The glottal stop, an interruption of breath produced by closing the vocal cords (as in the middle of English *oh-oh!*) is a common consonant. A related phenomenon, widespread in western North America, is the use of glottalized consonants, as when a *t* is produced with near simultaneous closure and reopening of the vocal cords. This is recorded with an apostrophe; it differentiates terms like Hupa (Athabascan) *teew* "underwater" from *t'eew* "raw."

The number of consonantal contrasts is also frequently expanded by distinguishing a larger number of tongue positions than do most European languages. Many languages distinguish two types of velar sounds (sounds made with the back of the tongue)—a *k* much like an English *k*, and a uvular *q*, produced further back in the mouth. Some languages even differentiate three such *k* sounds—front, middle, and back. Labiovelars, velar sounds that have simultaneous lip-rounding, are also common. Thus Tlingit has 21 phonemes made in the velar area alone: *g*, *k*, uvular *G*, *q*, glottalized *k'*, *q'*, labiovelar g^w, k^w, k'^w, G^w, q^w, $q^{w'}$, in addition to the corresponding fricatives *ɣ* and *x*, with uvular *X*, glottalized *x'*, *X'*, and labiovelar x^w, X^w, x'^w, X'^w. In comparison, English has only two sounds, *k* and *g*, made in the same area of the mouth.

Another class of sounds common in North America, especially in the West, is that of the laterals, which are produced by stopping the breath with the central part of the tongue but allowing it to escape at the sides. Alongside the common lateral *l*, such as exists in English, many Indian languages have a voiceless counterpart, similar to the Welsh *ll;* this sound is approximated by the *thl* in northwestern place names such as Cathlamet. To this some languages also add glottalized varieties, as well as a close-knit *tl* unit, which may in turn be aspirated or glottalized, so that there may result, as in Navajo, a total of five distinguishable lateral sounds.

In some Indian languages, as in English, stress is significant in distinguishing the meaning of words. In others, musical pitch plays a linguistic function, as it does in Chinese; *e.g.,* in Navajo, *bíní'* is "his nostril," *bìnì'* is "his face," and *bìní'* is "his waist." (High and low pitches are indicated with the acute and grave accents, respectively.)

A peculiarity of some northwest coast languages is their use of complex consonant clusters, as in Bella Coola *tlk'*^{w}*ix*^{w} "don't swallow it." Some words even lack vowels entirely; *e.g., nmnmk'* "animal."

Phonological change

Processes of phonological change, in which differences of sound are associated with grammatical distinctions (as with English *f* and *v* in "half," "halves," "to halve"), are also found in North American languages. In some languages, for example, consonantal change is related to diminutive meaning: thus Luiseño *r* changes to *đ* in *ŋarúŋru-š* "pot," *ŋadúŋdu-mal* "pot-small." Vowel harmony, a process whereby vowels change to resemble adjacent ones, is further attested in North America. Yurok in northwestern California, for example, has an unusual *r* vowel, comparable to the sound in English "bird"; when this occurs in a suffix, stem vowels change to agree with it, thus *lo'σγ̥e* "black" + *-'r'y* (animate suffix) yields *lr'rγ̥r'r'y* "black animal."

Vocabulary. The word stock of American Indian languages, like those of other languages, is composed both of simple stems and of derived constructions; the derivational processes commonly include affixation (the use of prefixes, suffixes, etc.) in addition to compounding in some languages. A few languages use internal sound change, similar to the case of English "song" from "sing"; *e.g.,* Yurok *pontet* "ashes," *prncrc* "dust," *prncrh* "to be gray." New vocabulary items are also acquired by borrowing, as mentioned above.

It should be noted that, in languages generally, the meaning of a vocabulary item cannot be adequately inferred from a knowledge of its historical origin or from knowing the meaning of its parts. For example, the name of an early 19th-century trapper, McKay, entered Karok as *mákkay,* but with the extended meaning of "white man." It was then compounded with a native noun *váas* "deerskin blanket" to give the neologism *makáy-vaas* "cloth"; this in turn was compounded with *yukúkku* "moccasin" to give *makayvas-yukúkku* "tennis shoes." At each stage of vocabulary formation, meaning is determined not simply by etymology but also by arbitrary extensions or limitations of semantic value.

Semantic structure of Indian languages

It is in the area of semantic structure that American Indian vocabulary is likely to present some surprises to the investigator. It is frequently observed that the immense diversity of the physical universe is reduced by every society to a manageable set of classifications embodied in its vocabulary. But there are few universals in such classification, and every language makes its unique semantic divisions. One language may make many specific discriminations in a particular area, while another is content with a few general terms; the difference is correlated with the importance of the semantic area for the particular society. Thus English is highly specific in classifying bovines (bull, cow, calf, heifer, steer, ox), even to the point of lacking a general cover term in the singular (what is the singular of cattle?), but for other species it has only cover terms like camel, llama. North American Indian vocabularies, as would be expected, embody semantic classifications that reflect native American environmental conditions and cultural traditions.

Interest in the semantic classifications of American Indian languages, especially in Hopi, has been particularly stimulated by the work of the American investigator Benjamin Lee Whorf. When English discriminates "air-plane," "aviator," and "flying insect," Hopi generalizes with a single term *masa'ytaka,* roughly "flier"; but when English uses a single general term, "water," Hopi differentiates *pāhe* "water in nature" from *kēyi* "water in a container."

The vocabularies of different languages may differ not only in the categorization of particular items but also in the general principles of semantic organization; such differences may be found even between neighbouring languages in a single culture area. English, for example, tends to exhaust the universe of flora and fauna with multi-levelled hierarchical classifications such as "plant, bush, berry bush, gooseberry bush" or "animal, insect, louse, body louse," but the languages of northwestern California, by contrast, have relatively few generic terms and many vocabulary items that do not fall into any such hierarchy. The generic terms of Yurok refer, roughly, to "quadruped mammal," "fish," "snake," "bird," "tree," "bush," "grass," "flower," and "berry"; the organization in the neighbouring Tolowa language is simpler, lacking "quadruped mammal" and "fish." In such frameworks, a term like Yurok *wrrγ̥r* "body louse" cannot be subsumed in the larger classes of "louse" or "insect" because none exist. The placing of terms in semantic pigeonholes tends to be replaced, in these semantic systems, by identifying them in terms of similarity. A Yurok speaker, asked to identify a flowering bush for which he knows no name will describe it not as "a kind of bush," but as *sahsip seγ̥on* "similar to wild lilac." Such evidence suggests that the semantic structures of some American Indian vocabularies are based on classes defined less by their boundaries than by their centres.

Kinship terms

Another type of semantic structuring is illustrated by certain systems of kinship terms. In Fox, an Algonkian language, the term for maternal uncle also includes maternal grandmother's sister's son's son (a kind of second cousin). This can be accounted for by recognizing some very simple rules, rules that apply to the other terms of the kinship system as well: (1) siblings of the same sex, as linking relatives, are reckoned as equivalent; (2) a father's sister, as a linking relative, is equivalent to a sister, and conversely, a mother's brother's child is equivalent to a mother's brother. Then a mother's mother's sister's son's son, by rule 1, is equivalent to a mother's mother's son's son; but because one's mother's son is one's brother, this is the same as a mother's brother's son; and this in turn, by the converse of rule 2, is equivalent to a mother's brother. It is clear that the semantic systems of American Indian languages exhibit not only structures of hierarchy and similarity but also rules of semantic equivalence.

Language and culture. The exotic character of American Indian semantic structures, as manifested not only in their vocabularies but also in the relationships expressed by their morphological categories and syntactic patterns, has led a number of scholars to speculate on the relationships between language, culture, and habitual thought patterns or "world view." It was hypothesized that the unique organization of the universe that is embodied in each language might act as a determining factor in the individual's habits of perception and of thought, thus forming and maintaining particular tendencies in the associated nonlinguistic culture. As Edward Sapir put it,

> Human beings do not live in the objective world alone, . . . but are very much at the mercy of the particular language which has become the medium of expression for their society . . . The fact of the matter is that the "real world" is to a large extent unconsciously built up on the language habits of the group . . . We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation.

The Whorfian hypothesis

This idea was further developed, largely on the basis of work with American Indian languages, by Sapir's student Benjamin Lee Whorf, and is now often known as the Whorfian hypothesis. Whorf's initial arguments focussed on the strikingly different organization of experience that can be found between English and Indian ways of saying "the same thing." From such linguistic differences, Whorf infers underlying differences in habits of thought. It then remains to show how these habits are manifested in nonlinguistic cultural behaviour. Thus, Whorf points out that, in Hopi, words referring to units of time (*e.g.,* "day") differ from other nouns in that they have no plural form; furthermore, they cannot be counted with the cardinal numerals ("one," "two," etc.) but only with the ordinals ("first," "second," etc.). From this he infers that when the English speaker speaks of "ten days," as if the days were an aggregate of separate units, the Hopi speaker, on the other hand, thinks in terms of the cyclic recurrence of a single phenomenon. Whorf attempts to support this idea by reference to Hopi ceremonial behaviour, which involves repeated preparation for future events. If, in the Hopi view, each day is really a recurrence, rather than something new, then it is reasonable to believe that the daily repetition of ceremonial acts will have a cumulative effect on the future. As Whorf says, the Hopi belief is diametrically opposed to the English proverb that "Tomorrow is another day."

More investigation is necessary to either prove or disprove the Whorfian hypothesis. In any case, the diversity of American Indian languages and cultures has continued to provide a rich laboratory for investigation. A particularly interesting problem is found in the area of northwestern California, where several small tribes have very similar cultures, but use languages of very diverse types. These are Karok, genetically classified as Hokan; Yurok and Wiyot, which are Algonkian; and Hupa and Tolowa, Athabascan languages. By the Whorfian hypothesis, one might expect that the difference in languages would have produced a greater diversity in the cultures; or failing that, one might expect the languages to have grown more similar to each other. In fact, both linguistic diversity and cultural uniformity seem to have made modest accommodations to each other. As an example of Whorfian linguistic determinism, the systems of biological taxonomy of Yurok and Tolowa, referred to in the previous section, may be noted. The Yurok have a larger number of generic classifications, which means they have more choice in nomenclature, because either a generic or a specific term can be used. This is consistent with the high degree of choice afforded in Yurok grammar, in which word order is nearly free and many morphological categories are optional. The sparser taxonomy of Tolowa offers less choice, corresponding to a much more rigid grammatical structure.

Language, culture, and prehistory

A different kind of relationship between language and culture is of more interest to the student of North American prehistory, namely, the fact that language retains traces of historical changes in culture and so aids in reconstructing the remote past. Here again the pioneering work was done by Sapir, who pointed out, for instance, that the original home from which a group of related languages or dialects has dispersed is more likely to be found in the area of great linguistic diversity; *e.g.,* there are much greater differences in the English dialects of the British Isles than of the more recently settled areas such as North America or Australia. To take an American Indian example, the Athabascan languages are now found in the Southwest (Navajo, Apache), on the Pacific Coast (Tolowa, Hupa), and in the Western Subarctic. The greater diversity of the Subarctic languages leads to the hypothesis that the original centre of Athabascan migration was from that area. This northern origin of the Athabascans was further confirmed in a classic study by Sapir in which he reconstructed parts of prehistoric Athabascan vocabulary, showing, for example, how a word for "horn" had come to mean "spoon" as the ancestors of the Navajo migrated from the far north (where they made spoons of deerhorns) into the Southwest (where they made spoons out of gourds). The correlation of such linguistic findings with the data of archaeology holds great promise for the study of American Indian prehistory.

Writing and texts. Although a writing system was in use among the Mayas of Meso-America at the time of first European contact, none was known in North America. All writing systems that have been used for North American Indian languages have resulted from the stimulus of European writing, or have actually been invented and introduced by whites. Perhaps the most famous system is that invented by Sequoyah, a Cherokee, for his native language. It is not an alphabet but a syllabary, in which each symbol typically stands for a consonant-vowel sequence. The forms of characters were derived in part from the English writing system, but without regard to their English pronunciation. Well suited to the language, the syllabary fostered widespread literacy among the Cherokee until their society was disrupted by government action; its use, however, has never died out, and attempts are now being made to revive it.

Other writing systems, invented by missionaries, teachers, and linguists, have also included syllabaries; *e.g.,* for Cree, Winnebago, and some northern Athabascan languages. Elsewhere, alphabetic scripts have been used, adapted from the Roman alphabet by the use of additional letters and diacritics. White educational policy, however, has generally not encouraged literacy in Indian languages. A rich oral literature of American Indian myths, tales, and song texts has been in part published by linguists and anthropologists, and there is now increasing encouragement for the training of Indians to transcribe their own traditions—*e.g.,* among the Navajo. It is possible that there may yet be a flowering of American Indian literature, not only in spoken but also in written form. (W.O.B.)

Meso-American Indian languages

Meso-American, or Middle American, Indian languages are spoken in an area of the aboriginal New World that includes central and southern Mexico, Guatemala, Belize (British Honduras), El Salvador, parts of Honduras and Nicaragua, and part of northwest Mexico. Though various centres of civilization have flourished in the area, sometimes concurrently, from 1000 BC down to the time of the Spanish conquest of Mexico in 1519, Meso-America as a whole has had a more or less common cultural history for 2,500 years.

Treatments of the languages of Meso-America are customarily organized on the basis of their genetic relationships, and only secondarily on that of geographical distribution. Thus, some languages treated as Meso-American are not in fact spoken in Meso-America proper but form linguistic families with languages that are spoken there. For information about languages of northeast, north central, and northwest Mexico that are not dealt with in this section, see above *North American Indian languages.* For languages of Central America not treated here, see below *South American Indian languages.*

Numbers of languages and speakers

Some 70 Indian languages are spoken today in Meso-America by perhaps 7,500,000 people. When the Spanish conquered Mexico in 1519, there may have been 20,000,000 people in Meso-America. Within 100 years of the conquest, the Indian population had decreased by 80 percent as a result of war, disease, forced labour, and starvation. Since then the Indian population has gone back to a higher level, but several languages—have become extinct. Meso-American languages with the greatest number of speakers in the mid-20th century are:

language	number of speakers	family
Aztec	1,200,000	Uto-Aztecan
Yucatec	600,000	Mayan
Quiché-Tzutujil-Cakchiquei	1,200,000	
Mam	450,000	
Kekchi	375,000	
Mixtec	350,000	Oto-Manguean
Zapotec	400,000	
Otomi	450,000	

The study of the Meso-American languages. During the 16th and 17th centuries, some Dominican and Franciscan missionaries devoted themselves to the study of native languages so that priests could deal in religious matters with monolingual Indians. They wrote grammars following a Latin model, devised orthographies applying values used in Spanish or Latin (occasionally inventing new letters), made dictionaries (usually vocabularies or glossaries), and translated Christian texts (confessionals, sacraments, and sermons) into Indian languages. Except for one heroic figure, the Spanish missionary priest Bernardino de Sahagùn, they neither collected nor fostered the collection of folklore. During this period grammars and dictionaries were written for such languages as Aztec, Zapotec, Mixtec, Tzeltal, Yucatec, Quiché-Tzutujil-Cakchiquel, Chortí, and Northwestern Otomí. These collections of data served the successors of the first missionaries. During the 18th century, the momentum of such work decreased, and,

Table 61: Meso-American Indian Languages

family, branch (or group), language	location	number of speakers
1. Uto-Aztecan (Uto-Nahuan) family 48c*		
Shoshonean (*Yutan, Otegonian*) division† *34c*		
A. *Plateau group 18c*		
1. Mono, N Paiute—Bannock (complex?)		
2. Shoshoni—Gosiute (Goshiute), Comanche (complex?)		
3. Ute-Chemehuevi, S Paiute (complex?)		
B. Tubatulabal		
C. *Southern California branch 24c*		
1. Serrano		
2. Luiseño, Juaneño		
3. *Gabrieleño complex 10c*		
Gabrieleño		
Fernandeño		
4. *Cahuilla complex*‡		
Cahuilla		
Cupeño		
D. Hopi		
Sonoran (*Mexican*) *division 39c*		
E. *Piman group*		
1. *Piman complex 8c*		
Papago (Pima)	Arizona; Sonora	500
Lower Pima (Nevome)	Sonora	900
Tepecano§	Jalisco	a few
2. *Tepehuán complex*‡		
N Tepehuán§	Sonora	6,300
S Tepehuán§	Jalisco	17,700
F. *Yaquian* (*Taracahitian*) *branch 23c*		
1. *Tarahumara complex 7c*		
Tarahumara (Rarámuri)	Chihuahua	36,600
Guarillo		7,200
2. Tubar		extinct
3. *Cáhita complex 15c*		
Eudeve (Heve)		extinct
Ópata, Jova		extinct
Yaqui, Mayo (Cáhita)§	Arizona; Sonora, Sinaloa	26,500
G. *Coran group 15c*		
Cora§		?
Huichol§	Nayarit	10,900
H. *Nahuan group 15c*		
1. *Aztec complex 11c*		
C, N Aztec (Nahuatl)§	State of Mexico, Puebla, Hidalgo	1,200,000
W Aztec (Nahual)§	Michoacán	
E Aztec (Nahuat)§	Veracruz	
Pipil§	C America	2,000
2. Pochutec§	Oaxaca coast	extinct
2. Cuitlatec (Teco) isolate	Guerrero	extinct
3. Seri isolate	Sonora coast	400
4. Tequistlatec complex or group	SE Oaxaca	
Huamelultec	coastal region	5,000
Tequistlatec	mountain region	5,000
5. Jicaque isolate (several dialects or languages)	NW Honduras	300
6. Tlapanec (Subtiaban, Tlapanecan) complex 8c		
A. Tlapanec (Yope)	Guerrero	44,300
B. Subtiaba (Nagrandan) ‖	Nicaragua	extinct?
Maribio ‖	El Salvador	extinct?
7. Oto-Pamean stock 55c		
A. Chichimec (Meco, Jonaz)	Guanajuato	1,000
B. *Pamean group 18c*		
N Pame	San Luis Potosi	3,600
S Pame	Hidalgo	
C. *Matlatzinca complex 10c*	State of Mexico	
Matlatzinca (Pirinda)		2,800
Ocuiltec (Atzingo)		a few
D. *Otomian group 16c*		
1. *Otomi complex 9c*	Hidalgo, Guanajuato, State of Mexico, Querétaro	432,000
NW Otomi		
NE Otomi		
SW Otomi		
Ixtenco Otomi		
2. Mazahua	Michoacán, State of Mexico	221,000
8. Popolocan (Mazatecan) family 25c		
A. *Chochoan group 13c*		
1. Ixcatec	NW Oaxaca (Santa Maria Ixcatian)	200
2. *Chocho complex 8c*		
Popoloc	SE Puebla, NW Oaxaca	34,000
Chocho	NW Oaxaca	2,500
B. *Mazatec complex 10c*		145,500
Mazatec (1)	SE Puebla, N Oaxaca	
Mazatec (2)		
9. Mixtecan family 42c		
A. Amuzgo	E Guerrero, W Oaxaca	20,100
B. *Greater Mixtecan branch 25c*		
1. *Mixtec group 15c*		335,100
Mixtec (1)	E Guerrero,	
Mixtec (2)	S Puebla,	
Mixtec (3)	W Oaxaca	
2. Cuicatec	NE Oaxaca	20,200
C. Trique	W Oaxaca	18,700
10. Zapotecan family 24c	Oaxaca	
A. *Zapotec group 14c*		407,600
Juárez Zapotec	Ixtlán	
Villalta Zapotec	Yatzachi	
S Mountain Zapotec	Cuixtla	
Valley Zapotec	Mitla, Tehuantepec	
B. Soltec	Sola de Vega	extinct
C. Papabuco	Elotepec	extinct
D. Chatino	southwest	27,500

*Indicates centuries of separation. †Not spoken in Meso-America. ‡There is some doubt whether these groups should be given the status of complexes. §Sonoran languages spoken in Meso-America. ‖ Varieties of the same language spoken in different countries (and having different names).

after Mexico became independent in the first part of the 19th century, Spanish clerics were ousted, leaving further work on indigenous languages to travellers and gentlemen scholars—mostly people poorly qualified for such a task.

Extent of the studies of Meso-American languages

Modern linguistic techniques for the description of languages were not applied to Meso-American languages until North Americans turned their attention to the area in the 1930s and 1940s. Since then, much professional linguistic work has been done on these languages, especially those of Mexico. Almost every language of Meso-America has been worked on by at least one linguist, but the time spent and level of linguistic competence of the investigators have varied greatly. For most of the languages, grammatical and lexical data have been collected, much of which remains unpublished. A number of competent grammars and dictionaries have appeared; none of them however, is exhaustive or definitive. Folktales have been collected for a smaller number of languages. Spanish-based orthographies have been devised for most of the Meso-American languages in the 20th century, but not much reading matter is available in them. In short much work remains to be done.

CLASSIFICATION

Modern genetic groupings. The classification of Meso-American Indian languages presented here reflects generally accepted genetic groupings (as of the early 1970s), based on similarities in vocabulary and grammar and on the establishment of regular correspondences between sounds in cognate (related) words among the several languages. The languages grouped together are presumed to have developed from a common ancestor, called a protolanguage. Not all of the languages of Meso-America have been convincingly assigned to a specific group. A few of these languages are currently thought to be unrelated to any of the established genetic groupings and are listed individually in the table; these solitary languages are called isolates.

Within a given genetic grouping, there may be several levels of relatedness. Glottochronology (or lexicostatistics), developed by two United States linguists—Morris Swadesh and Robert Lees—is a controversial and not universally accepted procedure for measuring degrees of difference between related languages in terms of years of separation. Based on the assumption that all languages change more or less to the same degree in a given period of time the method employs a list of 100 items of "basic" or "noncultural" concepts, which are assumed to be expressible by vocabulary items in any language. Over a period of 1,000 years, different words will have been substituted to express 14 percent of the 100 concepts every 1,000 years, two languages that separated 1,000 years ago will share 74

Table 61: Meso-American Indian Languages (continued)

family, branch (or group), language	location	number of speakers
11. Chinantecan group 15c*	N Oaxaca	80,000
Chinantec (1)		
Chinantec (2)		
Chinantec (3)		
Chinantec (4)		
12. Manguean (Chorotegan, Chiapanec-Mangue) group 13c		
A. Chiapanec	Chiapas	extinct
B. Mangue (Dirian, Nagrandan) ‖	Nicaragua	extinct
Chorotega ‖	Honduras	extinct
Nicoya (Orotiña) ‖	Costa Rica	extinct
13. Huave isolate	SE Oaxaca	25,300
14. Mixe-Zoque (Zoquean, Mixean, Zoque-Mixe) family 36c		
A. *Zoquean group 14c*		
Zoque	Tabasco, Chiapas, Oaxaca	37,600
Sierra Popoluca	Veracruz	25,300
Texistepec	Veracruz	3,000
B. *Mixean group 13c*		
1. Sayula	Veracruz	1,000
Oluta	Veracruz	1,000
E, W Mixe	E Oaxaca	77,500
2. Tapachultec	SE Chiapas coast	extinct
15. Totonacan family 2bc		
Totonac	Veracruz, Puebla	239,000
Tepehua	Veracruz, Hidalgo	18,800
16. Mayan family 41c		
A. *Huastec complex 9c*		
Huastec	San Luis Potosi, N Veracruz	101,000
Chicomuceltec (Coxoh)	Chiapas	a few?
B. *Yucatec* (*Maya*) *Complex 10c*		
Yucatec	Yucatán, Campeche, Quintana Roo, N Guatemala, Belize	605,000
Lacandón	Chiapas	200
Itzá	N Guatemala	500
Mopán	N Guatemala, Belize	6,000
C. *Western division 30c*		
1. *Greater Tzeltalan branch 19c*		
a. *Cholan proper group 14c*		
Chontal (Yocotán)	Tabasco	51,000
Chol	Tabasco, Chiapas	109,000
Chorti	Honduras, E Guatemala	64,000
b. *Tzeltalan group 14c*	Chiapas	
Tzotzil (Quelén)		123,000
Tzeltal		123,000
2. *Greater Kanjobalan branch 21c*		
a. *Chujean group 16c*		
Tojolabal (Chaneabal)	Chiapas	19,000
Chuj	NW Guatemala	30,000
b. *Kanjobal proper group 15c*		
i. *Kanjobal complex 7c*	NW Guatemala	
Kanjobal (Conob, Solomec)		62,000
Acatec		13,000
Jacaltec		21,000
ii. *Mochó complex* (Cotoque)	SE Chiapas	
Motozintlec		500
Tuzantec		100
D. *Eastern division 34c*		
1. *Greater Mamean branch 26c*		
a. *Mamean proper group 15c*		
Teco	SE Chiapas, W Guatemala	5,000
Mam	W Guatemala	434,000
b. *Ixilan group 14c*	NW Guatemala	
Aguacatec		15,000
Ixil		60,000
2. *Greater Quichéan branch 26c*		
a. Uspanteco	NW Guatemala	15,000
b. *Quiché complex 10c*	C Guatemala	
Quiché (Achi)		680,000
Sacapultec		3,000
Sipacapa		3,000
Cakchiquel		434,000
Tzutujil		50,000
c. *Pocom complex 10c*	EC Guatemala	
Pocomam		30,000
Pocomchi		75,000
d. Kekchi		374,000
17. Tarasco isolate	SW Michoacán	72,000
18. Xinca complex 10c	SE Guatemala	
Eastern Xinca	Yupiltepeque, Jutiapa	extinct
Northern Xinca	Jumaytepeque	50
Southern Xinca	Chiquimulilla	100
Western Xinca	Guazacapan	100
19. Lencan family 20c		
Lenca	SW Honduras	25
Chilanga	E El Salvador	a few
20. Paya complex 10c	N Honduras	300
21. Misumalpan (Misuluan) family 43c		
A. Mosquito (Miskito)	Nicaragua, Honduras	115,000
B. *Matagalpa complex 10c*		
Matagalpa	Nicaragua, Honduras	100
Cacaopera	El Salvador	?
C. *Sumo complex 11c*		
Sumo, Úlua, Tahuajca	Nicaragua	200

*Indicates centuries of separation. †Not spoken in Meso-America. ‡There is some doubt whether these groups should be given the status of complexes. §Sonoran languages spoken in Meso-America. ‖Varieties of the same language spoken in different countries (and having different names).

percent cognates (86 percent of 86 is 74 percent). The following are terms and categories for degree of relatedness, correlated with glottochronological time depths, that will be used to describe the various Meso-American language groups. The figures given are minimal bounds.

term	centuries of separation	percentage of cognates
dialects	0–5	86–100
language complex	7–11	71–81
language group	13–17	60–68
branch (or family if there is no superordinate category)	19–26	45–56
language family	35–45	26–35
stock or phylum	55–65	14–19

In Table 61 every family (group) and isolate has a separate number from 1 to 21. Each of the 21 headings specifies the name of a grouping, with alternative names. Numbers in parentheses following language names indicate that there are several closely related languages all referred to by the same name. For each language grouping the various levels of relatedness are specified, including glottochronological figures (*c* = centuries), which are Swadesh's, except for Mixe-Zoque, Mayan, and Xincan, which are those of the U.S. linguist Terrence Kaufman. Family and stock names are formed in the following ways: (1) A typical language, usually the most widely spoken, is suffixed with *-an* (*e.g.,* Mixtecan). (2) Two typical names are chosen and compounded (*e.g.,* Mixe-Zoque). (3) Parts of two or more language names are joined, and *-an* is suffixed (*e.g.,* Oto-Manguean, Oto-Pamean, Mis-Uluan/Misumalpan).

Group names end in *-an* if the groups are further subgrouped but do not end in *-an* if they are immediately divided into discrete languages.

Outliers

The map gives the approximate geographical distribution of the 21 language groupings and isolates of Meso-America. None of the extinct undocumented languages is indicated. Except for some outliers, separate languages within a grouping are not localized. An outlier is a language that has been carried into a foreign cultural and linguistic context by migration; *e.g.,* Mangue is a Chiapanec outlier in Misumalpan territory, Subtiaba is a Tlapanec outlier in Misumalpan territory, Pipil is a Nahua (Aztec) outlier in Quichéan, Xinca, Lencan, and Misumalpan territories.

In the following paragraphs the numbers in parentheses refer to groupings in Table 61.

Uto-Aztecan (1). The Uto-Aztecan family consists of some 27 languages that are universally recognized to fall into eight groups or branches—the Plateau group, Tubatulabal, the Southern California branch, Hopi, the Piman group, the Yaquian branch, the Coran group, and the Nahuan group. Tubatulabal and Hopi contain just one language each. The first four groups are commonly, but not universally, recognized as forming a Shoshonean division within the family. None of the Shoshonean languages is spoken in Meso-America, and no distribution or population data is cited for them in Table 61 (see above *North American Indian languages*). There are two common ways of grouping the remaining languages, depending on the position assigned the Nahuan group. Either Nahuan is considered as separate and the rest as forming a Sonoran division, thereby producing three divisions—Shoshonean, Sonoran, and Nahuan—or else Nahuan is included within Sonoran, thereby producing a Shoshonean versus Sonoran dichotomy, which is the arrangement used in this article. Several scholars believe that the "division" concept is faulty here and that Uto-Aztecan contains eight groups and branches that are not to be further grouped in any special way.

Only some Sonoran languages are spoken in Meso-America (indicated by signs [§] in Table 61). The extinct Tubar belongs to the Yaquian branch, but whether to the Tarahumara complex, the Câhita complex, or neither, is not clear. The Nahuan group includes the extinct Pochutec, formerly spoken on the coast of Oaxaca, Mexico, and poorly documented; Pochutec is clearly very divergent from the rest of the group. The Aztec complex is considered by some to be a single language with several dialects. The three Aztec languages were spoken within the Aztec Empire as it was constituted in 1519. Pipil speakers, who also refer to their language as *nawat*, were not a part of the Aztec culture and probably represent a Toltec expansion from several centuries earlier.

Various scholars' work on Uto-Aztecan

In 1859, Johann Karl Buschmann, a German philologist, correctly identified all the then-known Uto-Aztecan languages as forming a family. In 1883 a French philologist, Hyacinthe de Charencey, divided Uto-Aztecan into Oregonian (=Shoshonean) and Mexican (=Sonoran), and, in 1891, in the United States, anthropologist Daniel Brinton recognized Shoshonean and divided the Sonoran division (of this article) into Nahuatlan (=Nahuan) and Sonoran (=the Sonoran of this article minus Nahuan). Brinton's division was followed by the United States biologist John Wesley Powell in his classification of North American languages.

Buschmann in 1859 and United States anthropological linguist Edward Sapir in 1915 contributed to the comparative study of Uto-Aztecan by assembling sizable numbers of cognate sets.

A number of now-acculturated and racially absorbed Indian ethnic groups of northern Mexico are believed by many to have spoken Uto-Aztecan languages, although only the language names are known, and not the languages themselves. These are: Suma, Jumano, Lagunero, Cazcán, Tecuexe, Guachichil, and Zacatec.

Uto-Aztecan is generally accepted by specialists as related to the Kiowa-Tanoan family of North America and with it to form the Aztec-Tanoan stock (or phylum).

Cuitlatec (2). The now extinct Cuitlatec language has not been linked convincingly with any other language or family, though the idea that it might be related to Uto-Aztecan has been entertained.

The Hokan hypothesis (3–5). In 1919 two United States anthropologists, Roland Dixon and Alfred Kroeber, tried to improve on an older North American classification by reducing the multiplicity of language groupings in California (about 50) to a manageable number of families and stocks. Working over a period of several years, they developed the hypothesis that most California languages belong to one of two great groupings (called phyla or superstocks), Hokan and Penutian. The formulation was accepted and extended by others. Hokan included Shasta, Achumawi, Atsugewi, Chimariko, Karok, Yanan, Pomoan, Washo, Esselen, Yuman, Salinan, and Chumashan. By 1891/92 it had been suggested that Yuman, Seri (3), and Tequistlatec (4) were related. In 1915 the matter was re-examined in the light of the Hokan hypothesis, and it was concluded that all of the languages named above are related. Since then most scholars familiar with Yuman languages have believed that Seri and Yuman are related, and many who accept the Hokan hypothesis believe that Seri and Yuman form a special group within Hokan.

Jicaque (5), which is very poorly documented, though still spoken, has plain, aspirated, and glottalized stops (different varieties of consonant sounds), as do many Hokan languages. In 1953 it was suggested that Jicaque is a Hokan language. The general acceptance of the proposition may have been uncritical, because the available data on Jicaque is hardly reliable.

Documented dead languages

Extinct languages of northeast Mexico. All of the several languages once spoken in northeast Mexico and South Texas have become extinct. Documented languages of Mexico are: Coahuilteco, Comecrudo, Cotoname, Naolan, and Tamaulipec (or Maratino). Those of Texas are Karankawa (and Klamkosh), Atakapa, and Tonkawa. John Wesley Powell classified the first three as forming a Coahuiltecan family. The other Mexican languages were unknown until recently. Each of the three Texan languages was considered by Powell to be an isolate. In 1920 Coahuiltecan was redefined to include Karankawa and Tonkawa and to be coordinate with Hokan in a Hokan-Coahuiltecan (=Hokaltecan) superphylum.

Tlapanec (6). The Tlapanec complex was first correctly identified by Walter Lehmann, a German physician, in 1920. In 1925 Edward Sapir tried to establish Subtiaba as a Hokan language, proposing some Proto-Hokan reconstructions that could account for the Subtiaba forms. This classification is generally accepted. More recently, however, Calvin Rensch, a U.S. missionary linguist, tried to validate the Oto-Manguean hypothesis (see below) by means of full-scale phonological reconstruction. He believed Tlapanec to be Oto-Manguean; others considered it to be intermediate between Oto-Manguean and Hokan. It must be kept in mind that most of the specialists who have immersed themselves in the study of large numbers of American Indian languages believe that almost all of them are genetically related to one another. This relationship derives from a period, perhaps 20,000 to 30,000 years ago, when some of the languages were still spoken in Asia. With such a point of view, correct grouping (or degree of relationship) is a more interesting question than genetic relatedness.

Oto-Pamean (7). The Oto-Pamean stock contains four groups and complexes, Chichimec, Pamean, Matlatzinca, and Otomían, of which only the last two are spoken within Meso-America. The exact number of languages within the Otomí complex is not yet determined, though there seem to be four. Oto-Pamean was first correctly identified in 1892.

Popolocan (8). The Popolocan family (which might more appropriately be called Mazatecan) was correctly identified in 1926. The exact number of languages within the Mazatec complex has not yet been determined, though there are at least two.

Problem of classifying Amuzgo

Mixtecan (9). There is some difference of opinion as to how the various languages here included within Mixtecan are to be grouped. The main problem is whether Amuzgo is Mixtecan or a separate branch within Oto-Manguean. It has been included within Mixtecan in some systems and excluded from it in others. There seem to be three languages within the Mixtec group, a subdivision of Mixtecan.

Zapotecan (10). The Zapotecan family was correctly identified by William Mechling in 1912, but only Francisco Belmar, a Mexican philologist, correctly recognized that Papabuco is a separate language, neither Zapotec nor Chatino (in 1905). Belmar, however, incorrectly included Chinantec within Zapotecan. The Chatino language has several dialects. Within the Zapotec complex there are at least four languages, and perhaps more.

Chinantecan (11). The Chinantecan group contains approximately four languages, the exact number as yet undetermined. The separateness of Chinantecan within Oto-Manguean was recognized in 1912.

Manguean (12). The Manguean group was correctly identified by Belmar in 1905. Its members, formerly spoken in Chiapas (Mexico), and in Nicaragua, Honduras, and Costa Rica, are now extinct.

The Oto-Manguean hypothesis (7–12 or 6–13). Ever since 1891, it has been proposed that two or more of the above families (7–12) should be linked. Since about 1925,

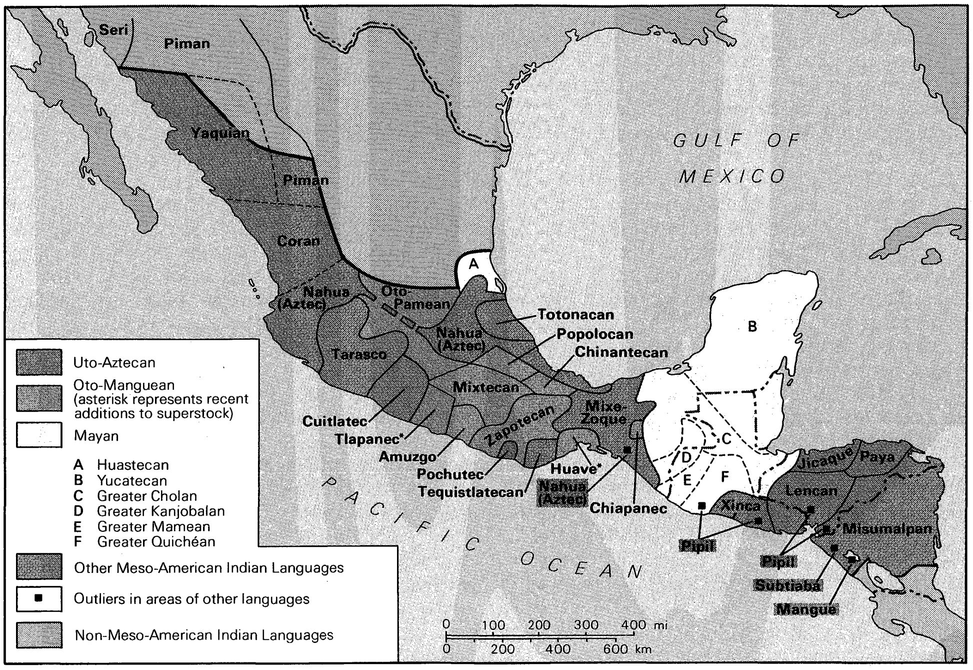


Figure 35: Distribution of Meso-American Indian languages *c.* AD 1500. Boundaries are schematic.

it has been generally accepted by specialists that the Oto-Pamean, Popolocan, Mixtecan, Zapotecan, Chinantecan, and Manguean groups form a larger genetic grouping (phylum), commonly labelled Oto-Manguean. This may be called the "classical Oto-Manguean formulation." Since 1950, work has been going on in the reconstruction of parent languages for each of the constituent families and groups. Since 1961, two revisions have been proposed in the formulation of what constitutes Oto-Manguean: the Tlapanec language complex has been recognized as included in or closely related to Oto-Manguean, and Huave has been proposed as an Oto-Manguean language. In the early 1970s, therefore, most Oto-Manguean specialists considered the grouping to consist of groups 6–13.

Reconstruction of Oto-Manguean phonology

The comparative study of the Oto-Manguean phylum has resulted in the first case in the Western Hemisphere in which the remote common ancestor of several language families has been phonologically reconstructed. Comparative linguistics at the phylum level has been largely unsuccessful with other postulated superstocks because of the relatively small number of cognates that can be identified. Except for Manguean, all Oto-Manguean languages are spoken in central Mexico.

Huave (13). Early proposals linked Huave to Mixe-Zoque and Mayan. Although this has not been generally accepted by many specialists, it has been uncritically repeated in most compilations. Recently, Morris Swadesh presented a reasonably well documented proposal for Huave as an Oto-Manguean language.

Mixe-Zoque (14). The Mixe-Zoque family consists of eight languages, which, comparative phonology and grammar suggest, form two branches—a Zoquean group, and a Mixean group including Tapachultec. Glottochronological figures, however, suggest a three-way division, as shown in the Table. The Mixe-Zoque family was correctly identified by Hyacinthe de Charencey in 1883. The Texistepec, Sayula, and Oluta languages of this family are all locally called Populuca.

Totonacan (15). The Totonacan family contains just two languages, of which one (Totonac) has at least three dialects. Possibly, Totonac is a complex.

Mayan (16). The Mayan family was correctly identified by a German ethnographer, Otto Stoll, in 1884. This family, with 24 languages and nearly 3,500,000 speakers, is the most diversified and populous language family of Meso-America. The Huastec language is separated by more than 1,000 miles from the nearest other Mayan language. Taken with the fact that the Huastecs did not share in the Classic Maya civilization, this requires a historical explanation involving the separation of Huastec from the rest of the family more than 2,500 years ago. Though the geographical extent of the Mayan languages is considerable, the Mayan peoples, languages, and cultures (as contrasted with those of the Aztecs), have never been particularly expansionist.

A number of attempts have been made to classify the Mayan languages, each one availing itself of more data than the last. The classification given here as of 1971 recognizes, at the lowest level, ten groupings. Specialists have disagreed on the precise positions of Tojolabal and Chuj, Motozintlec, Aguacatec, Uspantec, and Kekchí and have held no firm opinions about the Yucatec or Huastec complexes. Not much comparative work on the Mayan family has seen print, but much data has recently been collected. The main contributors to Mayan comparative studies have been the U.S. linguists Norman McQuown (1950s and 1960s) and Terrence Kaufman (1960s).

Hypotheses concerning superstocks

The Macro-Mayan and Macro-Penutian hypotheses. In 1931 L.S. Freeland, a U.S. anthropological linguist, tried to show that Mixe (Zoque) is related to the "Penutian" languages, a superstock that up until then had been limited to California, Oregon, Washington, and British Columbia. In 1935 it was suggested that the similarities between Uto-Aztecan, Tanoan, Kiowa, Penutian, Mixe-Zoque, and Mayan were such as to indicate the existence of a superstock, which it was proposed to call Macro-Penutian. This hypothesis had favour for a period but was never demonstrated nor taken very seriously by specialists. Since then the first three have been generally joined in Aztec-Tanoan. In 1942 it was suggested that Mixe-Zoque and Totonacan might be related genetically to each other and the two in turn might be related to Mayan, the resultant superstock to be called Macro-Mayan. Recently it has been claimed that Tarasco (17) probably belongs in Macro-Mayan as well, though the attempt to prove this has not been convincing to most Mayanists, to whom, minus Tarasco, the Macro-Mayan hypothesis seems as reasonable as the Hokan hypothesis.

Tarasco (17). Tarasco has been linked genetically by some not only to Marco-Mayan but also to both Zuni (in North America) and Quechua (in South America), but without general scholarly acceptance.

Xinca and Lencan (18–19). It has been suggested that Xinca and Lencan are related and that one or both of them is related to Mayan (16), Chibchan (in South America), or Uto-Aztecan (1). None of these hypotheses has been demonstrated as probable.

Languages outside Meso-America proper. The Paya language (20) and the Misumalpan family (21) are Central American languages spoken outside of the cultural area of Meso-America proper, though they have Meso-American outliers in their territory. Paya (20) has been linked in hypotheses to Chibchan and Cariban (both in South America), and perhaps to others, but not convincingly. The Misumalpan family (21) has been recognized since 1895. Since that date some scholars have believed that the three languages and complexes listed are coordinate, and others have believed that the first two constitute one group and the other consitutes a second group. Although the family relationship can be verified on inspection, no supporting comparative work has been published. Previous comprehensive classifications of the Meso-American Indian languages were presented by the U.S. anthropologists Cyrus Thomas and John R. Swanton in 1911 in *Indian Languages of Mexico and Central America and Their Geographical Distribution,* by Edward Sapir in the 14th edition of *Encyclopædia Britannica* (1929), and by Morris Swadesh in 1967 in *Handbook of Middle American Indians.*

Newly discovered languages and reconstructions. Although there are probably no uncharted areas in Meso-America, it is not necessarily the case that all the Indian languages of Meso-America have been correctly identified, and there are probably some multilingual Indian communities as well that are not known to be such. In 1967 Terrence Kaufman discovered a hitherto undocumented Mayan language spoken by several hundred Indians in four or five towns in southeast Chiapas and west central Guatemala. Although it appears to be closely related to Mam, Kaufman considered it a separate language and christened it Teco. Kaufman identified two more new Mayan languages in the course of a linguistic survey of Guatemala. These two new languages—Sacapultec (formerly considered Quiché) and Sipacapa (formerly assumed to be Mam)—are not documented in print and both belong to the Quiché complex.

Reconstruction of earlier forms of the Meso-American Indian languages has focussed primarily on phonology and vocabulary. Phonological and lexical comparative studies as well as reconstruction have been done for the following groups: Uto-Aztecan; Oto-Manguean—Oto-Pamean, Popolocan, Mixtecan, Zapotecan, Chinantecan, Manguean; Mixe-Zoque; and Mayan (in part). A small amount of grammatical comparison has been done within Oto-Manguean and Mixe-Zoque. In addition, some studies have been done of reconstructed vocabulary for the purpose of hypothesizing about the culture of the speakers of the protolanguages.

RELATION OF LANGUAGES TO HISTORICAL AND CULTURAL INFLUENCES

Pre-Columbian diffusion. The following are some of the important civilizations that have flourished in Meso-America:

civilization	period	location
Olmec	1200 BC–400 BC	Gulf Coast, Mexico
Monte Albán	400 BC–AD 700	Oaxaca, Mexico
Teotihuacán	AD 100–600	Central Mexico
Classic Maya	AD 300–900	Chiapas, Mexico; Petén, Guatemala
Toltec	900–1200	Central Mexico
Aztec	1300–1500	Central Mexico

The Aztecs spoke Nahuatl, as did the Toltecs. The Classic Maya probably spoke two or three Mayan languages, and the people of Monte Albán probably spoke one or more Zapotecan languages. No one knows what either the Teotihuacán people or the Olmecs spoke, but it has been surmised that at least some Olmecs spoke Mixe-Zoque languages and that the Teotihuacán people may have spoken Otomían languages (though an Aztec tradition says Totonac).

In the pre-Columbian period, there was naturally contact among Meso-American languages and occasional borrowing of vocabulary and other linguistic features. Partly because of the unavailability of grammars and dictionaries, actual cases of such diffusion have not been much studied.

Known contacts between various Indian groups

Some of the known contacts resulting in borrowing are the following: (1) Mixe-Zoque languages (Olmecs?) have given words to Mayan, Mixtecan, Zapotecan, Otomían, Aztec, Lencan, Xinca, and Jicaque; (2) Zapotecan languages (Monte Albán) have given words to Huastec and Yucatec; (3) Mayan languages (Mayas) have given words to Xinca, Lencan, and Jicaque; and (4) Nahuatl (Toltecs and Aztecs) has given words to Mayan, Lencan, other Uto-Aztecan languages, as well as to other Meso-American languages. Words diffused from these sources provide evidence that contact took place. Scholars know that contact must have taken place at particular times and places, and therefore can form hypotheses about where certain languages may have been spoken in the more remote past.

External relationships and contacts. Various scholars have suggested that some Meso-American language or family is related to a language or family (other than Uto-Aztecan) outside of Meso-America. These suggestions are mostly parts of larger attempts to synthesize the language classification of the New World, or of the whole world, and are usually based on the sometimes unexpressed view that all the languages of the Western Hemisphere or even of the whole world are ultimately genetically related. Although the assumption may be true, the proposed connections have been unconvincing to specialists in Meso-American languages. The only generally accepted larger groupings are Hokan and Penutian. Most scholars do not have the breadth of knowledge to be able to evaluate these vast proposals.

Prehistoric linguistic connection between Meso-America and the Andes

One proposal of external relationship probably has some merit. In 1961 it was suggested that Chipaya—a language spoken on the shores of Lake Titicaca in Bolivia—is genetically related to the Mayan languages. The hypothesis, proposed by Ronald Olson, a U.S. missionary linguist, was based on 120 sets of lexical comparisons between Chipaya and Proto-Mayan. The data cited are subject to more than one interpretation, because many of the comparisons involve semantic notions and word forms that are widespread in the Western Hemisphere; also, Chipaya has been so influenced grammatically by Aymara (which all Chipayas can speak) that any grammatical peculiarities it may once have shared with Mayan have disappeared. Because a core of data showing regular sound correspondences remains, it is probably necessary to assume that there is a historical connection between Chipaya and Mayan, possibly, but not demonstrably, a genetic relationship. The connection may have been direct—presumably from Meso-America to Bolivia via land—or there may be other languages in western South America that show prehistoric contacts with Mayan. The acceptance of a prehistoric linguistic connection, neither extremely remote nor extremely recent, between Meso-America and the Andes is quite provocative, inasmuch as other evidence exists for early culture contact between Meso-America and the Andes, Meso-America generally being the donor and the Andes generally being the beneficiary; *e.g.,* in the case of corn. Later diffusion from South America to Meso-America also occurred; *e.g.,* witness the transference of peanuts, metallurgy, hammocks.

Interaction between Spanish and Indian languages. In modern Meso-America, the dominant European language is Spanish. The speakers of all Meso-American Indian languages include some who are bilingual; and a few languages are spoken by almost totally bilingual populations. Most Indian languages spoken by sizable populations have at least 50 percent monolingual speakers. All Meso-American languages with a significant number of bilingual speakers have been influenced by Spanish, primarily in the areas of vocabulary, particles, and word order. Since the Spanish conquest, Meso-American languages have been borrowing words from Spanish, and, because the kind of Spanish spoken has changed somewhat over the years, both in vocabulary and pronunciation, different historical periods are usually distinguishable in lexical borrowings. For a variety of reasons, certain function words, primarily conjunctions and adverbs, are frequently borrowed from Spanish; *e.g., ya* "already," *pero* "but," *hasta* "until," *y* "and," *o* "or," *ni* "not even," *hasta* "even," *si* "if," *cuando* "when," *porque* "because," *por eso* "therefore, so," *entonces* "then." Some languages have assimilated the Spanish word order of subject–verb–object.

Conversely, the Spanish of Meso-America has been the recipient of vast amounts of lexical material from local languages, primarily Nahuatl. The borrowing has provided names of plants, animals, artifacts, and social forms indigenous to Meso-America and lacking names in Spanish. Among the reasons that Nahuatl has been the primary source is that the Aztecs were the first Meso-American people conquered by the Spaniards; the Aztecs had outposts in many parts of Meso-America; the Spaniards recruited Aztecs, particularly as guides, into their military force to assist their venture of subduing the rest of Meso-America; and, for several decades, Aztec, written in Roman orthography, was used in many parts of Meso-America to keep official records, such as deeds, wills, and censuses.

Many of the words borrowed into Spanish from Aztec have since passed in turn into English; *e.g.,* chili, chile, or chilli (Spanish *chile*), avocado (Spanish *aguacate*), chicle, chocolate, peyote, coyote, tomato (Spanish *tomate*), ocelot (Spanish *ocelote*), guacamole, mescal.

Bilingualism among Indians

In some parts of Meso-America, because of economic and social conditions, an Indian may speak one or more Indian languages besides his own. This is common in Guatemala, where some areas have been recently colonized by speakers of more than one language, or some communities have received outside settlers in the more remote past.

The names used in this article for the Meso-American Indian languages are English versions of the Spanish terms for them. Only in a few cases are these names the ones actually used by the people who speak the languages in question. First, most of the names are of Aztec origin, because at the outset the Spanish learned of local phenomena primarily via Aztec. Secondly, some languages have no special name of their own, simply being called "our language."

Pre-Columbian writing. Most of the Meso-American cultures shared a mathematical notation and calendrical system that had been developed and diffused in the distant past, probably before 500 BC. At the time of European contact the Aztecs, Zapotecs, Mixtecs, Otomís, Mayans, and perhaps some others were all producing records on stone (inscriptions) and on a type of homegrown paper (produced from the amate tree, *Ficus glabrata*), these latter being commonly called codices. Except for the Mayan system, which probably originated before AD 1, the records cannot properly be called writing, in that it was not possible to represent all of speech, but only numbers, dates, and names (pictographically). The Mayan system, besides representing all these, was also used to represent morphemes (words and word elements) and phonemes (distinctive sounds). Presumably the symbols used in this system (called glyphs) represent individual phonemes, syllables, and morphemes; and they give semantic information as well to take the ambiguity out of homophonous readings. Several scholars have devoted much time to the study of Mayan writing, but, to date, the results have not been very impressive. A few scholars outside the Meso-American field believe the Mayan writing system is purely ideographic and hence inherently undecipherable without a bilingual inscription or text in a known language. All specialists within the Mayan field hold that the Mayan is a mixed ideographic and phonological system.

What may be delaying progress in the deciphering of Mayan writing is the absence of reconstructions for intermediate groupings within the Mayan family (*e.g.,* Proto-Yucatecan, Proto-Cholan, and others) and ignorance of Mayan languages other than colonial Yucatec on the part of the investigators. Efforts are being made to correct these deficiencies, particularly by Mexican specialists. It is not known whether Mayan writing was used to write more than one language and, if so, what the languages were. If only one, it was probably either Proto-Cholan or Proto-Yucatecan. The symbols used in all the pre-Columbian notation systems are obviously pictographic in origin, as was the case in the ancient Egyptian, Sumerian, ancient Chinese, and Indus Valley writing systems.

LINGUISTIC CHARACTERISTICS

Diverse types of sound systems and grammars

In general, all the languages of a particular family are typologically similar to one another both in phonology and grammar. Among the 21 language groupings in Meso-America, there are several types of sound systems and grammatical systems. Because study in this area has hardly begun, nothing very secure can be asserted here, but some general characteristics can be outlined on the basis of data for the following reasonably well-documented languages: Tequistlatec, Otomí, Mazatec, Mixtec, Zapotec, Chinantec, Aztec, Zoque, Totonac, Quiché, and Tarasco.

Phonologically, there is a wide diversity among Meso-American languages. Voiced spirants—*i.e.,* sounds like English *v, z,* or *th* in "then"—are missing from all Meso-American languages. Other phonological features in these languages include a voiceless lateral spirant sound, *lh* (in Tequistlatec and Totonac); a lateral affricate, *tl* (in Aztec and Totonac); a postvelar stop, *q,* in contrast with a velar stop, *k* (in Quiché and Totonac); glottalized vowels (in Zapotec, Zoque, Aztec, and Totonac); glottalized consonants (in Tequistlatec, Quiché, Otomí, and Mazatec); aspirated stops (in Tarasco, Otomí, and Mazatec); voiced stops (in Tequistlatec, Otomí, Mazatec, and Chinantec); prenasalized stops (in Otomí, Mazatec, and Mixtec); nasalized vowels (in Otomí, Mazatec, Mixtec, and Chinantec); a labiovelar stop, *kw,* sometimes contrasting with a bilabial stop, *p* (in Otomí, Mazatec, Mixtec, Aztec); tone and stress accent (tone in Otomí, Mazatec, Mixtec, Chinantec, Zapotec; stress in Tarasco and Tequistlatec); and initial and final consonant clusters (in Tequistlatec).

Grammatically, Meso-American languages are rather diverse, but, according to available data, they fall into three main types: Type A, an Oto-Manguean type, is rightward expanding (*i.e.,* modifiers follow the elements they modify) and synthetic to a low degree (*i.e.,* characterized by relatively few morphemes per word). It employs prefixes and prepositions, and it seldom uses compounding to form words. Type B, an intermediate type, is prepositional, like A, and averagely synthetic, making some use of prefixes (subjects, objects, and possessors) and much use of suffixes. It is mildly leftward expanding (*i.e.,* modifiers precede the elements they modify) and is mainly represented by Mayan and Uto-Aztecan languages but partially by Mixe-Zoque and Totonacan. Type C, a leftward expanding type, is highly synthetic with great use of suffixes and postpositions and active ablaut (an interchange among consonants and vowels for the purpose of derivation or inflection). It is represented by Tarasco and, partially, by Totonacan and Mixe-Zoque.

Grammatical generalizations

There are a number of grammatical generalizations that can be made about all, or most, Meso-American Indian languages. (1) The genitive relationship between nouns or noun phrases is (except for Tarasco) expressed by means of a possessive pronoun with the possessed noun; *e.g.,* "the dog's fleas" is expressed as "his fleas the dog." (2) Locative notions, such as "above," "below," "in," "on," "beside," are not expressed by prepositions and adverbs, as in European languages, but by means of location nouns (meaning "aboveness," "belowness," "belly," "surface," "side," and so forth), which are always combined with a possessive pronoun, the function of which is to indicate the "object" of the prepositional–adverbial notion. Most languages, however, have at least one generic relational particle that is combined in a phrase with a location noun and its object and has "generic prepositional" function; thus "on the table" is expressed "at (generic particle) its-top the table," or "in the box" is expressed "at its-inside the box." Whereas in most languages the generic relational particles are prepositions, Zoque and Tarasco have postpositions, which are in part related to location nouns.

(3) Within the verbal system, aspect (type of action—*e.g.,* ongoing, habitual, finished, potential, and so forth) is well developed, and tense (time—*e.g.,* now, in the past, in the future) is generally weakly developed. (4) The copula, or equational verb "be," is not expressed in most Meso-American languages. (5) Case suffixes are generally absent,

being present in just three languages: Tarasco has a genitive case, an objective case, and various locational cases; Aztec and Zoque have only locational cases, and these are usually related to location nouns. (6) A relative clause that modifies a noun follows it in all the languages of the sample above; *e.g.,* "the man whom I saw (on the street yesterday)." (7) Some Oto-Manguean languages and some Mayan languages distinguish an inclusive pronoun "we" ("I and you") from an exclusive "we" ("I and he/they").

(8) Gender, or inflectional agreement of other word classes in the noun phrase with the noun itself, is rare in Meso-American languages and is limited to some Oto-Manguean languages. (9) Noun subclassification in the context of possession is not uncommon. In some languages, some nouns undergo form changes when possessed; these languages, therefore, have at least two classes of nouns. In other languages, the possessive pronouns differ in form according to how they are associated with different classes of nouns. In languages in which the semantic motivation for such a subdivision is clear, the main kind of distinction is between intimate possession (body parts, kinship terms, articles of clothing) and casual possession (domestic animals, tools). (10) Some languages (Mayan, Mixe-Zoque) distinguish between the subject (actor) of a transitive verb and that of an intransitive verb by the form of the associated affixed pronoun. (11) Most Meso-American languages average more than one morpheme per word, and Tarasco and Totonac average more than two morphemes per word. (12) Most Meso-American languages (except Aztec) have consonantal or vocalic ablaut, or else show in their vocabulary sets of words that seem to be related through a formerly functional ablaut system.

(13) The numeral systems are vigesimal–decimal; that is, counting is from 1 to 10, then from 11 to 20, then from 21 to 40 (adding 1–20 to 20), then from 41–60 (adding 1–20 to 40), and so on, with special terms for 400 (20 × 20), 8,000 (20 × 20 × 20), 160,000 (20 × 20 × 20 × 20), and so on. In most languages (except Mayan) the numeral expressions for 6 through 9 (sometimes 5 through 9) are compounds of 5 + 1, 5 + 2, 5 + 3, 5 + 4, or the like. (14) In all the languages referred to here, a numeral precedes the noun it quantifies. (Te.K.)

South American Indian languages

South American Indian languages once covered and today still partially cover all of South America, the Antilles, and Central America to the south of a line from the Gulf of Honduras to the Nicoya Peninsula in Costa Rica. Estimates of the number of speakers in that area in pre-Columbian times vary from 10,000,000 to 20,000,000. In the early 1980s there were approximately 15,900,000, more than three-fourths of them in the central Andean areas. Language lists include around 1,500 languages, and figures over 2,000 have been suggested. For the most part, the larger estimate refers to tribal units whose linguistic differentiation cannot be determined. Because of extinct tribes with unrecorded languages, the number of languages formerly spoken is impossible to assess. Only between 550 and 600 languages (about 120 now extinct) are attested by linguistic materials. Fragmentary knowledge hinders the distinction between language and dialect and thus renders the number of languages indeterminate.

Number of languages and speakers

Because the South American Indians originally came from North America, the problem of their linguistic origin involves tracing genetic affiliations with North American groups. To date only Uru-Chipaya, a language in Bolivia, is surely relatable to a Macro-Mayan phylum of North and Meso-America. Hypotheses about the probable centre of dispersion of language groups within South America have been advanced for stocks like Arawakan and Tupian, based on the principle (considered questionable by some) that the area in which there is the greatest variety of dialects and languages was probably the centre from which the language groups dispersed at one time; but the regions in question seem to be refugee regions, to which certain speakers fled, rather than dispersion centres.

South America is one of the most linguistically differentiated areas of the world. Various scholars hold the plausible view that all American Indian languages are ultimately related. The great diversification in South America, in comparison with the situation of North America, can be attributed to the greater period of time that has elapsed since the South American groups lost contact among themselves. The narrow bridge that allows access to South America (*i.e.,* the Isthmus of Panama) acted as a filter so that many intermediate links disappeared and many groups entered the southern part of the continent already linguistically differentiated.

Investigation and scholarship. The first grammar of a South American Indian language (Quechua) appeared in 1560. Missionaries displayed intense activity in writing grammars, dictionaries, and catechisms during the 17th century and the first half of the 18th. Data were also provided by chronicles and official reports. Information for this period was summarized in Lorenzo Hervás y Panduro's *Idea dell' universo* (1778–87) and in Johann Christoph Adelung and Johann Severin Vater's *Mithridates* (1806–17). Subsequently, most firsthand information was gathered by ethnographers in the first quarter of the 20th century. In spite of the magnitude and fundamental character of the numerous contributions of this period, their technical quality was below the level of work in other parts of the world. Since 1940 there has been a marked increase in the recording and historical study of languages, carried out chiefly by missionaries with linguistic training, but there are still many gaps in knowledge at the basic descriptive level, and few languages have been thoroughly described. Thus, classificatory as well as historical, areal, and typological research has been hindered. Descriptive study is made difficult by a shortage of linguists, the rapid extinction of languages, and the remote location of those tongues needing urgent study. Interest in these languages is justified in that their study yields basic cultural information on the area, in addition to linguistic data, and aids in obtaining historical and prehistorical knowledge. The South American Indian languages are also worth studying as a means of integrating the groups that speak them into national life.

Research on the languages

Classification of the South American Indian languages. Although classifications based on geographical criteria or on common cultural areas or types have been made, these are not really linguistic methods. There is usually a congruence between a language, territorial continuity, and culture, but this correlation becomes more and more random at the level of the linguistic family and beyond. Certain language families are broadly coincident with large culture areas—*e.g.,* Cariban and Tupian with the tropical forest area—but the correlation becomes imperfect with more precise cultural divisions—*e.g.,* there are Tupian languages like Guayakí and Sirionó whose speakers belong to a very different culture type. Conversely, a single culture area like the eastern flank of the Andes (the Montaña region) includes several unrelated language families. There is also a correlation between isolated languages, or small families, and marginal regions, but Quechumaran (Kechumaran), for instance, not a big family by its internal composition, occupies the most prominent place culturally.

Most of the classification in South America has been based on inspection of vocabularies and on structural similarities. Although the determination of genetic relationship depends basically on coincidences that cannot be accounted for by chance or borrowing, no clear criteria have been applied in most cases. As for subgroupings within each genetic group, determined by dialect study, the comparative method, or glottochronology (also called lexicostatistics, a method for estimating the approximate date when two or more languages separated from a common parent language, using statistics to compare similarities and differences in vocabulary), very little work has been done. Consequently, the difference between a dialect and language on the one hand, and a family (composed of languages) and stock (composed of families or of very differentiated languages) on the other, can be determined only approximately at present. Even genetic groupings recognized long ago (Arawakan or Macro-Chibchan) are probably more differentiated internally than others that have been questioned or that have passed undetected.

Extinct languages present special problems because of poor, unverifiable recording, often requiring philological interpretation. For some there is no linguistic material whatsoever; if references to them seem reliable and unequivocal, an investigator can only hope to establish their identity as distinct languages, unintelligible to neighbouring groups. The label "unclassified," sometimes applied to these languages, is misleading: they are unclassifiable languages.

Problem of language names

Great anarchy reigns in the names of languages and language families; in part, this reflects different orthographic conventions of European languages, but it also results from the lack of standardized nomenclature. Different authors choose different component languages to name a given family or make a different choice in the various names designating the same language or dialect. This multiplicity originates in designations bestowed by Europeans because of certain characteristics of the group (*e.g.,* Coroado, Portuguese "tonsured" or "crowned"), in names given to a group by other Indian groups (*e.g.,* Puelche, "people from the east," given by Araucanians to various groups in Argentina), and in self-designations of groups (*e.g.,* Carib, which, as usual, means "people" and is not the name of the language). Particularly confusing are generic Indian terms like Tapuya, a Tupí word meaning enemy, or Chuncho, an Andean designation for many groups on the eastern slopes; terms like these explain why different languages have the same name. In general (but not always), language names ending in *-an* indicate a family or grouping larger than an individual language; *e.g.,* Guahiboan (Guahiban) is a family that includes the Guahibo language, and Tupian subsumes Tupí-Guaraní.

There have been many linguistic classifications for this area. The first general and well-grounded one was that by U.S. anthropologist Daniel Brinton (1891), based on grammatical criteria and a restricted word list, in which about 73 families are recognized. In 1913 Alexander Chamberlain, an anthropologist, published a new classification in the United States, which remained standard for several years, with no discussion as to its basis. The classification (1924) of the French anthropologist and ethnologist Paul Rivet, which was supported by his numerous previous detailed studies and contained a wealth of information, superseded all previous classifications. It included 77 families and was based on similarity of vocabulary items. Čestmír Loukotka, a Czech language specialist, contributed two classifications (1935, 1944) on the same lines as Rivet but with an increased number of families (94 and 114, respectively), the larger number resulting from newly discovered languages and from Loukotka's splitting of several of Rivet's families. Loukotka used a diagnostic list of 45 words and distinguished "mixed" languages (those having one-fifth of the items from another family) and "pure" languages (those that might have "intrusions" or "traces" from another family but totalling fewer than one-fifth of the items, if any). Rivet and Loukotka contributed jointly another classification (1952) listing 108 language families that was based chiefly upon Loukotka's 1944 classification. Important work on a regional scale has also been done, and critical and summarizing surveys have appeared.

Current classifications

Current classifications are by Loukotka (1968); a U.S. linguist, Joseph Greenberg (1956); and another U.S. linguist, Morris Swadesh (1964). That of Loukotka, based fundamentally on the same principles as his previous classifications, and recognizing 117 families, is, in spite of its unsophisticated method, fundamental for the information it contains. Those of Greenberg and Swadesh, both based upon restricted comparison of vocabulary items but according to much more refined criteria, agree in considering all languages ultimately related and in having four major groups, but they differ greatly in major and minor groupings. Greenberg used short lexical lists, and no evidence has been published in support of his classification. He divided the four major groups into 13 and these, in turn, into 21 subgroups. Swadesh based his classification upon lists of 100 basic vocabulary items and made groupings according to his glottochronological theory (see above). His four groups (interrelated among themselves and with groups in North America) are subdivided into 62 subgroups, thus, in fact, coming closer to more conservative classifications. The major groups of these two classifications are not comparable to those recognized for North America, because they are on a more remote level of relationship. In most cases the lowest components are stocks or even more distantly related groups. It is certain that far more embracing groups than those accepted by Loukotka can be recognized—and in some cases this has already been done—and that Greenberg's and Swadesh's classifications point to many likely relationships; but they seem to share a basic defect, namely, that the degree of relationship within each group is very disparate, not providing a true taxonomy and not giving in each case the most closely related groups. On the other hand, their approach is more appropriate to the situation in South America than a method that would restrict relationships to a level that can be handled by the comparative method.

At present, a true classification of South American languages is not feasible, even at the family level, because, as noted above, neither the levels of dialect and language nor of family and stock have been surely determined. Beyond that level, it can only be indicated that a definite or possible relationship exists. In the accompanying chart—beyond the language level—recognized groups are therefore at various and undetermined levels of relationship. Possible further relationships are cross-referenced. Of the 82 groups included, almost half are isolated languages, 25 are extinct, and at least 10 more are on the verge of extinction. The most important groups are Macro-Chibchan, Arawakan, Cariban, Tupian, Macro-Ge, Quechumaran, Tucanoan, and Macro-Pano-Tacanan.

Macro-Chibchan. Macro-Chibchan languages, which form the linguistic bridge between South and Central America, are spoken from Nicaragua to Ecuador. Spread compactly in Central America and in western Colombia and Ecuador, they include approximately 40 languages spoken by more than 400,000 speakers. The group is probably more differentiated than a stock, languages not belonging to Chibchan being strongly differentiated. In the Colombian Andes a now extinct Chibchan language was the language of the highly developed Muisca culture. Important present-day languages include Guaymí (about 20,000 speakers) and Move (about 15,000) in Panama, Cuna (600) and Páez (37,000) in Colombia, and Cayapa, or Colorado (4,000), in Ecuador. A connection with Cariban has been suggested, and it is possible that such a relationship could be found through Warao (Warrau) and Waican (Waikan) on the one hand and through Chocó (Cariban) on the other.

Extent of the Arawakan languages

Arawakan. Arawakan languages formerly extended from the peninsula of Florida in North America to the present-day Paraguay–Argentina border, and from the foothills of the Andes eastward to the Atlantic Ocean. More than 55 languages are attested, many still spoken. Around 40 groups still speak Arawakan languages in Brazil, and others are found in Peru, Colombia, Venezuela, Guyana, French Guiana, and Surinam. Taino predominated in the Antilles and was the first language to be encountered by Europeans; although it rapidly became extinct, it left many borrowings. As did most languages of the tropical forest, the Arawakan languages receded with the influx of Spanish and Portuguese, mainly through group extinction; thus, 14 groups became extinct in Brazil between 1900 and 1957. Important languages still spoken are Goajiro (52,000 speakers) in Colombia, Campa (41,000) and Machiguenga (11,000) in Peru, and Mojo (more than 15,000) and Bauré (4,500) in Bolivia. Although most Arawakan languages have been recognized as such for a long time, they are greatly differentiated. They are most probably related to both the Macro-Pano-Tacanan and Macro-Mayan language groups.

Cariban. Cariban languages, numbering approximately 50, were spoken chiefly north of the Amazon but had outposts as far as the Mato Grosso in Brazil. The group has undergone drastic decline, and only about 22,000 people speak Cariban languages today, mostly in Venezuela and Colombia; they have disappeared from the Antilles and have been much reduced in Brazil and the Guianas. The

Table 62: South American Indian Language Groups*

language	location
1. Alacalufan (47): Aksanas or Kaueskar, Alacaluf, Caucau or Caucawe	Chile
2. Andoque (9)	Colombia
3. Araucanian or Mapuche (32, 37)	Chile, Argentina
4. Arawakan (43, 44)	
A. Amuesha	Peru
B. Apolista or Lapachu†	Bolivia
C. Arauan: Araua†, Curina, Madihá, Paumarí, Yamadí	Brazil
D. Chamicuro	Peru
E. Maipurean	
1. Achagua, Amarizana, Capite Minanei, Cauyarí, Guarú, Guayupé, Maipure†, Piapoco, Resígaro, Tariana, Warakena†, Yucuna	Colombia
Anauya, Baré, Curipaco, Guinau, Mandawaca, Parau-jano [Parauhano]	Venezuela
Araikú†, Aruan†, Cariay†, Carútana, Catapolítani, Cawishana†, Hohodene, Manao†, Mapanai, Marawá†, Mariate†, Maulieni, Moriwene, Pasé†, Siusí, Wainumá†, Wiriná, Yabaana, Yumana	Brazil
Arawak or Lokono	Guyana, Fr. Guiana
Goajiro	Colombia, Venezuela
Adzaneni, Ipeca	Brazil, Colombia
Island Carib	Dominica
2. Atorai, Mapidian	Guyana
Wapishana	Guyana, Brazil
3. Baníva, Yavitero	Venezuela
4. Bauré, Mojo or Ignaciano, Muchojeone, Pauna, Paicone	Bolivia
5. Campa, Machiguenga, Piro	Peru
Canamari, Chontaquiro [Chontakiro], Cuniba†, Cushineri†, Ipuriná	Brazil
Inapari	Bolivia
6. Caripuna, Marawan	Brazil
7. Chané†	Argentina
Guaná	Paraguay, Brazil
Quiniquiano, Tereno	Brazil
8. Paressí	Brazil
Sarave	Bolivia
F. Taino†	Antilles
G. Morique [Morike] or Mayoruna	Peru
5. Atacama or Cunza or Lincan Antai†	Argentina, Chile
6. Auake or Arutani	Brazil, Venezuela
7. Auishiri [Awishira]	Peru
8. Baenan	Brazil
9. Bora-Huitotoan (2)	
A. Boran: Bora or Miraña, Emejeite, Muinane	Colombia
B. Huitotoan [Witotoan]:	
Andoquero, Huitoto [Witoto]	Colombia
Coeruna†	Brazil
Ocaina, Orejone†, Nonuya or Achote [Achiote]	Peru
10. Canichana	Bolivia
11. Capixana or Canoe	Brazil
12. Cariban (70)	
A. 1. Acawai (Arecuna, Camaracoto, Ingarico) Waica	Venezuela
Taulipang or Ipuricoto or Pemon	Brazil, Venezuela
2. Apalai, Aracajú, Upurui	Brazil
Oyana	Suriname
Rucuyen	Fr. Guiana
3. Apiacá or Apingi†, Arara†, Parirí†	Brazil
4. Atroahi (Yauaperi), Quirixana, Waimiri [Waimiry]	Brazil
5. Bakairi† [Bacairi], Nahukua (Nahucua), Yaruma†	Brazil
6. Bonari†, Hishkariana, Parucoto, Waiboi, Waiwai	Brazil
7. Cachuene or Caxuiana, Mutuan†, Pauxí†, Saluma, Wayewé	Brazil
Chiquena [Chikene] or Shikiana	Brazil, Guyana
8. Carare, Opone	Colombia
9. Caribe or Calina or Galibi	Antilles, Guianas
Carif	Belize, Honduras
10. Carijona (Guaque [Guake], Umaua)	Colombia
11. Colima†, Muzo, Pijiao	Colombia
12. Cumanagoto† (Chayma†, Tamanaco), Tivericoto†	Venezuela
13. Keseruna, Macushí, Paraviyana†, Purukoto†, Zapara	Brazil
14. Mariquitare or Decuana	Brazil, Venezuela
Yecuana or Mayongong (Cunuana, Ihuruana)	Venezuela
15. Mapoyo or Nepoyo, Yauarana	Venezuela
16. Motilon	Colombia, Venezuela
17. Palmela† [Palmella]	Brazil
18. Panare	Venezuela
19. Patagon	Peru
20. Pawishana	Brazil
21. Pianocoto	Brazil
Tliometesen	Suriname
Trio (Ocomayana, Urucuyena [Urucuena], Wama)	Suriname, Brazil
22. Pimenteira†	Brazil
23. Yao†	Fr. Guiana, Trinidad
B. Chocó: Chamí, Sambú [Sambo], Waunana	Colombia
13. Cariri or Kiriri	Brazil
14. Catacao†: Catacao†, Colan	Peru
15. Catembri or Mirandela†	Brazil
16. Catuquina [Catukina]: Bendiapa (Canamari, Parawa), Catuquina [Catukina] or Wiridyapa, Catauxi, Tucundiapa [Tucundyapa]	Brazil
17. Cayuvava (70)	Bolivia
18. Chapacura: Chapacura or Huachi, Itene or Moré, Itoreauhip, Nape, Pacahanovo. Quitemo†	Bolivia
Cumaná (Abitana), Torá, Urupá (Yarú [Jarú]), Wañám [Wanyam]	Brazil
19. Chiquito or Tarapecosi (42)	Bolivia
20. Cholonan: Cholona or Seeptsá†, Híbito†	Peru
21. Cofán	Colombia, Ecuador
22. Culli or Ilinga†	Peru
23. Erikbaktsa or Canoeiro	Brazil
24. Gamela†	Brazil
25. Gorgotoqui	Brazil
26. Guahiboan (4): Chiricoa, Guahibo	Venezuela, Colombia
Churuya, Guayabero	Venezuela
27. Guaycurú-Charruan	
A. Charruan†: Chaná†	Uruguay
Charrúa† (Guenoa or Minuan†)	Uruguay, Argentina, Brazil
B. Guaycuruan: Abipón or Callaga†	Argentina, Paraguay
Caduveo or Mbayá or Guaycurú	Brazil, Argentina, Paraguay
Guachí†	Brazil
Mocoví	Argentina
Payaguá or Lengua	Paraguay
Toba-Pilagá	Argentina, Bolivia
28. Guamo†	Venezuela
29. Guató	Bolivia, Brazil
30. Guennaken or Gununa-Kune or Puelche†	Argentina
31. Huarian: Huari or Corumbiara, Masaca or Aicana	Brazil
32. Huarpean† (37): Allentiac†, Millcayac†	Argentina
33. Iranxe (4)	Brazil
34. Jirajaran: Ayomán, Gayón†, Jirajara	Venezuela
35. Kaliana or Sape	Venezuela
36. Koaia	Brazil
37. Jebero-Jivaroan	
A. Jeberoan or Cahuapanan: Cahuapana or Chuncho [Concho], Chayavita, Jebero [Chébero], Miquirá, Yamorai	Peru
B. Jívaroan: Aguaruna	Peru
Jívaro or Shuara (Achual, Huambisa [Wambisa])	Ecuador
38. Kukura†	Brazil
39. Leco	Bolivia
40. Lulean (48)	
A. Lule or Tonocoté†	
B. Vilela or Chunupí (Atalalá, Ocolé, Uacambabelé)†	Argentina
41. Macro-Chibchan	
A. Chibchan	
1. Abiseta or Orosi or Tucurrique [Tucurrike], Boruca or Turucaca†, Bribrí or Lari, Cabecar, Chiripó, Estrella†, Terraba or Brurán†, Tirub or Rayado†	Costa Rica
2. Andaquí	Colombia
3. Atanque or Busintana, Bintucua or Ijca, Cágaba or Koghi, Guamaca or Arsario, Tairona or Teyuna†	Colombia
4. Bairira or Cunaguasaya, Motilón or Dobocubí	Colombia
Mape	Venezuela
5. Betoi†	Colombia
6. Cara or Imbaya†, Cayapa or Nigua, Colorado or Campaz or Colima or Satxila, Pasto†	Ecuador
Cuaiquier, Muellama†, Telembi†	Colombia
7. Catio†, Nutabé†	Colombia
8. Chibcha or Muisca†, Tunebo	Colombia
9. Chimila, Malibú†	Colombia
10. Changuena†, Chumula† (Gualaca†)	Panama
11. Coconuco, Guambiana or Silviano, Moguex, Totoró	Colombia
12. Corobisi, Guatuso, Guetar, Suerre or Camachi	Costa Rica
13. Cuna	Panama
Cueva	Colombia
14. Guaymí, Move (Penonemeño†)	Panama
15. Panzaleo or Quito†	Ecuador
Paez	Colombia
16. Rama	Nicaragua
17. Sebondoy or Kamsá or Coche or Mocoa, Quillasinga†	Colombia
B. 1. Esmeralda or Atacame†	Ecuador
2. Yaruro	Venezuela
C. Itonama	Bolivia
D. Paya	Honduras
E. Sumo-Mosquito-Matagalpa	
1. Matagalpa† (Cacaopera)	Ecuador, Suriname
Jinotega or Chingo	Nicaragua
2. Mosquito [Miskito]	Honduras, Nicaragua
Ulua, Sumo	Nicaragua
F. Waican or Yanoaman: Karimé [Carimé], Pakidai-Surara, Paucosa	Brazil
Sanemá or Samatari (Pubmatari)	Venezuela
Shamatari, Shirianá or Casapare, Waica [Waika]	Brazil, Venezuela
G. Warao [Warrau] or Guarauno	Venezuela
42. Macro-Ge (70)	
A. Bororoan	
1. Bororo	Brazil
2. Otuké	Bolivia
B. Botocudo or Aymoré†	
C. Fulnió†	Brazil
D. Ge	
1. Akroá†, Xakriaba† [Shacriaba], Xavante [Shavante], Xerente [Sherenté]	Brazil
2. Apinayé-Kayapó, Eastern Timbira, Suyá	Brazil
3. Caingang [Kaingang], Xokleng [Shocleng]	Brazil
4. Southern Kayapó	
E. Jeikó [Jeico]†	Brazil
F. Kamakán†	Brazil
G. Karajá	Brazil
H. Kapoxo† (Kumanaxo†), Malalí†, Maxakalí†, Monoxo†, Patashó†	Brazil
I. Ofayé or Opayé-Shavante†	Brazil
J. Purí-Coroado†: Coroado†, Coropó†, Purí†	
43. Macro-Mayan (4): Uru-Chipaya (Uru, Chipaya)	Bolivia
44. Macro-Pano-Tacanan (4)	
A. Chon: Haush or Manekenken†, Ona or Shelknam†, Tehuelche, Teushen or Tehuesh†	Argentina

Table 62: South American Indian Language Groups* (continued)

language	location
B. Mosetene: Chimane, Mosetene† [Moseten]	Bolivia
C. Pano-Tacanan	
1. Panoan: Amahuaca [Amawaca], Cashinahua [Cashinawa]	Brazil, Peru
Capanahua [Capanawa], Cashibo, Conibo-Shipibo (Chama, Setebo, Sensi), Marinahua [Marinawa], Marobo, Nocamán, Pano or Pánobo	Peru
Culino (Curina), Jaminahua, Mayoruna or Maruba, Nastanahua, Nixinahua, Parannahua, Poyanahua, Remo, Shaminahua, Tushinahua [Tushinawa], Waninahua or Catoquino, Yahuanahua (Yawanawa), Yumanahua	Brazil
Arazaire [Arasaire], Atsahuaca [Atsawaca] or Chaspa, Yamiaca or Haauñeiri	Peru
Caripuna	Brazil
Chácobo, Pacahuara [Pacawara]	Bolivia
2. Tacanan: Arasa, Cavineña, Chama or Esseejja, Guarizo, Huarayo (Tianinagua), Mabenaro, Maropa or Reyesano, Sapiboca [Sapiboka], Tacana (Araona, Toromona)	Bolivia
D. Yuracare	Bolivia
45. Makú	Venezuela, Brazil
46. Mascoy [Mascoi] or Lengua: Angaité (Sanapá), Kashiká, or Guaná, Lengua or Enslet or Cocoloth (Mascoy)	Paraguay
47. Mataco-Maccá [Macá]	
1. Ashluslay or Chulupí	Paraguay
Chorotí or Solote or Yofuaha	Paraguay, Argentina
Choropí (Suhin, Sotirai), Mataco or Mataguayo (Guisnay, Nocten, Vejoz)	Argentina
2. Enimagá or Cochaboth†	Paraguay
3. Maccá [Macá] or Towothli	Paraguay
48. Movima (27)	Bolivia
49. Munichi [Muniche]	Peru
50. Mura-Matanawí	
A. Bohurá, Mura, Pirahá	Brazil
B. Matanawí†	Brazil
51. Murato or Candoshi or Shapra	Peru
52. Nambikwara [Nambicuara]: Central, Eastern Nambikwara	Brazil
53. Omurano or Mayna†	Peru
54. Otomaco-Taparita†: Otomaco†, Taparita†	Venezuela
55. Pankarurú [Pancararú]	Brazil
56. Puinave-Maku: Makú, Marahan, Querari	Brazil
Puinave	Colombia
57. Puquina†: Pohena or Callahuaya†, Puquina	Bolivia
58. Quechumaran	
A. Aymaran: Aymara, Cauqui or Jaqaru	Bolivia, Peru
B. Quechuan: Almaguero, Inga	Colombia
Ancash, Ayacucho, Cajamarca, Chasutino, Huánuco, Junín, Lamano, Lima, Mayna, Pasco, Ucayali	Peru
Catamarca-La Rioja, Santiago del Estero	Argentina
Cuzco-Bolivian	Peru, Bolivia
Ecuadorian, Quijos, Tena	Ecuador
Tuichi	Bolivia
59. Sabelan: Sabela or Auca or Huarani, Tiwituey	Peru
60. Sáliva-Piaroan: Maco (Macu), Piaroa	Venezuela
Sáliva	Colombia
61. Sec or Sechura or Tallán or Atalán†	Peru
62. Simacu or Itucale or Arucuaya or Urariña	Peru
63. Tarariu (Tarairiu) or Ochukuyana† [Ochukayana]	Brazil
64. Taruma†	Brazil
65. Tikuna [Ticuna] or Tukuna [Tucuna]	Brazil, Colombia
66. Timote†	Venezuela
67. Tiniguan: Pamigua, Tinigua†	Colombia
68. Trumai	Brazil
69. Tucanoan	
A. Western: Amaguaje†, Coto, Piojé	Peru
Coreguaje [Correguaje], Dätuana, Icaguaje†, Macaguaje, Macuna, Sära, (Ömöa, Buhagana), Siona [Sioni], Tama, Tanimuca, Uantia, Yahuna, Yupua	Colombia
Coretu	Brazil
Secoya	Ecuador
B. Eastern: Bara, Erulia (Paneroa, Tsölöa), Karapana [Carapaná], (Möchda), Pamöá, Siana or Chiranga, Tatapuyo, Waiana, Yarutí or Patsoca	Colombia
Desana, Tuyuka (Tuyuca), Wanana (Waikina)	Colombia, Brazil
Kubeo (Cubeo), Tucano	Brazil
70. Tupian	
A. 1. (Tupí-Guaraní): Apiaká† [Apiacá], Awetí, Canoeiro, Kamayurá [Camayura], Kawaíb [Cawahíb] (Pawate, Parintintin, Wirafed) Kayabí (Cayabí), Shetà†, Takuñapé† [Tacunyapé], Tapirapé, Tenetéhara (Anambé, Guayayara [Guajajára], Manajé, Tembé, Turiwara†, Urubú), Tupí-Guaraní (Tupí-nambá), Neengatú	Brazil
Oyampí-Emerillon†	Brazil, Fr. Guiana
Pauserna	Bolivia, Brazil
Guaraní	Argentina, Brazil, Paraguay
Kaiwá	Brazil, Paraguay
Chiriguano, Guarayú	Bolivia
Tapieté, Chané	Paraguay
2. Cocama	Peru
Omagua	Brazil, Peru
3. Guayakí	Paraguay
4. Mawé	Brazil
5. Sirionó	Bolivia
B. Arara, Ramarama (Itogapid), Urukú, Urumí	Brazil
C. Arikem†, Kabishiana†, Karitiana†	Brazil
D. Arué, Digüt, Mondé†	Brazil
E. Guarategaya† (Amniapé†, Kanoe, Mekens), Kepkiriwat†, Makurap†, Tuparí, Wayoró (Apichum)	Brazil
F. Kuruaya† (Curuaya), Mundurukú	Brazil
G. Manitsawá†, Shipaya†, Yuruna	Brazil
H. Puruborá	Brazil
71. Tushá	Brazil
72. Tuyoneiri or Arasairi or Huachipairi	Peru
73. Uman or Huamoi	Brazil
74. Xukurú or Ichikile	Brazil
75. Yabutí†: Aricapú†, Mashubi†, Yabutí or Quipiu†	Brazil
76. Yagua: Masamae, Peba or Nijamo, Yagua or Mishara, Yameo or Camuchivo	Peru
77. Yámana† or Yaghan (41)	Chile
78. Yunca or Chimú or Muchic or Chincha†: Puruhá-Cañari†	Ecuador
Yunca†	Peru
79. Yurí†	Brazil, Colombia
80. Yurimangui†	Colombia
81. Zamuco: Chamacoco (Ebidoso, Tumrahá), Zamuco (Ayoré, Moro)	Paraguay
82. Záparo: Arabela, Iquito (Cahuarano), Shimigae† (Semigae)†, Andoa, Záparo†	Peru

*Only languages attested linguistically are included. Extinct languages are marked with †. A number in parentheses after the name of a group indicates a possible relationship with the group identified by that number. Languages are separated by commas, names in parentheses are of dialects, and names in brackets are alternative spellings. Except for Arawakan, Macro-Ge, and Tupian, most groupings are geographical, but those identified by capital letters represent in general markedly differentiated groups. Spelling follows the most common usage for each language or group, thus it is not consistent. Equivalent spellings: *b* = *v*; *g* = *j* = *y*; *gu* = *hu* = *u* = *w*; *i* = *y*; *h* = (nothing); *h* = *j*; *k* = *c* (before *a*, *o*, *u*); *k* = *qu* (before *i*, *e*); *sh* = *x* = *ch*; *s* = *z*; *ñ* = *nh* = *ny*; *x* = *j*. Names are arranged alphabetically within each subdivision.

most important group today—Chocó in western Colombia—is distantly related to the rest of the stock. Other languages are Carib in Suriname, Trio in Suriname and Brazil, and Waiwai, Taulipang, and Makushí (Macusí) in Brazil. A relationship with Tupian seems certain.

The Tupí and Guaraní languages

Tupian. With the exception of Emerillon and Oyampi of French Guiana and northeastern Brazil, Tupian languages were spoken south of the Amazon, from the Andes to the Atlantic Ocean and down to the Río de la Plata. There are approximately 50 attested languages related on the stock level and subdivided into eight families. Tupinambá, the language spoken along the Atlantic coast at the time of discovery, became important in a modified form as a lingua franca, and the closely related Guaraní became the national language in Paraguay, being one of the few Indian languages that does not seem to yield under the influence of Spanish or Portuguese. At the time of discovery, Tupí-Guaraní tribes were moving everywhere south of the Amazon, subjugating other tribes; some of these tribes adopted Tupí-Guaraní. Both Tupí and Guaraní are among the languages that have exerted a great influence on Portuguese and Spanish language. Tupí groups have declined markedly, 26 groups becoming extinct in Brazil between 1900 and 1957, and at least 14 languages disappearing during the same period. The westernmost language, Cocama in Peru, is still spoken by about 19,000 speakers, and Chiriguano in Bolivia has about 20,000 speakers. Other languages have a much smaller number of speakers; there are 19,000 speakers for the 26 surviving groups in Brazil. The total number of Indian speakers of Tupian languages is approximately 60,000, but there are also about 3,000,000 culturally non-Indian speakers of Guaraní in Paraguay. Besides the connection with Cariban, further relationships possibly exist with Macro-Ge, various small families like Zamuco and Mataco-Maccá and isolated languages like Cayuvava.

Macro-Ge. Macro-Ge is geographically the most compactly distributed of the big South American language families. Ge proper extends uninterruptedly through inland eastern Brazil almost as far as the Uruguayan border. There are about ten Ge languages with a total of 2,000 speakers. Most of the other families, now extinct, were located closer to the Atlantic coast, from where they probably were displaced by Tupian expansion. The Bororan family is represented by Bororo in Brazil and by the Otuké language in Bolivia. It seems likely that Macro-Ge has its closest relationship with Tupian.

Quechumaran. Quechumaran, which is composed of the Quechuan and Aymaran families, is the stock with the largest number of speakers—7,000,000 for Quechuan and 1,000,000 for Aymaran—and is found mainly in the Andean highlands extending from southern Colombia to northern Argentina. The languages of this group have also resisted displacement by Spanish, in addition to having gained in numbers of speakers from the time of the Incas to the present as several other groups adopted Quechuan languages. Cuzco-Bolivian Quechua is spoken by well over 1,000,000 speakers, and there are around seven Quechuan languages in Peru with almost 100,000 speakers each. Although most Quechuan languages have been influenced by Spanish, Quechuan in turn is the group that has exerted the most pervasive influence on Spanish. No convincing further genetic relationship has been yet proposed.

Tucanoan. Tucanoan, which is spoken in two compact areas in the western Amazon region (Brazil, Colombia, and Peru), includes about 30 languages with a total of

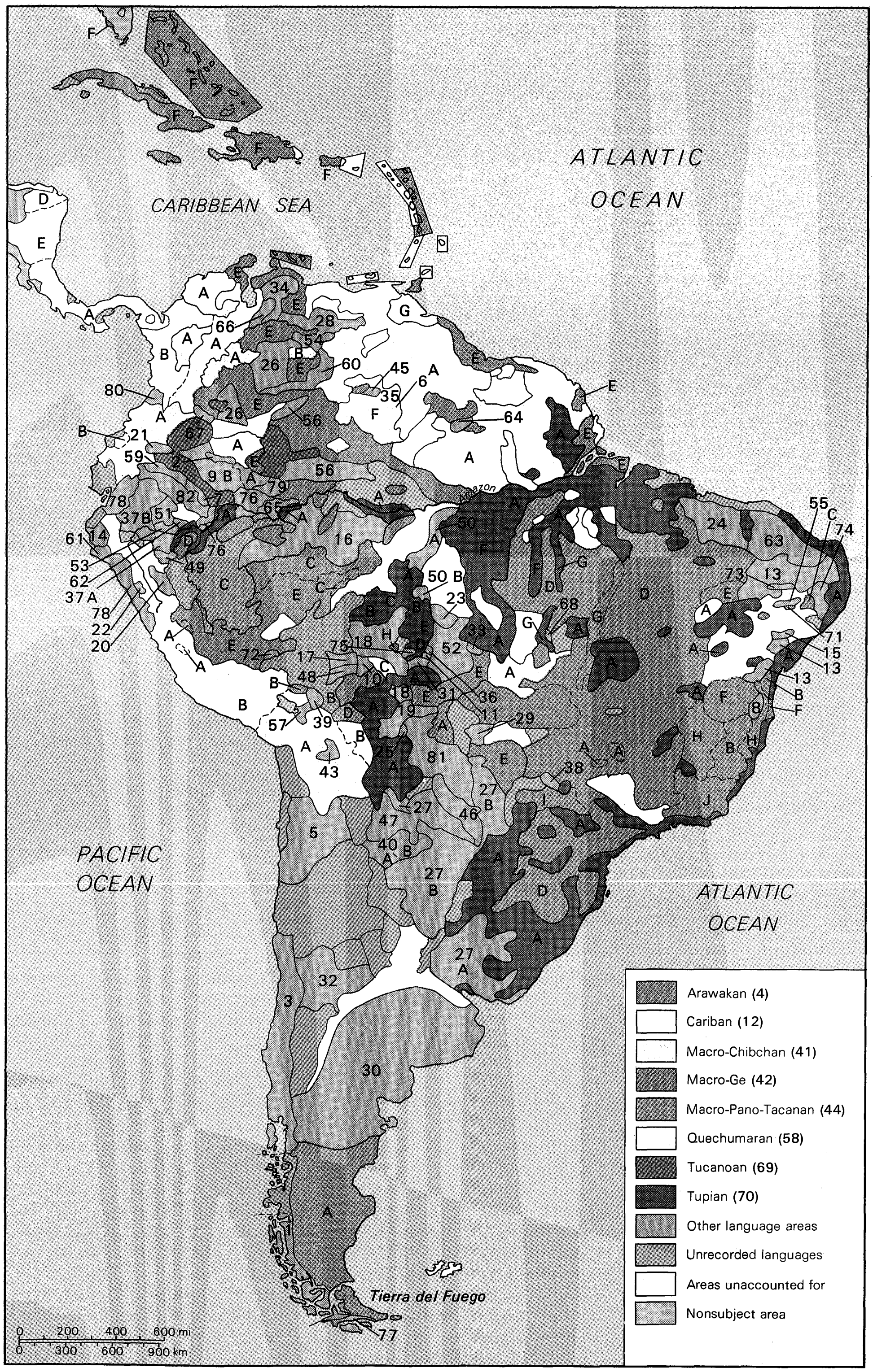


Figure 36: Distribution of the South American Indian languages. The numbers and letters refer to the numbers and letters in Table 62.

over 30,000 speakers. One of the languages is a lingua franca in the region.

Macro-Pano-Tacanan. Macro- Pano-Tacanan, a group more distantly related than a stock, includes about 30 languages, many of them still spoken. The languages are located in two widely separated regions: lowland eastern Peru and adjoining parts of Brazil and lowland western Bolivia on the one hand, and southern Patagonia and Tierra del Fuego on the other. In the latter region the languages are practically extinct.

Other language groups

By number of component languages, or by number of speakers, or by territorial extension, the other language groups are not as significant as those just listed. Most of these small families and isolated languages are located in the lowlands, which form an arch centred on the Amazon from Venezuela to Bolivia and include the bordering parts of Brazil.

Lingua francas and cultural tongues. Lingua francas as well as situations of bilingualism arose mainly under conditions furthered or created by Europeans, although a case like that of the Tucano language, which is used as a lingua franca in the Río Vaupés area among an Indian population belonging to some 20 different linguistic groups, may be independent of those conditions. Quechua, originally spoken in small areas around Cuzco and in central Peru, expanded much under Inca rule, coexisting with local languages or displacing them. It was the official language of the Inca Empire, and groups of Quechua speakers were settled among other language groups, although the language does not seem to have been systematically imposed. The Spaniards, in turn, used Quechua in a great area as a language of evangelization—at one period missionaries were required to know the language—and continued to spread it by means of Quechua speakers who travelled with them in further conquests. During the 17th and 18th centuries it became a literary language in which religious, historical, and dramatic works were written. Today its written literary manifestations are not spontaneous, but there is abundant oral poetry, and in Bolivia radio programs are broadcast entirely in this language.

Dispersion of Tupí-Guaraní dialects, taking place shortly before the arrival of Europeans and even after it, resulted not from imperial expansion—as for Quechua—but from extreme tribal mobility and the cultural and linguistic absorption of other groups. Under Portuguese influence the modified form of Tupinamba known as *língua-geral* ("general language") was the medium of communication between Europeans and Indians and among Indians of different languages in Brazil. It was still in common use along the coast in the 18th century, and it is still spoken in the Amazon. Tupí, now extinct, was an important language of Portuguese evangelization and had a considerable literature in the 17th and 18th centuries. Another dialect, Guaraní, was the language of the Jesuit missions and also had abundant literature until the middle of the 17th century when the Jesuits were expelled and the missions dispersed. Nevertheless, Guaraní survived in Paraguay as the language of a culturally non-Indian population and is today the only Indian language with national, although not official, status—persons not speaking Guaraní being a minority. Paraguayan Guaraní is also a literary language, not so much for learned works—for which Spanish is used—but for those of popular character, especially songs. There is a more or less standardized orthography, and persons literate in Spanish are also literate in Guaraní. A great mutual influence exists between Guaraní and Spanish.

Grammatical diversity

Grammatical characteristics. Diversity rather than common traits characterizes the grammar of South American Indian languages. Features commonly encountered seem to reflect facts of frequency in general typology rather than traits specific to this area. The greatest number of languages are probably suffixing languages like Quechumaran and Huitotoan, or use many suffixes and some prefixes like Arawakan and Panoan. Also very numerous are those languages having few prefixes and suffixes, such as Ge, Carib, or Tupian. Languages employing only prefixes to show grammatical distinctions have not been reported. There are a few with many prefixes but still more suffixes (Jebero, or Chêbero); others, like Ona and Tehuelche, with almost no affixing, are also rare.

Similarly, the complexity of words varies a great deal. In Guaraní words with three components and in Piro (Arawakan) words with six elements are of average complexity for the respective languages. In languages like the Cariban or Tupian ones, word roots are nominal (nouns) or verbal (verbs) and may be converted into the other class by derivational affixes; in languages like Quechua or Araucanian, many word roots are both nominal and verbal. Languages like Yuracare form many words by reduplication (the repetition of a word or a part of a word), a process that does not occur systematically in the Tupian languages. Compounding, the joining of two or more words to form new words, is a very widespread type of word formation, but it can be nearly absent, as in the Chon languages. Verb stems in which the nominal (noun) object is incorporated are also rather frequent. Many languages are of the agglutinative type (Quechuan, Panoan, Araucanian); *i.e.,* they combine several elements of distinctive meaning into a single word without changing the element. Others (Cariban, Tupian) show a moderate amount of change and fusion of the elements when combined in words.

Grammatically marked gender in nouns occurs in Guaycuruan (Guaicuruan), and a difference in masculine and feminine gender in the verb occurs in Arawakan, Huitotoan (Witotoan), and Tucanoan, but genderless languages are more common. Singular and plural in the 3rd person ("he, she, it") is not obligatorily distinguished in Tupian and Cariban, but languages like Yámana and Araucanian have singular, dual, and plural. A very common distinction is that between inclusive 1st person ("you and I," hearer included) and exclusive 1st person ("he and I," hearer exluded). Pronominal forms differentiated according to categories that indicate whether the person is present or absent, sitting or standing, and so forth occur in Guaycuruan languages and Movima. Case relations in nouns are generally expressed by suffixes or postpositions; the use of prepositions is rare. Possession is indicated predominantly by prefixes or suffixes, and systems in which possessive forms are the same as those used as the subject of intransitive verbs and as the object of transitive ones are rather common. Classificatory affixes that subclassify nouns according to the shape of the object occur in the Chibchan, Tucanoan, and Waican groups.

Indication of case relationships

Very frequently the verbal forms express the subject, object, and negation in the same word. The categories of tense and aspect seem to be about evenly represented in South American languages, but the specific categories expressed vary a great deal from language to language: Aguaruna (Jívaroan) has a future form and three past forms differentiated as to relative remoteness, while in Guaraní the difference is basically between future and nonfuture. Other languages like Jebero express fundamentally modal categories. Very common are affixes indicating movement, chiefly toward and away from the speaker, and location (*e.g.,* in Quechumaran, Záparo, Itonama), and in some stocks like Arawakan and Panoan there are many suffixes in the verb with very concrete adverbial meaning, such as "by night," "during the day." Classificatory affixes indicating the way the action is performed—by biting, striking, walking—occur in Jebero and Tikuna (Ticuna). Actions done individually or collectively are differentiated paradigmatically in Carib, while in Yámana and Jívaro different verbal stems are used according to whether the subject or the object is singular or plural. There are also various languages (Guaycuruan, Mataco, Cocama) in which some words have different forms according to the sex of the speaker.

Equational sentences are very common. These are formed by juxtaposing two nominal expressions (nouns) without a linking verb, a fact that usually correlates with the absence of a verb "be" for expressing identification or location (*e.g.,* "John good man," "my house there"). Sentences in which the predicate is a noun inflected like a verb with the meaning "being" or "having" that thing designated by the noun also occur in Bororo and Huitoto (Witoto); *e.g.,* "I-knife" = "I have a knife." Sentences in which the subject is the undergoer of the action are frequent, but true passive sentences in which the undergoer and the agent are expressed are rare, though they do occur in Huitoto. Subordinate sentences are rarely introduced by conjunctions; subordination is usually expressed by postposed elements or special forms of the verbs such as gerunds, participles, or subordinate conjugations.

Phonological characteristics. As in grammar, there are no phonological features common to all South American languages that would be specific to them alone. The number of distinctive sounds (phonemes) may vary from 42 in Jaqaru (Quechumaran) to 17 in Campa (Arawakan). Jaqaru has 36 consonants, while Makushí (Cariban) has 11; some Quechuan languages have only three vowels, whereas Apinayê (Macro-Ge) has ten oral vowels and seven nasal ones. A dialect of Tucano (Tucanoan) exhibits three contrasting points of articulation, while Chipaya (Macro-Mayan) has nine. Many types of contrasting sounds occur although not with equal frequency. Voiceless stops (*e.g.*, *p*, *t*, *k*) occur everywhere, but voiced stops (*e.g.*, *b*, *d*, *g*) may be absent, and fricatives (*e.g.*, *f*, *v*, *s*, *z*) may be few in number. Glottalized voiceless stops—consonants made with simultaneous closure of the glottis and without vibration of the vocal cords—are rather common (Quechumaran, Chibchan), but not glottalized voiced stops (in which the vocal cords vibrate). Also less frequent are aspirated (Quechumaran) and palatalized sounds (Puinave); glottalized nasal sounds (Movima) and voiceless laterals (*l*-like sounds, as in Vilela) are rare. A distinction between velar and postvelar sounds occurs in Quechumaran and Chon, between velar and labiovelar in Tacana and Siona (Sioni); palatal retroflex consonants, made with the tip of the tongue turned up touching the palate, occur in Pano-Tacanan and Chipaya.

Great variety in distinctive sounds

Systems with nasal vowels are common (Macro-Ge, Sabelan), but in several languages (Tupian, Waican) nasalization is a feature not of vowels and consonants but of whole words. There is an apparent absence of front rounded vowels (*ü*, *ö*), but central or back unrounded vowels (ɨ, *i*) are common. Systems with long vowels occur in Chipaya and some Cariban languages, and glottalized vowels occur in Tikuna and Chon languages. Very common are pitch-stress systems with high and low tones on stressed syllables; *e.g.*, in Panoan, Huitotoan, and Chibchan. More complex systems with three tones as in Acaricuara, four as in Mundurukú (Mundurucú), and five as in Tikuna are rare. Syllables are generally without complex consonant clusters.

The typology proposed by Tadeusz Milewski, a Polish linguist, classifies American Indian languages into three types: (1) Atlantic, with few oral consonants but complex systems of nasal consonants, and oral and nasal vowels, of which the Ge languages would be typical; (2) Pacific, with complex systems of oral consonants (many contrasting points and modes of articulation) but with few nasal consonants and few vowels, as exemplified by Quechumaran; and (3) Central, with consonant systems more like the Pacific type and vowel systems like the Atlantic, of which Chibcha would be typical. The typology is probably too gross to accommodate meaningfully every language type found in South America, but it holds to a certain extent, especially for the Atlantic type (Macro-Ge, Tupian, and Cariban).

Vocabulary. Indian languages vary significantly in the number of loanwords from Spanish and Portuguese. Massive borrowing has taken place in areas where languages have been in intense and continued contact with Spanish or Portuguese, especially where groups are economically dependent on the national life of the country and there is a considerable number of bilingual persons, as in Quechuan, or where no cultural differences correlate with language differences, as in Paraguayan Guaraní. Borrowings have not been limited to designations of artifacts of European origin but affect all spheres of vocabulary, having displaced native terms in many cases. Neither are they limited to lexical items; they include function elements such as prepositions, conjunctions, and derivative suffixes. Sound systems have also been modified. In some contact situations in which the Indian group displayed an antagonistic attitude toward the European conquest, purism developed and loans are comparatively few; *e.g.*, Araucanian. When contact has been frequent but superficial, loanwords are usually scant, but the meaning of native terms has shifted or new descriptive terms have been coined to designate new cultural traits, as in Tehuelche.

Loanwords from Spanish and Portuguese

Borrowings among Indian languages may have been more numerous than yet reported, judging from the wide and rapid diffusion that loans from Spanish and Portuguese had through the central part of South America. Borrowings between Quechua and Aymara have occurred in great number, but the direction of borrowing is difficult to determine. Many Indian languages in the Andes and the eastern foothills have borrowed from Quechua either directly or through Spanish. In Island Carib (an Arawakan language), borrowings from Carib (a Cariban language) have formed a special part of the vocabulary, properly used only by men; these words were adopted after the Island Carib speakers were subjugated by Caribs.

In turn, some Indian languages have been a source of borrowing into European languages. Taino (Arawakan), the first language with which Spaniards had contact, furnished the most widespread borrowings, including "canoe," "cacique," "maize," and "tobacco," among many others. No other South American Indian language has furnished such widespread and common words, although Quechua has contributed some specialized items such as "condor," "pampa," "vicuña." The larger number of Arawakan borrowings results from these languages having been predominant in the Antilles, a region where Dutch, French, English, Portuguese, and Spanish were present for a long time. Cariban languages, the other important group in that region, do not seem to have furnished many words, but "cannibal" is a semantically and phonetically modified form of the self-designation of the Caribs. The influence of some Indian languages on regional varieties of Spanish and Portuguese has been paramount. Thus Tupí accounts for most Indian words in Brazilian Portuguese, Guaraní in the Spanish of Paraguay and northeast Argentina; and Quechua words are abundant in Spanish from Colombia to Chile and Argentina. In addition, Quechuan and Tupí-Guaraní languages account for most place-names in South America.

Language and culture

No detailed studies are available concerning the relationship of the vocabularies of Indian languages to the culture. Certain areas of vocabulary that are particularly elaborated in a given language may reflect a special focus in the culture, as for example the detailed botanical vocabularies for plants of medical or dietary importance in Quechua, Aymara, and Araucanian. Shifts in cultural habits may also be reflected in the vocabulary, as in Tehuelche, which formerly had a vocabulary designating different kinds of guanaco meat that is now very much reduced, because the group no longer depends on that animal for subsistence. Kinship terminology is usually closely correlated with social organization so that changes in the latter are also reflected in the former: in Tehuelche, former terms referring to paternal and maternal uncles tend to be used indiscriminately, even replaced by Spanish loans, because the difference is not functional in the culture any more.

Proper names, to which different beliefs are attached, offer a variety of phenomena, among them the practice of naming a parent after a child (called teknonymy) in some Arawakan groups; the repeated change of name according to various fixed stages of development, as in Guayaki; word taboo, forbidding either the pronunciation of one's own name or the name of a deceased person, or both, as in the southernmost groups (Alacaluf, Yâmana, Chon) and in the Chaco area (Toba, Terena); and the use of totemic names for groups, as in Panoan tribes.

Writing and texts. The existence of pre-Columbian native writing systems in South America is not certain. There are two examples, that of the Cuna in Colombia and an Andean system in Bolivia and Peru, but in both cases European influence may be suspected. They are mnemonic aids—a mixture of ideograms and pictographs—for reciting religious texts in Quechua and ritual medical texts in Cuna. The Cuna system is still in use.

Although the linguistic activity of missionaries was enormous and their work, from a lexicographic and grammatical viewpoint, very important, they failed to record texts reflecting the native culture. The texts they left for most languages are, with a few exceptions, of a religious nature. Most of the folklore has been collected in the 20th century, but many important collections (*e.g.,* for the Fuegian and Tacanan tribes) are not published in the native language but rather in translation. There are good texts recorded in the native language for Araucanian, Panoan, and Cuna, for instance, and more are being recorded by linguists now, though not necessarily analyzed from a linguistic point of view.

Efforts are being made in several areas to introduce literacy in the native Indian languages. For some, practical orthographies have existed since the 17th century (Guaraní, Quechua); for several others, linguists have devised practical writing systems and prepared primers in recent years. The success of these efforts cannot yet be evaluated. (J.A.S.)

LANGUAGE ISOLATES

Sumerian language

Sumerian is the oldest written language in existence. First attested about 3100 BC in southern Mesopotamia, it flourished during the 3rd millennium BC. About 2000 BC, Sumerian was replaced as a spoken language by Semitic Akkadian (Assyro-Babylonian) but continued in written usage almost to the end of the life of the Akkadian language, around the beginning of the Christian era. Sumerian never extended much beyond its original boundaries in southern Mesopotamia; the small number of its native speakers was entirely out of proportion to the tremendous importance and influence Sumerian exercised on the development of the Mesopotamian and other ancient civilizations in all their stages.

History. Four periods of Sumerian can be distinguished: Archaic Sumerian, Old or Classical Sumerian, New Sumerian, and Post-Sumerian.

Archaic Sumerian covered a period from about 3100 BC, when the first Sumerian records make their appearance, down to about 2500 BC. The earliest Sumerian writing is almost exclusively represented by texts of business and administrative character. There are also school texts in the form of simple exercises in writing signs and words. The Archaic Sumerian language is still very poorly understood, partly because of the difficulties surrounding the reading and interpretation of early Sumerian writing and partly because of the meagreness of sources.

Old, or Classical, Sumerian

The Old, or Classical, period of Sumerian lasted from about 2500 to 2300 BC and is represented mainly by records of the early rulers of Lagash. The records are business, legal, and administrative texts, as well as royal and private inscriptions, mostly of votive character; letters, both private and official; and incantations. These sources are much more numerous than those of the preceding period, and the writing is explicit enough to make possible an adequate reconstruction of Sumerian grammar and vocabulary.

During the period of the Sargonic dynasty, the Semitic Akkadians took over the political hegemony of Babylonia, marking a definite setback in the progress of the Sumerian language. At this time the Akkadian language was used extensively throughout the entire area of the Akkadian empire, while the use of Sumerian gradually was limited to a small area in Sumer proper. After a brief revival during the 3rd dynasty of Ur, the New Sumerian period came to an end about 2000 BC, when new inroads of the Semitic peoples from the desert succeeded in destroying the 3rd dynasty of Ur and in establishing the Semitic dynasties of Isin, Larsa, and Babylon.

The period of the dynasties of Isin, Larsa, and Babylon is called the Old Babylonian period, after Babylon, which became the capital and the most important city in the country. During this time the Sumerians lost their political identity, and Sumerian gradually disappeared as a spoken language. It did, however, continue to be written to the very end of the use of cuneiform writing. This is the last stage of the Sumerian language, called Post-Sumerian.

In the early stages of the Post-Sumerian period the use of written Sumerian is extensively attested in legal and administrative texts, as well as in royal inscriptions, which are often bilingual, in Sumerian and Babylonian. Many Sumerian literary compositions, which came down from the older Sumerian periods by way of oral tradition, were recorded in writing for the first time in the Old Babylonian period. Many more were copied by industrious scribes from originals now lost. The rich Sumerian literature is represented by texts of varied nature, such as myths and epics, hymns and lamentations, rituals and incantations, and proverbs and the so-called wisdom compositions. For many centuries after the Old Babylonian period, the study of Sumerian continued in the Babylonian schools. As late as the 7th century BC, Ashurbanipal, one of the last rulers of Assyria, boasted of being able to read the difficult Sumerian language, and from an even later period, in Hellenistic times, there are some cuneiform tablets that show Sumerian words transcribed in Greek letters.

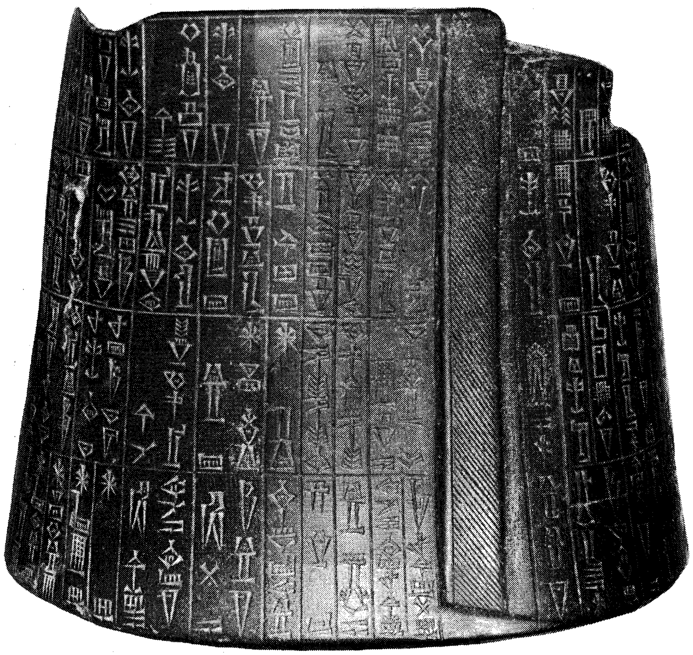


Figure 37: Sumerian inscription, detail of a diorite statue of Gudea of Lagash, 22nd century BC. In the Louvre, Paris.
Archives Photographiques

Rediscovery. Around the time of Christ, all knowledge of the Sumerian language disappeared along with that of cuneiform writing, and in the succeeding centuries even the name Sumer vanished from memory.

Unlike Assyria, Babylonia, and Egypt, whose histories and traditions are amply documented in biblical and classical sources, there was nothing to be found in non-Mesopotamian sources to make one even suspect the existence of the Sumerians in antiquity, let alone fully appreciate their important role in the history of early civilizations.

Decipherment of Sumerian cuneiform

When the decipherment of cuneiform writing was achieved in the early decades of the 19th century, three languages written in cuneiform were discovered: Semitic Babylonian, Indo-European Persian, and Elamite, of unknown linguistic affiliation. Only after the texts written in Babylonian had become better understood did scholars become aware of the existence of texts written in a language different from Babylonian. When the new language was discovered it was variously designated as Scythian, or even Akkadian (that is, by the very name now given to the Semitic language spoken in Babylonia and Assyria). It was only after knowledge of the new language had grown that it was given the correct name of Sumerian.

Characteristics. The linguistic affinity of Sumerian has not yet been successfully established. Ural-Altaic (which includes Turkish), Dravidian, Brahui, Bantu, and many other groups of languages have been compared with Sumerian, but no theory has gained common acceptance. Sumerian is clearly an agglutinative language in that it preserves the word root intact while expressing various grammatical changes by adding on prefixes, infixes, and suffixes. The difference between nouns and verbs, as it exists in the Indo-European or Semitic languages, is unknown to Sumerian. The word *dug* alone means both "speech" and "to speak" in Sumerian, the difference between the noun and the verb being indicated by the syntax and by different affixes.

The distinctive sounds (phonemes) of Sumerian consisted of four vowels, *a, i, e, u,* and 16 consonants, *b, d, g, ŋ, h, k, l, m, n, p, r, s, ś, š, t, z.* In Classical Sumerian, the contrast between the consonants *b, d, g, z* and *p, t, k, s* was not between voiced (with vibrating vocal cords) and voiceless consonants (without vibrating vocal cords) but between consonants that were indifferent as to voice and those that were aspirated (pronounced with an accompanying audible puff of breath). The semivowels *y* and *w* functioned as vocalic glides.

In the noun, gender was not expressed. Plural number was indicated either by the suffixes *-me* (or *-me + esh*), *-hia,* and *-ene,* or by reduplication, as in *kur + kur* "mountains." The relational forms of the noun, corresponding approximately to the cases of the Latin declension, include: *-e* for the subject (nominative), *-a*(*k*) "of" (genitive), *-ra* and *-sh*(*e*) "to," "for" (dative), *-a* "in" (locative), *-ta* "from" (ablative), *-da* "with" (commitative).

The Sumerian verb

The Sumerian verb, with its concatenation of various prefixes, infixes, and suffixes, presents a very complicated picture. The elements connected with the verb follow a rigid order: modal elements, tempo elements, relational elements, causative elements, object elements, verbal root, subject elements, and intransitive present–future elements. In the preterite transitive active form, the order of object and subject elements is reversed. The verb can distinguish, in addition to person and number, transitivity and intransitivity, active and passive voice, and two tenses, present-future and preterite.

Several Sumerian dialects are known. Of these the most important are *eme-gir,* the official dialect of Sumerian, and *eme-SAL,* the dialect used often in the composition of hymns and incantations (see also WRITING).

(I.J.G.)

Etruscan language

The Etruscan language was spoken by close neighbours of the ancient Romans. The Romans called them Etrusci or Tusci; in Greek they were called Tyrsenoi or Tyrrhenoi; in Umbrian, and Italic language, their name can be found in the adjective *turskum.* The Etruscans' name for themselves was *rasna* or *raśna.*

The Etruscans lived in Italy in the region of modern Tuscany, in an area bounded by the Arno River on the north, the Tiber River on the southeast, and the Tyrrhenian Sea on the west. At one time they controlled most of an area extending south from Milan through Marzabotto and Sarsina to the Adriatic Sea north of Ancona, and to the southwest their rule extended as far as Capua, Naples, and Pompeii. For the history of the Etruscans and Etruria, see GRECO-ROMAN CIVILIZATION: *Ancient Italic peoples.*

Inscriptions

Records and scholarship. The Etruscan language is known mainly from epigraphic records originating in the Tuscan area and dating from the 7th century BC to the first years of the Christian Era. There are some 10,000 of these inscriptions, mainly brief and repetitious epitaphs or dedicatory formulas, as well as votive or owner's inscriptions on paintings in tombs and accompanying engraved figures on small artifacts such as metal mirrors. There are, however, some remarkable exceptions to the general brevity of the inscriptions, and there are important differences in their origins. The longest single text, of 281 lines (about 1,300 words), now in the National Museum at Zagreb, is written on a roll of linen that had been cut into strips and used in Egypt as a wrapping for a mummy; a clay tablet found at Capua contains some 250 words; a stone slab from Perugia has two adjacent sides elegantly engraved with an inscription of 46 lines (some 125 words); a bronze model of a liver found at Piacenza, which probably represents the Etruscan microcosm in a form used for instruction in divination, has some 45 words; and a heavy rectangular block found on the island of Lemnos in the northern Aegean has an engraving of what is probably a warrior with one inscription of perhaps 18 words surrounding the head and another of 16 words in three lines on an adjacent side. In 1964 two inscriptions on gold tablets, one in Phoenician and the other in Etruscan, were unearthed at Pyrgi.

In the Museo Archeologico Nazionale dell'Umbria, Perugia, Italy

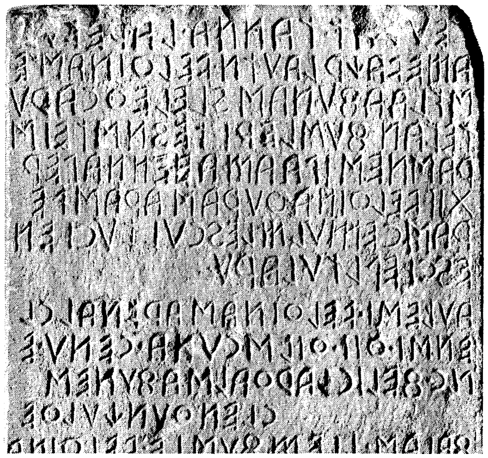


Figure 38: Etruscan inscription from a section of the "Cippus of Perugia."

Despite many attempts at decipherment and some claims of success, the Etruscan records still defy translation. While the possibility always remains that an imaginative conjecture or a brilliant inference will suddenly provide the key to the mystery, this now seems remote. The etymological method of investigation, which ultimately depends upon the recognition of presumed cognates from related languages, seems to have failed because no clear and certain relationship between Etruscan and any other language has ever been established. The procedure sometimes called the combinatory method now appears to be the most efficacious if not indeed the only useful one. It requires, first, that note be made of anything unusual in the provenance of the object on which Etruscan writing is found (such as that the mummy wrapping came from Egypt and the Lemnos inscription from the Aegean) and likewise of anything unusual in the object itself (*e.g.,* that it is a bronze replica of a liver or the representation of a god or mythological figure). Finally, each word and phrase and formula is compared with every recurrence of the same element or elements elsewhere, and all variations in the physical and the linguistic contexts are recorded. By this means it has been possible to assign some words to grammatical categories such as noun and verb, to identify some inflectional endings, and to assign meanings to a few words of very frequent occurrence.

Linguistic relationships

The problem of Etruscan origins is insoluble until the language can be translated. While nothing at all is certain other than the existence of Etruscans in Italy, some Etruscan writing in Egypt, and an Etruscan inscription on Lemnos, the weight of all the evidence seems to favour a non-Italic but certainly Mediterranean place of origin of the Etruscan people. It is unlikely, therefore, that the Etruscan language is genetically related to any language or language family existing in an area remote from the Mediterranean. On the other hand, it does not follow that Etruscan must be related to a language or language fragment that can be found in the Mediterranean area.

The Etruscan alphabet. Etruscan is written in an alphabet probably derived from one of the Greek alphabets.

It is of very great importance that Etruscan is written in a recognizable alphabet related to the Greek and Semitic because sound values can be assigned with some degree of precision to each symbol. Etruscan writing proceeded from right to left and in earliest times had no word division or punctuation. In about the 6th century BC a system of points, or dots, consisting of four, three, or two dots inscribed vertically, was introduced to mark word boundaries and, in some instances, apparently, to indicate syllables and possibly abbreviations.

By courtesy of the Soprintendenza alle Antichita, Florence

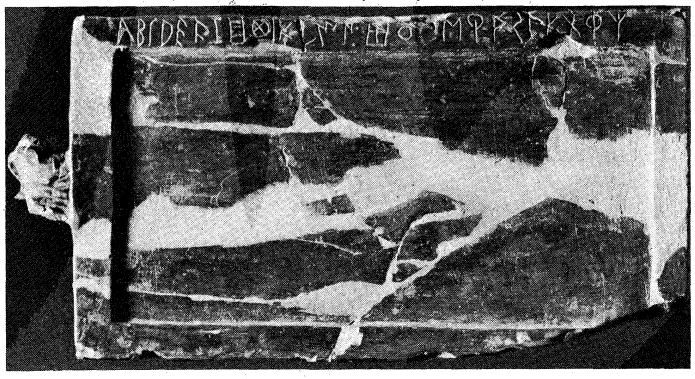


Figure 39: Etruscan alphabet on the edge of a writing tablet from Marsiliana d'Albegna, 7th century. In the Museo Archeologico, Florence.

There were four vowels in Etruscan, *i, e, a,* and *u* or *o,* and symbols in the alphabet for *p, t, c, m, n, l, r, z* and for the equivalents of the Greek *phi, theta,* and *chi,* which in Etruscan as in classical Greek were the aspirated stops *ph, th, ch* (pronounced as *p, t, k* with an added brief puff of air). There were two sibilants, written *s* and *ś,* for which the precise pronunciation is uncertain; two front fricatives, *f* and *v,* articulated either with the two lips (bilabial) or with the lower lip approaching the upper front teeth (labiodental); and an *h,* which nearly always occurs at the beginning of words and is used to represent, inconsistently, the rough breathing of Greek (*e.g.,* Greek *Hēraklēs,* Etruscan *hercle* or *ercle*). There were also a *k* and a *q,* of which the precise pronunciation is unknown. A marked tendency to make all vowels in a word similar or identical (qualitative vowel harmony) is characteristic: Greek *Klutaimēstra,* which if transliterated directly into Etruscan would be *cluthemestha,* actually occurs as *cluthumustha* and *clutmsta.*

Both historical changes and dialectal differences can be observed. Diphthongs became single letters. Thus Greek *Aiwas* became Etruscan *aivas, eivas,* and *evas,* successively; *au* alternated with *a*; *eu* (like *ai*) became *e* (Greek *Kleopatra* is Etruscan *clepatra;* Greek *Poludeukēs* is, with Etruscan vowel harmony, *Pultuce*). Among consonants the most noticeable changes are *c* to *ch* to *h* (*e.g., casri* becomes *chasri, caspr* becomes *haspr*); similarly, *p* changes to *ph* to *f* to *h* and *t* to *th* to *h.* Throughout the history of Etruscan, a first syllable usually remains unchanged, whereas later syllables tend to weaken or lose vowels, at least in writing. Older Etruscan *lavtun* "family" becomes in later Etruscan *lavtn;* other examples are *mutana* changing to *mutna,* Greek *Adonis* written *atunis* and then *atuns,* Greek *Alexandros* appearing as *elchsntre.* The consonant cluster of *elchsntre,* while extreme, is not untypical of Etruscan spelling; words thus written have led some to suggest that a very economical spelling system may have been used that was far removed from the reality of pronunciation, requiring the introduction of lightly stressed vowels in actual utterance. (For a short history and a chart of the Etruscan alphabet, see WRITING: *Alphabetic writing.*)

Grammatical characteristics. The minimal unit of meaning seems to have been a verbal root, such as *zic* or *zich,* meaning "write." The suffixing of any vowel or certain consonants (*c* or its variant *ch, t* or its variant *th, l, r,* or *n*) produced a noun. The vowel *u* was used to form a gerund that, without further change, could be used as an agent noun; thus *zicu* meant "writing," then, further, "scribe," and then the equivalent of the Roman name Scribonius. From these verbal nouns, denominative verbs could be made; thus from *zicu* plus *-ce,* a past tense of perfective affix, was made *zichuche* "he wrote, he has written."

The *genitivus genitivi*

Because of the large number of names occurring in the inscriptions, the noun declension system can be understood reasonably well. Similar to the process of word building is the construction called *genitivus genitivi,* or "genitive of the genitive," in which several possessive suffixes may be added to a word in succession. Thus, the simple genitive of *larth,* a proper name, is *larthal* "Larth's"; a second genitive suffix added gives *larthalisa* "of that which is Larth's."

There is apparently no grammatical gender in Etruscan. In late Etruscan an *-i* suffix marks some women's names; still later, apparently (or possibly in a different dialect area), *-ia* is similarly used. Although these suffixes appear to be the same as the final elements in some words designating exclusively female functions, such as *puia* "wife" (in one occurrence spelled *pui,* if it is the same word) and *ati* "mother" (in one occurrence *atiu,* if it is the same word), there is no evidence for any syntactic use of gender, and there is no formal marker that can be shown to have marked gender consistently.

Case endings do not differ from singular to plural; in the singular they are suffixed directly to the word stem, and in the plural they are added to the stem, along with one of the plural markers *ar, er, ur.* There is no distinctive nominative (subject) case marker, the word stem or, in some cases, the root alone serving as the nominative. A final marker *-s,* however, does appear to have been added in some instances of a probable nominative case.

The repetitive nature of most Etruscan inscriptions is such that very few distinctively different verb forms are available for analysis. Indeed, probably the only really certain verbal suffix is *-ce.* It must not be assumed, however, that the paucity of the verbal data from inscriptions reflects an impoverished verb system in the language; indeed, judging from the variety of verbal stems to which the recurring *-ce* is added, it is more likely that the Etruscan verb had a more complicated structure than the noun.

Vocabulary. Since the language is undeciphered, meaning can be assigned with certainty to only a few Etruscan words that occur very frequently in the texts. Some kinship terms are sure—among these are *ati,* "mother," *clan* "son," *śec* "daughter," *puia* "wife." Less certain but probably correct are words designating members of the larger societal organization: *lavtn* "family," *zilc* "official," *maru* "official," *spur* "city," *rasna* or *raśna* "Etruscan, Etruria." A pair of dice certainly have on them the names of the numbers from one through six. Although the order of these numbers has been and still is disputed, the arrangement most generally accepted is this: *thu* "one," *zal* "two," *ci* "three," *śa* "four," *mach* "five," *huth* "six."

Among the continuing mysteries of Etruscan are the reasons why the Etruscans left no written records of their great civilization other than inscriptions and occasional texts and why the Romans, who knew the Etruscans intimately, transmitted little or nothing to posterity about either Etruscan literature or their language, which must certainly have been spoken, or at least preserved, by some families in Rome long after the period of Etruscan greatness had passed. (M.Fo.)

Basque language

Basque, the only remnant of the languages spoken in southwestern Europe before the region was romanized, is currently used in a narrow area of approximately 10,000 square kilometres (3,900 square miles) in Spain and France. The number of Basque-speaking persons outside that territory, in Europe and in the Americas, however, is far from insignificant. In Spain the Basque-speaking region comprises the province of Guipúzcoa, parts of Vizcaya and Navarra, and a corner of Álava, and in France the western region of the *département* of Pyrénées-Atlantiques. Although few statistics are available, the number of speakers, who are largely bilingual, might be judiciously estimated at 1,000,000. Most of them live in the highly industrialized Spanish part of the Basque country.

The Basques have derived their name, Euskaldunak, from Euskara, the native word for their language. According to the classification of the 19th-century philologist Prince Louis-Lucien Bonaparte, there are eight modern dialects of Basque. Dialectal division is not strong enough to mask the common origin or to preclude mutual understanding. Basque attained official status for a short period (1936–37) during the Spanish Civil War, under Basque autonomous government. In 1978, Basque and Castilian Spanish became the official languages of the autonomous Basque Country, which includes Guipúzcoa, Vizcaya, and Álava provinces of Spain.

Origins and classification. Basque remains an isolated language with no known linguistic relatives. The hypothesis of the German philologist Hugo Schuchardt (1842–1927), which once had wide currency, posited an intimate genetic connection between Basque and Iberian (see below) and the Hamito-Semitic (Afro-Asiatic) language group. This theory was superseded by attempts to establish a more or less close link between Basque and Caucasian, the language group indigenous to the Caucasus region. A lack of common linguistic characteristics between the Basque and Hamito-Semitic languages makes Schuchardt's hypothesis extremely dubious. There are, however, some common features that favour the relationship between Basque and Caucasian. Still, proof of a genetic relationship beyond reasonable doubt appears remote. Perhaps the most promising theory involves the comparison of Basque with the long-extinct Iberian, the language of the ancient inscriptions of eastern Spain and of the Mediterranean coast of France. But, despite amazing phonological coincidences, Basque has so far contributed next to nothing to the understanding of the now-readable Iberian texts. Therefore, it is possible that the similarity may have resulted from close contact between Basques and Iberians and not from a genetic linguistic relationship.

Link between Basque and Iberian

History of the language. At the beginning of the Christian Era, dialects of Euskarian (Basque) stock were probably spoken north and south of the Pyrenees and as far east as the Valle de Arán in northeastern Spain. It is likely that only the disruption of Roman administration in these regions saved the Basque dialects from being completely overcome by Latin. It is also likely that the Basque tongue, which had a firm foothold in the country that then began to be called Vasconia, experienced a substantial expansion toward the southwest, which carried it to the Rioja Alta (High Rioja) region in Old Castile and near Burgos. The more eastern Basque dialects, separated from the main area by Romance-speaking populations, were doomed. During the Middle Ages, Basque, the language of a population more peasant than urban, could not possibly hold the field as a written language against Latin and its successors, Navarrese Romance and, to a certain extent, Occitan (the *langue d'Oc,* also called Provençal) in the kingdom of Navarre. Since the 10th century, Basque has slowly but steadily lost ground to Castilian Spanish; in the north, however, where French is a more modern rival, the Basque-speaking area is practically the same as it was in the 16th century. In the last two centuries, above all in industrial centres, Basque has had to fight for survival in the heart of the Basque-speaking country, as well as on the frontier of the Basque-speaking area.

Basque records and writing

Latin inscriptions from the Roman period, found mostly in southwestern France, record a handful of proper names of unmistakable Basque etymology. From AD 1000 on, records consisting chiefly of proper names but also of Basque phrases and sentences grew more numerous and reliable. The first printed Basque book, dating from 1545, began an uninterrupted written tradition. Scholarly Basque literature, with its prevailing religious interests, has been neither abundant nor varied until recent times. Intense efforts are now being made to introduce Basque as a vehicle of private primary education. In addition, a model of a unified, standard written language also seems to be gaining increasing acceptance.

Phonology. The sound pattern of Basque is, on the whole, similar to that of Spanish. The number of distinctive sounds is relatively low compared with other languages. Combinations of sound (*e.g.,* consonant clusters) are subject to severe constraints. It can confidently be asserted that certain types of consonant clusters, such as *tr, pl, dr,* and *bl,* were all but unknown about two millennia ago. The common sound system underlying the systems of the present Basque dialects has five (pure) vowels and two series of stopped consonants—one voiced (without complete stoppage in many contexts), represented by *b, d, g,* and the other voiceless, represented by *p, t, k.* Nasal sounds include *m, n,* and palatal *ñ,* similar to the sound indicated by *ny* in "canyon." In this respect, as in others, Basque orthography coincides with the Spanish norm. There are two varieties of *l,* the common lateral *l* and a palatal variety, *ll,* as in Spanish, that sounds similar to the *lli* in "million" (as *l* + *y*). The Basque *r,* made by a single tap of the tongue against the roof of the mouth, contrasts with a rolled or trilled *r,* written *rr.* Two phonological features are worthy of special attention. Sibilants (both fricatives and affricates) made with the area of the tongue directly before the dorsum (the back of the tongue) are distinct from the apical sibilants, produced with the tip of the tongue. The letter *z* in Basque symbolizes the predorsal fricative, and *tz,* the predorsal affricate sound; *s* and *ts* represent the apical fricative (similar to Castillian Spanish *s*) and affricate, respectively. (A fricative is a sound, such as English *f* or *s,* produced with friction and, hence, without complete stoppage in the vocal tract; an affricate is a sound, such as *ch* in "church" or *j* in "jam," that begins as a stop and ends as a fricative, with incomplete stoppage.) In addition to these hissing sibilants, Basque also includes the hushing ones, written as *x* and *tx*; they are like the English *sh* and *ch.* The *x* and *tx* sounds, along with the palatal sounds written as *ll* and *ñ,* often have an expressive value (diminutive, endearing) in comparison with their nonpalatal counterparts; *e.g., hezur* means "bone" and *hexur* "little bone" (fish bone, for example); *sagu* is "mouse" and *xagu* "little mouse."

Varieties of sibilants

The phonology of some Basque dialects may be more complex than that presented in the preceding paragraph. In the easternmost Souletin region, for example, the dialect has acquired, by internal development or by contact with other languages, a sixth oral vowel—rounded *e* or *i*—and nasal vowels, voiced sibilants, and voiceless aspirated stops. The aspiration accompanying stop consonants consists of a small puff of air. There is also, word-initially and between vowels, an aspirated *h,* once common but now peculiar to the northern dialects. It has also been retained in the proposed standard form of Basque.

Grammar. The mention of two features is unavoidable in describing Basque syntax. Basque is, in the first place, a language of the so-called ergative type. That is, it has a case denoting the agent of an action. Hence, what in English would stand for the subject of a transitive verb is expressed in Basque by means of a suffix *-k*; for example, in the sentence "the foot serves the hand, and the hand serves the foot," *oinak zerbitzatzen du eskua, eta eskuak oina,* the first word, meaning "the foot," is composed of three elements, *oin* "foot," *-a,* "the," and *-k,* which marks the Basque equivalent of the subject of the verb. The fourth word, meaning "the hand," does not have the *-k* ending. In the second clause, *eta eskuak oina,* the word for hand, *eskuak,* now has the ergative *-k* ending to indicate that the hand is the agent of the clause "the hand serves the foot." The subject of an intransitive verb, which is not distinguished from the object of a transitive verb, has no overt mark—*e.g.,* in "if the belly does not eat, the belly itself will fail," *sabelak jaten ez ba du, sabela bera ihartuko da,* the first term, *sabelak* "the belly," has the *-k* marker because it is the agent of a transitive verb "eat"; but, in the second clause, *sabela* is the subject of the intransitive verb "fail" and, therefore, has no overt grammatical mark.

The second characteristic feature of Basque concerns the finite verb, which acts as a summary of all the noun phrases in the sentence. It has markers for all three persons—the 1st, 2nd, and 3rd—and may contain as many as three personal references (for subject, direct object, and indirect object). *Da,* for example, means "is," *du* means "he has it," and *dio* means "he has it for him" in the sentence *oinari ez dio eskuak kolperik emaiten* "the hand does not give a blow [*kolpe*] to the foot [*oin-a-ri*]." In certain situations the interlocutor can also be referred to within the verb. Further, most Basque verbs have only a compound conjugation; *e.g., erori da* "he has fallen," literally, "he is fallen," and *jaten du* "he eats [is eating] it."

Although some ancient prefixes are still apparent in modern Basque, they are no longer productive, so that Basque can be characterized as an over-all suffixing language; that is, it appends suffixes to words. There is one declension with suffixes or postpositions to indicate number and case; *e.g., etxe-a* "the house," but *etxe berri-a* "the new house," and *etxe berri-a-ri* "to [for] the new house." Suffixes, under certain restrictions, may be heaped upon one another. Theoretically, genitival endings indicating possession may be added to one another without limit. This is similar to the case in English of "the button of the coat of the son of the Major of York"; in Basque, however, the phrase "of the" is indicated by an ending, *-(r)en,* added to the noun. Noun suffixes can also be attached to verb forms in order to express subordination of the clauses in which the verb forms appear; *e.g., da* "is," *den* "which is," *dena* "that (*-a*) which is," *denean* "when there is," literally, "in that which is." Prefixes are also used for that purpose; *e.g., ez du jaten* "he does not eat" with the particle *ba* "if" becomes "if [the belly] does not eat," *jaten ez ba du.*

Vocabulary. Basque has preserved a peculiar and distinctive appearance, despite the overwhelming pressure to which it has been subjected over a period of at least 2,000 years. Nevertheless, its borrowings from the neighbouring languages, especially of words and idioms, can hardly be underrated. Loanwords from the Romance languages are numerous. Some of them bear the unmistakable stamp of their archaic Latin ancestry; *e.g., bake* "peace" from Latin *pax, pacis, bike* "pitch" from Látin *pix, picis,* and *errege* "king" from Latin *rex, regis.* Contrary to a widely held opinion, Indo-European loanwords of non-Latin origin are extremely scarce. Derivation, the formation of new words by the use of suffixes, is accomplished partly through the use of borrowed suffixes. This practice, as well as the compounding of nouns to form new words, as in *bizkar-hezur* "backbone," has been very much alive throughout the history of the language. On the other hand, Basque itself has contributed but little vocabulary to the Spanish, Occitan, French, and English languages. But family and place names of Basque coinage are frequently encountered in Spain and in Latin America, where they can be found in such proper names as Aramburu, Bolívar, Echeverría, and Guevara.

Latin loanwords

(L.M.)

PIDGIN

The term pidgin is applied to a number of varieties of speech that have grown out of English or other languages and that have been used in various parts of the world since the 17th century. Often termed "bastard jargons," "mongrel lingos," or the like, these tongues in fact are languages like any others and can be accurately delimited and described.

Definitions of lingua franca, pidgin, and creole. When a language is used as a means of communication between persons having no other language in common (*e.g.,* French in 18th-century diplomacy) it is a lingua franca. A lingua franca native to none of those using it and with a sharply reduced grammar and vocabulary is called a pidgin. (This definition of pidgin excludes both the broken English of a beginning learner and the skillful but non-native use of English in such countries as India.) When a whole speech community gives up its former language or languages and takes a pidgin as its mother tongue, the pidgin becomes a creole (is creolized).

Situations in which pidgins developed

Origins. A number of pidgins and creoles have arisen on the basis of various European languages. The first known pidgin, Lingua Franca, or "Westerners' language," of the medieval Levant and the Barbary Coast, was based chiefly on Italian. The American Indians first encountered by Englishmen in the 17th century were a tribe known as Pidians near the mouth of the Orinoco; the reduced language that emerged was termed Pidgin (= Pidian) English. Later in the same century, other varieties of Pidgin English grew up in China as a result of English commercial contacts and in Africa in connection with slave-trading activities. (Some authorities derive the word "pidgin" from a variation of English "business".) Establishment of plantation economies in the Caribbean area, with large groups of Negro slaves from different language backgrounds in West Africa, led to a number of pidgins based on English, French, Spanish, and Portuguese. Many have survived as creoles; *e.g.,* Gullah off the Sea Islands of South Carolina, the Negro English of the Antilles, and Sranantongo (formerly called Taki-Taki) in Suriname (formerly Dutch Guiana), all based on English; the French-based creoles of Louisiana, Haiti, and the Lesser Antilles; and the Papiamento of Curaçao, an outgrowth of Pidgin Spanish and Portuguese. Early contacts between settlers and natives led to the formation of pidgins in Australia and New Zealand, whereas the Pidgin English of the South Seas (called Beach-la-Mar) grew out of whaling, trading, and recruiting native labour. Pidgin English is extinct in New Zealand and the Caroline Islands and moribund in Australia but still flourishes in Melanesia (New Hebrides, Solomon Islands, New Guinea) and has become creolized in Hawaii. One variety, that of the Rabaul area (New Britain), has become the widespread and indispensable lingua franca of Papua New Guinea because of the official sanction given its use under German rule (1884–1914) and later Australian administrations.

Development and persistence of pidgins

Survival. Pidgin languages spring from the initial, non-intimate contacts between speakers of different languages, when quick comprehension is more highly valued than grammatical correctness or fine shades of meaning. As contacts grow closer, normally one group learns the other's language more fully, and pidgins survive the stages of initial contact only in special circumstances. Pidgins persist where a dominant group regards another as childlike or capable only of a simplified version of the "superior" language, as in the relations between Europeans and American Indians, West Africans, or South Sea natives. On plantations and in other situations where European masters were in permanent contact with native servants or labourers, pidgins served as status languages, as in New Guinea. Caste distinction, however, is not a necessary function of a pidgin; Russonorsk, for instance, was a reduced language used by Russians and Norwegians in the Arctic at the beginning of the 20th century. Chinese Pidgin English survived for three centuries and not only in master–servant relations; it also was in use between English merchants and Chinese dignitaries, primarily because each side desired to keep the other at arm's length. Slaves on Caribbean plantations, New Guinea natives in newly founded multilingual villages, and others who have come to live together with no language in common save a pidgin have used the pidgin as the customary language of the group. In such instances, the resultant creole has usually re-expanded its structure and vocabulary by borrowing from the language of a culturally dominant group; *e.g.,* Haitian Creole from French, Sranantongo and Papiamento from Dutch, and Melanesian Pidgin from English.

Spelling and sounds

Orthographies. In its original function as a lingua franca among unlettered folk, a pidgin language is a medium of purely oral communication, as also are creoles in their initial stages. Only afterward, and usually in connection with missionary or other educational programs, are spelling systems devised for pidgins or creoles. Speakers of European languages have often applied the orthographical conventions of their own languages, as when the Melanesian Pidgin sentence for "Why did you hit this policeman?" is written *Belong what name you fight 'im dis fellow police boy?* Such a spelling embodies all the inconsistencies of English orthography and is therefore difficult to learn; it distorts the structure of pidgin; and it confirms the naïve European or American in his belief that pidgin is only a ridiculous reduction of English. Those who have devised orthographies, for ease of learning, accuracy in representing linguistic structure, and emphasis on the independent status of the language, have used phonemically based systems. The most effective orthographies of this type use the letters available on typewriters or in printshops, but consistently and predictably. Thus, the Melanesian Pidgin sentence just quoted reads, in the officially recognized orthography: *Bilong wonem yu faitim dispela plisboi?* In the following discussion, forms are cited only in this type of transcription.

Phonology. The simplification which characterizes pidgin extends to all aspects of linguistic structure (sounds, forms, constructions) as well as vocabulary. In some varieties stress is automatically on the first syllable; *e.g., bíkos* "because," *míshin* "machine." A minimum of five distinctive vowels is necessary, those represented in Latin or Italian pronunciations by *a* ("ah"), *e* ("eh"), *i* ("ee"), *o* ("oh"), *u* ("oo"), as in *antap* "up," *em* "he," *winim* "defeat," *kot* "coat," *tu* "also." The vowel sound *a* ("ah") combines with *i* ("ee") and *u* ("oo") in the basic diphthongs *ai* (English "i") and *au* ("ow"), as in *dai* "cease" and *nau* "now." Some, but not all, speakers make further distinctions—*e.g.,* between the *e* of *em* "he" and the *ei* ("ay") of *neim* "name." In almost all varieties of pidgin English, the two consonant sounds represented by English *th* have merged with *t* and *d* respectively: *saut* "south," *dispela* "this." Many speakers of Melanesian Pidgin merge *f* and *p* (in current official orthography, both are represented by *p*) and also merge *ch* and *sh* with *s*: *tumas* "very" (from English "too much"), *masin* "machine." Users of pidgin often carry over habits of sound production from their native languages; *e.g.,* many Melanesian languages have *mb, nd* as variants of *b, d* between vowels, and thus Melanesians often pronounce *tabak* "tobacco" as *tambak,* and *sidaun* "sit" as *sindaun.*

Morphology. Grammatical categories—such as number, gender, case, person, tense, mood, voice—are almost absent from pidgin and creole languages, as from many other languages of the world. Pidgin is not, however, "devoid of grammar," as is often asserted. Melanesian Pidgin, for example, has three inflected parts of speech: pronouns, adjectives, and verbs. The suffix *-pela* added to pronouns makes plurals, in *mi* ("I, me") versus *mipela* ("we, us") and *yu* ("you" singular) versus *yupela* ("you" plural). Another suffix, *-pela,* serves as a marker for adjectives of one syllable, demonstratives, indefinites, and numerals: *naispela* "pretty," *dispela* "this," *sampela* "some," *wanpela* "one." Verbs having a direct object (expressed or implied) take the suffix *-im,* and verbs without this suffix are intransitive or passive: *e.g., rausim* "eject, remove" versus *raus* "be out, come out." Other parts of speech—nouns, prepositions, adverbs, conjunctions—are invariable but are distinguished by the types of combinations in which they occur. In other varieties of Pidgin English, the specific criteria for distinguishing classes of linguistic forms are different, but the basic structure is similar. Chinese Pidgin nouns and pronouns, for example, can take the locative suffix *-said,* in *doksaid* "at the dock (docks)," *maisaid* "at my house," and the verb suffix *-em* forms passive participles, as in *dis tri blong spoilem* "this tree is rotten."

Phrases and clauses

Syntax. The basic types of combination—phrases and clauses—found in pidgin are the same as those of English; here again, however, many details of syntax are different. Melanesian Pidgin nouns are followed not only by possessive phrases (*haus bilong mi* "house of me, my house") but also by modifying nouns (*haus pepa* "house [for] paper, office"), verbs (*haus kuk* "house [for] cooking, kitchen"), adverbs (*man nogud* "evil man," *nogud* being an adverb meaning "undesirably"), and clauses (*man mi lukim em* "the man I saw"). In Chinese Pidgin, pronouns simply precede nouns to indicate possession (*hi fes* "his face"), and relative position is shown by a noun preceding an adjective (*Ning-Po mo fa* "further than Ning Po"). With third person subjects, Melanesian Pidgin predicates are normally preceded by the predicate-marker *i-*: *ol i-singaut* "they call"; *balus i-no kamap yet* "the plane hasn't arrived yet." Melanesian Pidgin has the clause type called "equational" (also found in such languages as Russian and Hungarian) in which no verb is present: *dispela kaikai i-gudpela* "this food [is] good." Chinese Pidgin, on the other hand, has the copulative, or linking, verb *blong* "be" with nouns and adjectives in the predicate (*e.g., yu fut blong plenti sor* "your foot is very sore") but uses no verb with an adverb in the predicate indicating location: *tumuchi dast tebal tapsaid* "a lot of dust [is] on the table."

Vocabulary

Lexicon. Since vocabulary is restricted (about 700 words in Chinese Pidgin, 2,000 in Melanesian), each word necessarily has a greater range of meaning than its English counterpart. The central meaning of Melanesian Pidgin *sori* is not "sorry" but "emotionally moved," as shown by its extension to "sympathetic, grateful, glad"; similarly, *dai* means "cease" ("die" is *dai tru* "stop for good"); and *stap* is "be located; remain; continue." For many concepts, pidgin uses phrases rather than single words: with *skru* "screw; joint" are formed *skru bilong arm* "elbow," *skru bilong leg* "knee," etc. Some pidgin words represent different parts of speech from their English counterparts; *e.g.,* the Melanesian Pidgin preposition *belong* "of" and the Chinese Pidgin copulative verb *blong* "be" from English "belong." Non-English meanings of pidgin words often reflect native social structure, as when *papa* means "uncle," since a boy's maternal uncle rather than his father (*papa tru*) is primarily responsible for his upbringing in New Guinea. Speakers of English are often naïvely amused or shocked by certain shifts of meaning, as when *ars* "buttocks" is extended to mean "bottom (of anything), foundation, reason, cause, source"; for example, *ars bilong diwai* "the base of the tree," or *God i-ars bilong olgeda samting* "God is the source of all things." In the context of native society and attitudes, however, these concepts are not taboo and no stigma attaches to the words or their use.

Non-English vocabulary. The proportion of vocabulary elements in Pidgin English derived from non-English sources is small. Of the approximately 2,000 words in Melanesian Pidgin, not over 10 percent are of non-English origin. Of these, perhaps half are Melanesian (such as *kiau* "eggbomb," *diwai* "tree," *malolo* "rest," *balus* "pigeon, airplane"), and one quarter were borrowed from German, such as *mark* "shilling," *tais* "pond," *langsam* "slow," and *beten* "prayer." The remainder are from various languages—a few from Malay (such as *karabau* "water buffalo") and three from Romance sources: *save* "know," *pikinini* "child," and *pato* "duck." The percentage of non-English elements in Chinese Pidgin is even smaller.

Structural changes from English

Restructuring. In the structural reduction from English to pidgin, the main grammatical characteristics have been kept (the part-of-speech system, the dichotomy between subject and predicate, the use of phrases functioning as single parts of speech), though often with different identifying features. The various kinds of Pidgin English are definitely English and Indo-European, not (as is often said) "native languages spoken with English words." A more sophisticated version of this latter theory is that indigenous vocabulary is simply replaced with new words by "relexification," with indigenous grammatical habits continuing. However, when new functional elements (*e.g.,* pronouns, inflectional suffixes, syntactic patterns) are also taken over, the process is one of complete language-substitution, involving replacement of grammatical structure ("regrammaticalization") as well. Nevertheless there have been extensive carryovers from non-English structural patterns because speakers of native languages have translated their own constructions into pidgin, especially in the early stages of its formation. For instance, in Chinese a numeral modifying a noun must be accompanied by a special word indicating a measure of quantity, called a "numeral-classifier," as in *sān ge rén,* literally "three piece man." This type of combination was reflected in older Chinese Pidgin *tufela man* "two men," *forpisi naif* "four knives," etc., with *-fela* for animate objects, *-pisi* for inanimate objects. However, in modern Chinese Pidgin, *-fela* has not survived, and *-pisi* has lost its independent status and become simply a numeral-suffix, as in *tupisi man* "two men." Similarly, the presence of separate pronouns for "we (excluding the hearer)" and "we (including the hearer)" in Melanesian languages has led to the establishment in Melanesian Pidgin of a parallel contrast between *mipela* "we (but not you)" and *yumi* "we (including you)."

Modern function. With the coming of modern civilization and technology to New Guinea and similar areas, pidgin has proved indispensable in education and political life. Earlier opposition to pidgin—partly on puristic, partly on anticolonialistic grounds—has proved unfounded. In New Guinea and the Solomons, and in many parts of West Africa, pidgin is no longer a status language or imposed on the people by white colonialists, but is the people's own lingua franca, indispensable for communication and easier to learn than English, which is both more complicated and more foreign to them. If skillfully used, pidgin can serve as both a medium of instruction and a bridge to English. In any case, it is clearly destined to remain as an increasingly useful lingua franca, with already manifest prospects of extensive creolization and resultant permanence as the native language of ever larger groups. (R.A.H.)

BIBLIOGRAPHY

Languages of the world. E.H. STURTEVANT, *Linguistic Change* (1917, reprinted 1961), was an important text in Indo-European comparative linguistics, until replaced by WINFRED P. LEHMANN, *Historical Linguistics: An Introduction* (1962). EDWARD SAPIR, *Language: An Introduction to the Study of Speech* (1921, reprinted 1957), presents a revision of the 19th-century whole-language typology. For an example of modern application of subsystem typology, see A.K. RAMANUJAN and COLIN MASICA, "Toward a Phonological Typology of the Indian Linguistic Area," in *Current Trends in Linguistics,* vol. 5, pp. 543–577 (1969); and for an example of lexical domain typology, see BRENT BERLIN and PAUL KAY, *Basic Color Terms: Their Universality and Evolution* (1969). ANTOINE MEILLET and MARCEL COHEN (eds.), *Les Langues du monde,* new ed. (1952), is a survey of languages of the world organized in terms of their genetic classification. Languages are also classified genetically in two more recent surveys: C.F. and F.M. VOEGELIN, "Languages of the World," in *Anthropological Linguistics,* vol. 6–8 (1964–66); and THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* 13 vol. (1963–75)—individual volumes deal with geographic areas, and typological information on particular languages is presented. JOSEPH H. GREENBERG, "The Indo-Pacific Hypothesis" (pp. 807–871), and S.A. WURM, "The Papuan Linguistic Situation" (pp. 541–657), both in *Current Trends in Linguistics,* vol. 8 (1971), exemplify variation in proposals for remote genetic relationships among the same set of languages. EDWARD SAPIR, "Central and North American Languages," in *Encyclopædia Britannica,* 14th ed. (1929), presented a classification of these languages in terms of six phyla; this classification was revised by a conference of specialists that resulted in the publication of C.F. and F.M. VOEGELIN (comps.), *Map of North American Indian Languages,* rev. ed. (1967). The culture areas of North American Indians are described in HAROLD E. DRIVER, *Indians of North America,* 2nd ed. rev. (1969).

Indo-European languages. The introductory, classic, and most comprehensive works on Indo-European languages are in German and French. Several publications, however, are written in English.

KARL BRUGMANN, *Grundriss der vergleichenden Grammatik der indogermanischen Sprachen,* 2nd ed., 3 vol. (1897–1916), the latest completed full treatment of the whole family; ANTOINE MEILLET, *Introduction à l'étude comparative des langues indo-européennes,* 8th ed. (1937, reprinted 1964), the best introduction to the subject; JERZY KURYLOWICZ (ed.), *Indogermanische Grammatik* (1968–), vol. 2, *Akzent, Ablaut,* by JERZY KURYLOWICZ (1968), and vol. 3, pt. 1, *Geschichte der indogermanischen Verbalflexion,* by CALVERT WATKINS (1969), when completed, this work will present an account of the entire family—less detailed than Brugmann's, but much more up-to-date; JULIUS POKORNY, *Indogermanisches etymologisches Wörterbuch,* 2 vol. (1951–69), the most recent etymological dictionary of the whole family; CARL DARLING BUCK, *A Dictionary of Selected Synonyms in the Principal Indo-European Languages* (1949), a mine of information about Indo-European words for several hundred basic concepts; HOLGER PEDERSEN, *Sprogvidenskaben i det nittende aarhundrede* (1924; Eng. trans., *Linguistic Science in the Nineteenth Century,* 1931; reissued as *The Discovery of Language,* 1962), a very good account of 19th-century work in the field; GEORGE CARDONA, HENRY M. HOENIGSWALD, and ALFRED SENN (eds.), *Indo-European and Indo-Europeans* (1970), a collection of recent papers on aspects of Indo-European language, culture, and mythology, especially valuable for the attempt to combine linguistic and archaeological evidence about Indo-European prehistory; FREDRIK OTTO LINDEMAN, *Einführung in die Laryngaltheorie* (1970), an excellent brief account of the main advance in Indo-European phonology since 1900.

Anatolian languages. The best general handbook on Anatolian history and civilization, in German, is A. GOETZE, *Kleinasien,* 2nd ed. (1957), with succinct chapters on thé Anatolian languages (pp. 45–63). An excellent, more popular account of Anatolia with marked emphasis on the Hittite period is O.R. GURNEY, *The Hittites,* rev. ed. (1961). This book deals with languages and races in ch. 6. The general handbook on the Anatolian languages, both Indo-European and non-Indo-European, written by scholars, is A. KAMMENHUBER *et al., Altkleinasiatische Sprachen* (1969), although in some chapters the book is perhaps too detailed for the general reader. J. FRIEDRICH (ed.), *Kleinasiatische Sprachdenkmäler* (1932), is still a very valuable collection of texts, chosen to illustrate the various languages of Anatolia. Although written from the viewpoint of the Indo-Hittite hypothesis, E.H. STURTEVANT and E.A. HAHN, *A Comparative Grammar of the Hittite Language,* rev. ed. (1951), was a major achievement. In many respects it is still the best grammar of Hittite. H. PEDERSEN, *Hittitisch und die anderen Indoeuropäischen Sprachen* (1938), was directed against the Indo-Hittite hypothesis and should be used with Sturtevant and Hahn's book. J. PUHVEL's contribution "Dialectal Aspects of the Anatolian Branch of Indo-European," in H. BIRNBAUM and J. PUHVEL (eds.), *Ancient Indo-European Dialects,* pp. 235–247 (1966), gives a good survey of the main characteristics of the Anatolian subgroup. The best recent introduction to the "Indo-European problem" with a notable emphasis on Anatolian matters is R.A. CROSSLAND, *Immigrants from the North,* fasc. 60 of the *Cambridge Ancient History,* rev. ed. (1967). The best introduction to the Anatolian languages and to Hittite literature may be found in two highly authoritative and very well written scientific contributions by H.G. GUTERBOCK: "Toward a Definition of the Term Hittite," *Oriens,* 10:233–239 (1957) and "A View of Hittite Literature," *Journal of the American Oriental Society,* 84:107–115 (1964). The history of the various decipherments has been dealt with by J. FRIEDRICH, *Entzifferung verschollener Schriften und Sprachen,* 2nd ed. (1966).

Indo-Iranian languages. *Indo-Aryan languages (general works):* JULES BLOCH, *L'Indo-aryen du veda aux temps modernes* (1934; rev. Eng. trans., *Indo-Aryan from the Vedas to Modern Times,* 1965), a masterly survey of Indo-Aryan throughout its history; R.L. TURNER, *A Comparative Dictionary of the Indo-Aryan Languages* (1966), an indispensable source in which Sanskrit word headings are given Middle Indo-Aryan forms and New Indo-Aryan cognates; M.B. EMENEAU, "The Dialects of Old Indo-Aryan," in HENRIK BIRNBAUM and JAAN PUHVEL (eds.), *Ancient Indo-European Dialects,* pp. 123–138 (1966), a good summary, with discussion of proposed theories and references; SURYAKANTA, *A Practical Vedic Dictionary* (1981).

Old Indo-Aryan: THOMAS BURROW, *The Sanskrit Language,* new and rev. ed. (1973), a summary of the prehistory and history of Sanskrit, with references to Middle Indo-Aryan, which contains somewhat personal views but is valuable for its discussion of non-Aryan influences on Sanskrit; LOUIS RENOU, *Histoire de la langue sanskrite* (1956), an insightful summary of the grammar, vocabulary, and style of different stages of Sanskrit, with text selections and translations; MANFRED MAYRHOFER, *Kurzgefasstes etymologisches Wörterbuch des Altindischen* (*A Concise Etymological Sanskrit Dictionary*), 4 vol. (1953–80), contains sober etymologies, full references, and a discussion of loanwords and words supposed to have been borrowed from Dravidian.

Middle Indo-Aryan: RICHARD PISCHEL, *Grammatik der Prākrit-Sprachen* (1900; Eng. trans., *Comparative Grammar of the Prākrit Languages,* 2nd ed., 1965), an encyclopaedic grammar of all the Prākrits except Buddhist Hybrid Sanskrit and Pāli, which includes a good discussion of the different Prākrits in the introduction (now in need of updating); S.M. KATRE, *Prakrit Languages and Their Contribution to Indian Culture,* 2nd ed. (1964), a general survey of the Prākrits, including Pāli; LUDWIG ALSDORF, *Apabhraṃśa-Studien,* pp. 5–17, 20–37 (1937, reprinted 1966, important studies discussing noun and verb inflection.

Modern Indo-Aryan: JOHN BEAMES, *A Comparative Grammar of the Modern Aryan Languages: To Wit, Hindi, Panjabi, Sindhi, Gujarati, Marathi, Oṛiya,* and *Bangali,* 3 vol. (1872–79, reprinted 1966); and A.F.R. HOERNLE, *A Comparative Grammar of the Gaudian Languages with Special Reference to the Eastern Hindi* (1880, reprinted 1975), general comparative grammars of the New Indo-Aryan languages—though in need of modernization, still indispensable; SIR GEORGE A. GRIERSON, *On the Modern Indo-Aryan Vernaculars* (1931), a reprint of two long articles, tracing the phonologic developments that led to New Indo-Aryan; S.K. CHATTERJEE, *Indo-Aryan and Hindi,* 2nd rev.

ed. (1960), a series of lectures briefly tracing the history of Indo-Aryan, with particular emphasis being given to Hindi and its relation to other Indo-Aryan languages, and to the general language problem in India.

Iranian languages (general works): An important comprehensive treatment of the Iranian languages in general is WILHELM GEIGER and ERNST KUHN (eds.), *Grundriss der iranischen Philologie,* 2 vol. (1895–1904, reprinted 1974). This invaluable work is now in many respects antiquated, and it contains no account of several Middle Iranian languages that have been made known only in this century. A more recent account in less detail is provided by the *Handbuch der Orientalistik,* vol. 1, sect. 4, *Iranistik,* pt. 1, *Linguistik* (1958). There is an introduction to the subject in Russian: ИОСИФ МИХАЙЛОВИЧ ОРАНСКИЙ, *Введение в иранскую филологию* (1960). Some useful bibliography with brief guidelines is given by D.N. MACKENZIE, "Iranian Languages," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 5, *Linguistics in South Asia,* pp. 450–477 (1969).

Old Iranian: All known Old Persian texts except for recent discoveries are given in transcription and translation in ROLAND KENT, *Old Persian,* 2nd ed. rev. (1953). Information on Old Persian linguistic problems is contained in WILHELM BRANDENSTEIN and MANFRED MAYRHOFER, *Handbuch des Altpersischen* (1964). On the Avestan language the article by GEORG MORGENSTIERNE, "Orthography and Sound-System of the Avesta," in *Norsk Tidsskrift for Sprogvidenskap,* 12:30–82 (1942), is of great importance. A useful bibliographical guide to recent work on Avestan is provided by J. DUCHESNE-GUILLEMIN, "L'Étude de l'iranien ancien au vingtième siècle," *Kratylos,* 7:1–44 (1962).

Middle Iranian: The verbal system of Parthian is described by A. GHILAIN, *Essai sur la langue parthe, son système verbal d'après les textes manichéens* (1939, reprinted 1966); and for Middle Persian by W. HENNING, "Das Verbum des Mittelpersischen der Turfanfragmente," *Zeitschrift für Indologie und Iranistik,* 9:158–253 (1933). A Pahlavi dictionary for students is D.N. MACKENZIE, *A Concise Pahlavi Dictionary* (1971). The grammar of Sogdian has received detailed treatment in ILYA GERSHEVITCH, *A Grammar of Manichean Sogdian* (1954); and of Khotanese in R.E. EMMERICK, *Saka Grammatical Studies* (1968). A brief sketch of Khwārezmian is given in W.B. HENNING, "The Khwarezmian Language," *Zeki Velīdi Togan'a Armağan,* pp. 421–436 (1955).

Modern Iranian: An important recent work is GILBERT LAZARD, *Grammaire du persan contemporain* (1957). For the early stages of Modern Persian, Lazard's *Langue des plus anciens monuments de la prose persane* (1963), is invaluable. The same author has provided a comprehensive guide to the most important linguistic features of Tazhik in "Caractères distinctifs de la langue tadjik," *Bulletin de la Société linguistique de Paris,* 52:117–186 (1956). A treatment of Baluchi dialects is J.H. ELFENBEIN, *The Baluchi Language* (1966). Of the many works describing Kurdish dialects, a comprehensive modern work is D.N. MACKENZIE, *Kurdish Dialect Studies,* 2 vol. (1961–62). For the history of the Pashto language, GEORG MORGENSTIERNE, *An Etymological Vocabulary of Pashto* (1927), remains standard. Morgenstierne's *Indo-Iranian Frontier Languages,* 4 vol, in 6, 2nd. rev. and with new material (1973), is the work most often quoted for most of the minor languages. Ossetic has been described by a native speaker: V.I. ABAEV, *A Grammatical Sketch of Ossetic* (1964; orig. pub. in Russian, 1959).

Greek language. *Ancient Greek:* M. VENTRIS and J. CHADWICK, *Documents in Mycenaean Greek: Three Hundred Selected Tablets from Knossos, Pylos, and Mycenae, with Commentary and Vocabulary* (1956), a study of both the writing system and the content of the tablets, by the authors of the decipherment; L.H. JEFFERY, *The Local Scripts of Archaic Greece* (1961), a description of all the local varieties of the Greek alphabet from the 8th to the 5th century BC; C.D. BUCK, *The Greek Dialects* (1955), a summary of the dialectal features of Ancient Greek within the scope of a traditional descriptive grammar; A. MEILLET, *Aperçu d'une histoire de la langue grecque,* 7th ed. (1965), the first and still fundamental endeavour to define the characteristics of Greek in a diachronic perspective; E. SCHWYZER and A. DEBRUNNER, *Griechische Grammatik,* 3 vol. (1939–53), a complete and accurate description with exhaustive bibliography; P. CHANTRAINE, *La Formation des noms en grec ancien* (1933), deals with the history of noun suffixes throughout the history of Greek; H. FRISK, *Griechisches etymologisches Wörterbuch,* 2 vol. (1954–70), up to date and wisely selective (but often underrating Mycenaean data).

Koine and Byzantine: For Koine, see F. BLASS and A. DEBRUNNER, *Grammatik des neutestamentlichen Griechisch,* 10th ed. (1959; Eng. trans., *A Greek Grammar of the New Testament and Other Early Christian Literature,* 1961), a translation and revision of a classic work by R.W. FUNK. T.M. DAWKINS, "The Greek Language in the Byzantine Period," in N.H. BAYNES and H.ST.L.B. MOSS, *Byzantium* (1948); and R. BROWNING, *Medieval and Modern Greek* (1969), cover later periods.

Modern Greek: S.A. SOFRONIOU, *Teach Yourself Modern Greek* (1962); and J.T. PRING, *A Grammar of Modern Greek on a Phonetic Basis* (1950), are good elementary introductions. There is no reasonably complete dictionary of literary demotic, but J.T. PRING, *The Oxford Dictionary of Modern Greek (Greek-English)* (1965), is adequate for the spoken language. In general, English–Greek dictionaries are intended for Greeks and fail to mark the stylistic level of Greek equivalents; A.N. JANNARIS, *A Concise Dictionary of the English and Modern Greek Languages* (1895, frequently reprinted), has not been improved on. F.W. HOUSEHOLDER, K. KAZAZIS, and A. KOUTSOUDAS, *Reference Grammar of Literary Dhimotiki* (1964), is useful; and for a convenient summary of the development of demotic, see R. Browning (cited above). For dialects, see B.E. NEWTON, *The Generative Interpretation of Dialect: A Study of Modern Greek Phonology* (1972).

Italic languages: General surveys of the history of Italic languages include: J. WHATMOUGH, *The Foundations of Roman Italy* (1937); G. DEVOTO, *Gli antichi Italici,* 3rd ed. (1967), in Italian; and E. PULGRAM, *The Tongues of Italy* (1958).

The fundamental works on Osco-Umbrian are R. VON PLANTA, *Grammatik der oskisch-umbrischen Dialekte,* 2 vol. (1892–97), in German; and R.S. CONWAY (ed.), *Italic Dialects,* 2 vol. (1897). The most complete edition of texts (with the exception of Venetic) is E. VETTER, *Handbuch der italischen Dialekte,* vol. 1, *Texte mit Erklärung, Glossen, Wörterverzeichnis* (1953), in German. The best introduction for beginners continues to be C.D. BUCK, *A Grammar of Oscan and Umbrian,* 2nd ed. (1928). For special studies of the Iguvine Tables, see G. DEVOTO, *Tabulae Iguvinae,* 3rd ed. (1962), in Latin; and J.W. POULTNEY, *The Bronze Tables of Iguvium* (1959). For Faliscan, see G. GIACOMELLI, *La lingua falisca* (1963); and for Venetic, see G.B. PELLEGRINI and A.L. PROSDOCIMI, *La lingua venetica,* 2 vol. (1967), both works in Italian.

Remains of other languages within ancient Italy, including Venetic, are treated in R.S. CONWAY, J. WHATMOUGH, and S.E. JOHNSON, *The Prae-Italic Dialects of Italy,* 3 vol. (1934).

Romance languages: I. IORDAN and J. ORR, *An Introduction to Romance Linguistics,* revised reprint of the 1937 edition, with an additional essay, "Thirty Years On" by REBECCA POSNER (1970), describes the work done in Romance during the 19th and 20th centuries and provides an extensive bibliography of all works written both on the family as a whole, and on individual languages. On the Romance languages in general, two books written in English will prove useful: the more philologically oriented *The Romance Languages* by W.D. ELCOCK (1960); and the more popular, linguistically oriented *The Romance Languages: A Linguistic Introduction* by REBECCA POSNER (1966). Outstanding is Y. MALKIEL, *Essays on Linguistic Themes* (1968), which includes several articles on the Romance languages and Romance linguistics. On the individual languages, among works in English particularly to be recommended, are L.R. PALMER, *The Latin Language* (1954); BRUNO MIGLIORINI, with T. GWYNFOR GRIFFITH, *Storia della lingua italiana* (1960, Eng. trans., *The Italian Language,* 1966); A.E. EWERT, *The French Language,* 2nd ed. (1961); and WILLIAM ENTWISTLE, *The Spanish Language* (1936).

Germanic languages. *History and classification of the Germanic languages:* The reconstruction of Proto-Germanic and the derivation of Proto-Germanic from Proto-Indo-European are treated in FRANS VAN COETSEM (ed.), *Toward a Grammar of Proto-Germanic* (1972); EDUARD PROKOSCH, *A Comparative Germanic Grammar* (1939); HERMANN HIRT, *Handbuch des Urgermanischen,* 3 vol. (1931–34); ANTOINE MEILLET, *Caractères généraux des langues germaniques,* 7th ed. (1949); WILHELM STREITBERG, *Urgermanische Grammatik* (1896, reprinted 1963). See also ANATOLY LIBERMAN, *Germanic Accentology,* vol. 1 (1982).

East Germanic and Gothic: (*General survey*): JAMES W. MARCHAND, "The Gothic Language," *Orbis,* 7:492–515 (1958). (*Texts*): WILHELM STREITBERG (ed.), *Die gotische Bibel,* 3rd ed. (1950). (*Grammar*): JOSEPH WRIGHT, *Grammar of the Gothic Language . . . ,* 2nd ed. (1954); FERNAND MOSSE, *Manuel de la langue gotique,* 2nd ed. (1956); THEODOR WILHELM BRAUNE, *Gotische Grammatik,* 16th ed. (1961); WOLFGANG KRAUSE, *Handbuch des Gotischen* (1953). (*Dictionary and concordance*): ERNST SCHULZE, *Gothisches Glossar* (1848). (*Etymological dictionary*): SIGMUND FEIST, *Vergleichendes Wörterbuch der gotischen Sprache,* 3rd ed. (1939). (*Ostrogothic*): FERDINAND WREDE, *Über die Sprache der Ostgoten in Italien* (1891). (*Vandalic*): FERDINAND WREDE, *Über die Sprache der Wandalen* (1886). (*Bibliography*): FERNAND MOSSE, "Bibliographia Gotica," *Mediaeval Studies,* 12:237–324 (1950), supplements in 15:169–183 (1953), and 19:174–196 (1957).

Frisian: (*General survey*): BO SJOLIN, *Einführung in das*

Friesische (1969). (*Old Frisian, grammar and phonology*): WALTHER STELLER, *Abriss der altfriesischen Grammatik* (1928). (*Dictionary*): FERDINAND HOLTHAUSEN, *Altfriesisches Wörterbuch* (1925). (*Modern West Frisian, grammar*): K. FOKKEMA, *Beknopte Friese Spraakkunst*, 2nd ed. (1967). (*Phonology*): ANTONIE COHEN *et al., Fonologie van het Nederlands en het Fries*, 2nd ed. (1961). (*Dictionary*): H.S. BUWALDA, G.A.G. MEERBURG, and Y.R. POORTINGA (eds.), *Frysk Wurdboek*, 2 vol. (1952–56), Netherlandic-Frisian, Frisian-Netherlandic. (*Dialects*): NILS ARHAMMER, "Friesische Dialektologie," in *Germanische Dialektologie*, vol. 1, pp. 264–317 (1968).

Netherlandic: (*General*): WALTER LAGERWEY, *Guide to Dutch Studies: Bibliography of Textual Materials for the Study of Dutch Language, Literature, Civilization* (1961); C.B. VAN HAERINGEN, *Netherlandic Language Research*, 2nd ed. (1960). (*History*): MORITZ SCHONFELD, *Historische Grammatica van het Nederlands*, 6th ed. (1960); C.G.N. DE VOOYS, *Geschiedenis van de Nederlandse Taal*, 5th ed. (1952). (*Old Low Franconian*): ROBERT L. KYES (ed.), *The Old Low Franconian Psalms and Glosses* (1969). (*Phonology*): ANTONIE COHEN *et al., Fonologie van het Nederlands en het Fries*, 2nd ed. (1961); B. VAN DEN BERG, *Foniek van het Nederlands*, 2nd ed. (1960); EDGAR BLANCQUAERT, *Practische Uitspraakleer van de Nederlandse Taal*, 4th ed. (1953). (*Grammar*): ETSKO KRUISINGA, *A Grammar of Modern Dutch* (1924). (*Dialects*): A.A. WEIJNEN, *Nederlandse Dialectkunde* (1966).

Afrikaans: (*History*): G.G. KLOEKE, *Herkomst en Groei van het Afrikaans* (1950). (*Phonology*): MEYER DE VILLIERS, *Afrikaanse Klankleer: Fonetiek, Fonologie en Woordbou*, 3rd ed. (1965). (*Grammar*): H.J.J.M. VAN DER MERWE, *An Introduction to Afrikaans* (1952). (*Dictionary*): D.B. BOSMAN, I.W. VAN DER MERWE, and L.W. HIEMSTRA, *Tweetalige Woordeboek*, vol. 1, *Afrikaans-Engels* and vol. 2, *Engels-Afrikaans*, 5th ed. (1964; 7th ed., 1 vol., 1967). (*Comparison with Netherlandic*): JAMES L. WILSON, *Some Phonological, Morphological and Syntactic Correspondences Between Standard Dutch and Afrikaans* (1967; available from University Microfilms).

German: (*History*): W.B. LOCKWOOD, *An Informal History of the German Language, with Chapters on Dutch and Afrikaans, Frisian and Yiddish* (1965); JOHN T. WATERMAN, *A History of the German Language* (1966); ADOLF BACH, *Geschichte der deutschen Sprache*, 8th ed. (1965). (*Pronunciation*): WILLIAM G. MOULTON, *The Sounds of English and German* (1962); THEODOR SIEBS, *Deutsche Aussprache: Bühnenaussprache*, 19th ed. (1969); MAX MANGOLD (ed.), *Duden Aussprachewörterbuch* (1962). (*Spelling*): *Duden: Rechtschreibung der deutschen Sprache und der Fremdwörter*, 16th rev. ed. (1967). (*Grammar*): HERBERT LEDERER, *Reference Grammar of the German Language* (1969), trans. and adapted from HEINZ GRIESBACH and DORA SCHULZ, *Grammatik der deutschen Sprache*, 6th ed. (1967); HERBERT L. KUFNER, *The Grammatical Structures of English and German* (1962); PAUL GREBE (ed.), *Duden Grammatik der deutschen Gegenwartssprache*, 2nd ed. (1966). (*Dialects*): R.E. KELLER, *German Dialects* (1961); ADOLF BACH, *Deutsche Mundartforschung*, 2nd ed. (1950); V.M. SCHIRMUNSKI, *Deutsche Mundartkunde* (1962; orig. pub. in Russian, 1956).

Yiddish: The main works up to 1958 are listed in URIEL and BEATRICE WEINREICH, *Yiddish Language and Folklore: A Selective Bibliography for Research* (1959). Significant recent works include: *For Max Weinreich on His Seventieth Birthday: Studies in Jewish Languages, Literature, and Society* (1964); *The Field of Yiddish: Studies in Language, Folkore and Literature*, 2nd collection, ed. by URIEL WEINREICH (1965); *The Field of Yiddish: Studies in Language, Folklore, and Literature*, 3rd collection, ed. by MARVIN I. HERZOG, WITA RAVID, and URIEL WEINREICH (1969); JOSHUA A. FISHMAN, *Yiddish in America: Socio-linguistic Description and Analysis* (1965); MARVIN I. HERZOG, *The Yiddish Language in Northern Poland: Its Geography and History* (1965); URIEL WEINREICH, *Modern English-Yiddish, Yiddish-English Dictionary* (1968), and "Yiddish Language," *Encyclopaedia Judaica*, vol. 16, col. 789–798 (1971).

Scandinavian: (*North Germanic*): A survey of research on Scandinavian languages since 1918 by EINAR HAUGEN and THOMAS L. MARKEY is presented in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 11 (forthcoming). Important recent contributions are treated in *The Nordic Languages and Modern Linguistics*, ed. by H. BENEDIKTSSON (1970). (*Histories*): Compact histories of all the languages are presented in ELIAS WESSEN, *Die nordischen Sprachen* (1968); and EINAR HAUGEN, *The Scandinavian Languages* (forthcoming). Detailed histories include: D.A. SEIP, *Norsk språkhistorie til omkring 1370*, 2 vol. (1955; German trans., 1971), for Old Norwegian; PETER SKAUTRUP, *Det danske sprogs historie*, 4 vol. and index (1944–68), for Danish; ELIAS WESSEN, *Svensk språkhistoria*, 3 vol. (1965), for Swedish; GUSTAV L. INDREBO, *Norsk målsoga* (1951), for Norwegian; and EINAR HAUGEN, *Language Conflict and Language Planning* (1966), for the modern Norwegian period. (*Grammars*): The classic grammars of the older languages are ADOLF G. NOREEN, *Altisländische und altnorwegische Grammatik*, 4th ed. (1923); and *Altschwedische Grammatik* (1904); and JOHANNES BRONDUM-NIELSEN, *Gammeldansk grammatik*, 5 vol. (1950–65). For introductory purposes the best grammar is E.V. GORDON, *An Introduction to Old Norse*, 2nd ed. rev. by A.R. TAYLOR (1981). Grammars of the modern languages are: PAUL DIDERICHSEN, *Elementaer dansk grammatik*, 2nd ed. (1957), for Danish, and Diderichsen's compendium *Essentials of Danish Grammar* (1964), as well as AAGE K. HANSEN, *Moderne dansk*, 3 vol. (1967); ADOLF G. NOREEN, *Vårt språk*, 9 vol. (1903–24), for Swedish; AUGUST WESTERN, *Norsk riksmåls grammatikk* (1921), for Dano-Norwegian; OLAV T. BEITO, *Nynorsk grammatikk* (1970), for New Norwegian; W.B. LOCKWOOD, *An Introduction to Modern Faroese* (1955), for Faeroese; STEFAN EINARSSON, *Icelandic: Grammar, Texts, Glossary* (1945). See also JONATHAN WYLIE and DAVID MARGOLIN, *The Ring of Dancers: Images of Faroese Culture* (1981), which treats the language. Introductory textbooks for English-speaking users are (beside the two just preceding) ELIAS BREDSDORFF, *Danish* (1956); EINAR HAUGEN and KENNETH G. CHAPMAN, *Spoken Norwegian*, rev. ed. (1964); NILS-GUSTAV HILDEMAN and ANN-MARI BEITE (eds.), *Learn Swedish*, 2nd ed. (1964); and volumes in the "Teach Yourself" series. (*Dictionaries*): There are many-volumed native dictionaries for each language. Only some bilingual dictionaries are listed here: HERMANN VINTERBERG and JENS AXELSEN, *Dansk-engelsk ordbog*, 7th ed., 2 vol. (1967); RICHARD CLEASBY and GUDBRAND VIGFUSSON, *An Icelandic-English Dictionary*, 2nd ed. (1957); W.E. HARLOCK, *Svensk-engelsk ordbok* (1944); the technical I.E. GULLBERG, *Svensk-engelsk fackordbok* (1964); and EINAR HAUGEN (ed.), *Norwegian-English Dictionary* (1965). EINAR HAUGEN, *Scandinavian Language Structure* (1982), is a comparative survey.

English language. *Dictionaries: The Oxford English Dictionary*, 13 vol. (1933), is the reissue of *The New English Dictionary on Historical Principles* (1884–1928); it is updated by a Supplement (1972–), projected to be four volumes upon completion. Derivative dictionaries include *The Shorter Oxford English Dictionary on Historical Principles*, 2nd ed., 2 vol. (1939); *The Concise Oxford Dictionary of Current English*, 6th ed. (1976); *The Pocket Oxford Dictionary of Current English* (1969); *The Little Oxford Dictionary of Current English*, 4th ed. (1969); *Oxford Illustrated Dictionary*, 2nd ed. (1975); and *The Oxford Advanced Learner's Dictionary of Current English*, 3nd ed. (1974); and the *Oxford American Dictionary* (1980). Other one-volume dictionaries include *Chambers' Twentieth Century* (1972); *The Universal Dictionary of the English Language*, rev. by E.H. PARTRIDGE (1952); *Longmans English Larousse* (1968); and P. HANKS, *Encyclopedic World Dictionary* (1971).

The leading American dictionary is *Webster's Third New International Dictionary of the English Language* (1961), actually 8th in the series since the first appeared in 1828; it is updated by a separately published "Addenda" section, *6,000 Words* (1976). *Webster's New Collegiate Dictionary* (1979) is an abbreviated version. Other comprehensive dictionaries are *The New Century Dictionary of the English Language*, 2 vol. (1959); and *Funk and Wagnalls New Standard Dictionary of the English Language* (1963). Two comprehensive dictionaries are outstanding: *The Random House Dictionary of the English Language* (1966); and *The American Heritage Dictionary of the English Language* (1969).

Reliable etymological dictionaries include ERNEST WEEKLEY, *An Etymological Dictionary of Modern English*, 2 vol. (1921, reprinted 1967); E.H. PARTRIDGE, *Origins*, 5th rev. ed. (1971); and ERNEST KLEIN, *A Comprehensive Etymological Dictionary of the English Language*, 2 vol. (1966–67). *The Oxford Dictionary of English Etymology* (1966) will long remain the most authentic work in this field.

The two great historical dictionaries of American English are: SIR WILLIAM A. CRAIGIE and JAMES R. HULBERT (eds.), *A Dictionary of American English on Historical Principles*, 4 vol. (1936–44); and MITFORD M. MATHEWS, (ed.), *A Dictionary of Americanisms on Historical Principles*, 2 vol. (1951).

Modern usage: H.W. FOWLER's somewhat eccentric *A Dictionary of Modern English Usage* (1926) was thoroughly updated by SIR ERNEST GOWERS (1965). It has its transatlantic counterpart in the following two works: BERGEN and CORNELIA EVANS, *A Dictionary of Contemporary American Usage* (1957); and MARGARET NICHOLSON, *A Dictionary of American-English Usage* (1957). See also *Roy H. Copperud, American Usage and Style* (1980).

Grammar and structure of English: A.A. HILL, *Introduction to Linguistic Structures: From Sound to Sentence in English* (1958); SAMUEL JAY KEYSER and PAUL M. POSRAL, *Beginning English Grammar* (1976); PAUL ROBERTS, *English Sentences* (1962); MARTIN JOOS, *The English Verb* (1964); H.A. GLEASON, *Linguistics and English Grammar* (1965); N.C. STAGEBERG, *An Introductory English Grammar*, 3rd ed. (1977); A.E. DAR-

BYSHIRE, *A Description of English* (1967); R. QUIRK *et al., A Grammar of Contemporary English* (1972); B.M.H. STRANG, *Modern English Structure,* 2nd ed. rev. (1968); R.W. ZANDVOOST, *A Handbook of English Grammar,* 7th ed. (1975).

Phonetics of English: Handbooks include HANS KURATH and R.I. MCDAVID, *The Pronunciation of English in the Atlantic States* (1961); and A.C. GIMSON, *An Introduction to the Pronunciation of English* (1963).

Histories: An excellent account of the "external history" of the language is given by A.C. BAUGH in *A History of the English Language,* 3rd ed. (1978). FERNAND MOSSÉ, *Esquisse d'une histoire de la langue anglaise* (1947), is a masterpiece—brief, lucid, and profound. KARL BRUNNER, *Die englische Sprache: Ihre geschichtliche Entwicklung,* 2nd ed., 2 vol. (1960–62), is indispensable to advanced students.

Two brief surveys written early in the 20th century are recognized classics and remain stimulating: HENRY BRADLEY, *The Making of English* (1904, rev. by SIMEON POTTER, 1968); and J.O.H. JESPERSEN, *Growth and Structure of the English Language* (1905, reprinted 1971). Other fairly substantial histories include STUART ROBERTSON, *The Development of Modern English,* 2nd ed. rev. by FREDERIC G. CASSIDY (1954); M.M. BRYANT, *Modern English and Its Heritage,* 2nd ed. (1962); M.W. BLOOMFIELD and L.D. NEWMARK, *A Linguistic Introduction to the History of English* (1963); W.N. FRANCIS, *The English Language, an Introduction* (1965); THOMAS PYLES, *The Origins and Development of the English Language,* 2nd ed. (1971); SIMEON POTTER, *Our Language,* rev. ed. (1968); J.W. CLARK, *Early English: A Study of Old and Middle English* (1967); A.C. PARTRIDGE, *Tudor to Augustan English* (1969); J.A. SHEARD, *The Words We Use,* rev. ed. (1970); JOSEPH M. WILLIAMS, *Origins of the English Language: A Social and Linguistic History* (1975); and B.M.H. STRANG, *A History of English* (1970). F.T. VISSER, *An Historical Syntax of the English Language,* 3 vol. (1963–73), provides copious illustrations and bibliographies.

Special studies: GEORGE W. TURNER, *The English Language in Australia and New Zealand* (1966); SIMEON POTTER, *Changing English* (1969); JOHN W. SPENCER (ed.), *The English Language in West Africa* (1971); MITFORD M. MATHEWS (ed.), *The Beginnings of American English* (1931); THOMAS PYLES, *Words and Ways of American English* (1952); ALBERT H. MARCKWARDT, *American English,* 2nd. ed. rev. by J.L. DILLARD (1980); and *Black English: Its History and Usage in the United States* (1972), also by Dillard.

Bibliographies: ARTHUR G. KENNEDY, *A Bibliography of Writings on the English Language from the Beginning of Printing to the End of 1922* (1927); HAROLD B. ALLEN, *Linguistics and English Linguistics,* 2nd ed. (1977). New books are recorded in the *Annual Bibliography of English Language and Literature,* edited for the Modern Humanities Research Association, and in *The Year's Work in English Studies* (annual), edited for the English Association. Books and contemporary studies are listed in the *MLA International Bibliography of Books and Articles on the Modern Languages and Literatures* (annual) of the Modern Language Association.

Armenian language. S.L. KOGIAN, *Armenian Grammar* (1949); H. HUBSCHMANN, "Armenische Grammatik," *Armenische Etymologie* (1897); A. MEILLET, *Esquisse d'une grammaire comparée de l'arménien classique,* new ed. (1936); GERHARD DEETERS, *Armenisch und Südkaukasisch* in *Caucasica* (1927); HEINRICH ZELLER, "Armenisch," in *Geschichte der indogermanischen Sprachwissenschaft,* vol. 4 (1927). For ancient Armenian, see A. MEILLET, *Altarmenisches Elementarbuch* (1913); and H. JENSEN *Altarmenische Grammatik* (1959). Medieval Armenian is treated in J. KARST, *Historische Grammatik des Kilikisch-Armenischen* (1901). For modern speech, see H. ADJARIAN, *Classification des dialectes arméniens* (1909); and A. ABEGHIAN, *Neuarmenische Grammatik* (1936).

Tocharian language. E. SIEG and W. SIEGLING (eds.), *Tocharische Sprachreste* (1921), gives the transcription of all the manuscripts in dialect A. A small companion volume, *Tafeln,* reproduces a number of the best preserved leaves in facsimile. E. SIEG, W. SIEGLING, and W. SCHULZE, *Tocharische Grammatik* (1931), is an exhaustive grammar of dialect A, with a verbal index identifying and listing all verb forms in that dialect. H. PEDERSEN, *Tocharisch vom Gesichtspunkt der indoeuropäischen Sprachvergleichung* (1941), is still the best overall comparative study from the Indo-European point of view, even though extensive published materials in dialect B were not available. E. SIEG and W. SIEGLING, *Tocharische Sprachreste, Sprache B,* 2 pt. (1949–53), contain all the Berlin manuscripts in dialect B plus a few from other collections (especially the Hoernle collection in London). Part 1 is an edition of the Udānālaṅkāra fragments with translation and glossary. W. KRAUSE, *Westtocharische Grammatik,* vol. 1, *Das Verbum* (1952), is indispensable for the verb in dialect B (a second volume was never published). W. KRAUSE and W. THOMAS, *Tocharisches Elementarbuch,* vol. 1, *Grammatik* (1960), vol. 2, *Texte und Glossar* (1964), including both dialects, now replace all earlier introductions to the study of Tocharian. Two articles by G.S. LANE, "On the Interrelationship of the Tocharian Dialects," in H. BIRNBAUM and J. PUHVEL (eds.), *Ancient Indo-European Dialects* (1966); and "Tocharian: Indo-European and Non-Indo-European Relationships," in G. CARDONA, H.M. HOENIGSWALD, and A. SENN (eds.), *Indo-European and Indo-Europeans* (1970), attempt to solve some of the problems concerning the varied uses of the two dialects and the general problem of the position of Tocharian within the Indo-European family of languages.

Celtic languages. H. LEWIS and H. PEDERSEN, *A Concise Comparative Celtic Grammar* (1937), is the most recent survey of the entire field. The early history of the British group is discussed in detail by K.H. JACKSON, *Language and History in Early Britain* (1953); for the later history of Welsh and Breton, J. MORRIS JONES, *A Welsh Grammar* (1913), and F. GOURVIL, *Langue et littérature bretonnes* (1952), give useful information. P. BERRESFORD ELLIS, *The Cornish Language and Its Literature* (1974), deals with the early period, as well as with the recent Cornish language revival movement. R. THURNEYSEN, *Grammar of Old Irish,* rev. ed. (1946), is a classic among linguistic handbooks; for the later development of Irish, T.F. O'RAHILLY, *Irish Dialects Past and Present* (1932), is full of information and contains chapters on Scottish Gaelic and Manx. B.O CUIV (ed.), *A View of the Irish Language* (1969), containing 12 essays by various hands, maps, and illustrations; and D. GREENE, *The Irish Language* (1966), are directed to the general reader rather than to the linguist. The relevant sections of GLANVILLE PRICE, *The Present Position of Minority Languages in Western Europe* (1969), give full bibliographies of works dealing with the political and social status of the surviving Celtic languages.

Baltic languages. There are very few works on the Baltic languages in English aside from LEONARDAS DAMBRIŪNAS, ANTANAS KLIMAS, and WILLIAM R. SCHMALSTIEG, *Introduction to Modern Lithuanian* (1966); TERĒZA BUDIŅA LAZDIŅA, *Teach Yourself Latvian* (1966); and JANIS ENDZELĪNS, *Baltų kalbų garsai ir formos* (1957; Eng. trans., *Comparative Phonology and Morphology of the Baltic Languages,* 1971). *Baltic Linguistics,* ed. by THOMAS F. MAGNER and WILLIAM R. SCHMALSTIEG (1970), is a collection of papers on various aspects of Baltic linguistics.

Works in other languages include REINHOLD TRAUTMANN, *Die altpreussischen Sprachdenkmäler* (1910); JOHANN ENDZELIN, *Lettische Grammatik* (1922; trans. into Latvian, 1951); KAZIMIERAS BŪGA, *Lietuvių kalbos žodynas* (1924–25), the introduction to this dictionary contains much valuable information on the history of the Baltic languages; *Rinktiniai raštai,* 3 vol. (1958–61); JANIS ENDZELĪNS, *Senprūšu valoda* (1943; trans. into German, 1944, without a glossary); *Ievads baltu filoloģijā* (1945), an introduction to Baltic linguistics, in Latvian; ERNST FRAENKEL, *Die baltischen Sprachen* (1950), a general introduction; ALFRED SENN, "Die Beziehungen des Baltischen zum Slavischen und Germanischen," *Zeitschrift für vergleichende Sprachforschung,* vol. 71 (1954), discusses the relationship of Baltic to Slavic and Germanic; *Mūsdienu latviešu literārās valodas gramatika,* 2 vol. (1959–62), a grammar of the modern Latvian literary language; ARTURS OZOLS, *Veclatviešu rakstu valoda* (1965), treats Old Latvian; JAN OTREBSKI, *Gramatyka języka litewskiego,* 3 vol. (1956–65), a Lithuanian grammar, in Polish; MARTA RUDZĪTE, *Latviešu dialektoloģija* (1964), treats Latvian dialectology; *Lietuvių kalbos gramatika,* 2 vol. (1956, 1971), the most authoritative grammar of the Lithuanian language; ZIGMAS ZINKEVICIUS, *Lietuvių dialektologija* (1966), a valuable treatment of Lithuanian dialectology; CHRISTIAN S. STANG, *Vergleichende Grammatik der Baltischen Sprachen* (1966), the only scholarly comparative grammar of the Baltic languages; VYTAUTAS J. MAZIULIS (comp.), *Prūsų kalbos paminklai* (1966), contains and discusses all the photographed Old Prussian texts; ALGIRDAS SABALIAUSKAS, "Lietuvių kalbos leksikos raida," *Lietuvių kalbotyros klausimai,* 8:5–140 (1966), treats the development of the vocabulary of Lithuanian; JONAS PALIONIS, *Lietuvių literatūrinė kalba XVI–XVII a.* (1967), a treatment of the Lithuanian literary language in the 16th and 17th centuries; JONAS KAZLAUSKAS, *Lietuvių kalbos istorinė gramatika* (1968), the only historical grammar of Lithuanian; VYTAUTAS MAZIULIS, *Baltų ir kitų indoeuropiečių kalbų santykiai* (1970), treats the relationship of Baltic and the other Indo-European languages.

The largest dictionaries of Baltic languages are: K. MULENBACHS, *Latviešu valodas vārdnīca,* 4 vol. (1923–32); *Lietuvių kalbos žodynas,* 8 vol. (1941–70); and ERNST FRAENKEL, *Litauisches etymologisches Wörterbuch,* 2 vol. (1955–65).

Slavic languages. ROMAN JAKOBSON, *Slavic Languages,* 2nd ed. (1955), is the best short structural sketch. A useful general survey is REINHOLD TRAUTMANN, *Die slavische Völker und Sprachen* (1947). The comparative grammar is described in the following works: ANTOINE MEILLET, *Le Slave com-*

mun, 2nd ed. rev. (1934), the best introduction to Proto-Slavic from the point of view of Indo-European; VACLAV VONDRAK, *Vergleichende slavische Grammatik,* 2 vol. (1924–28), rich in material; JOOSEPPI J. MIKKOLA, *Urslavische Grammatik,* 3 vol. (1913–50), condensed, but not quite up-to-date now. The general drift of the languages is seen in RAJKO NAHTIGAL, *Slovanski jeziki,* 2nd ed. (1952). A traditional approach to the history of the phonemic systems is exemplified in САМУИЛ БОРИСОВИЧ БЕРНШТЕЙН, *Очерк сравнительной грамматики славянских языков,* vol. 1 (1961); ANDRE VAILLANT, *Grammaire comparée des langues slaves,* vol. 1 (1950); and PEETER ARUMAA, *Urslavische grammatik,* vol. 1 (1964); recent developments in the diachronic phonology are summarized in ROMAN JAKOBSON, *Selected Writings,* 2nd ed., 6 vol. (1971–); FRANTISEK MARES, *The Origin of the Slavic Phonological System and Its Development Up to the End of Slavic Language Unity* (1965); GEORGE SHEVELOV, *A Prehistory of Slavic* (1964); CHRISTIAN S. STANG, *Slavonic Accentuation* (1957); and NICHOLAS VAN WIJK, *Les Langues slaves, de l'unité à la pluralité,* 2nd ed. rev. (1956). The dialectal differentiation of Proto-Slavic from the point of view of comparative phonology is analyzed in ANTONI FURDAL, *Rozpad języka prasłowjańskiego w świetle rozwoju głosowego* (1961); and HENRIK BIRNBAUM, "The Dialects of Common Slavic," in HENRIK BIRNBAUM and JAAN PUHVEL (eds.), *Ancient Indo-European Dialects* (1966).The problem of the original territory and migrations of the speakers of Proto-Slavic is discussed in the light of linguistic and archaeological evidence in the following works: TADEUSZ LEHR-SPLAWINSKI, *O pochodzeniu i praojczyzńie Słowian* (1964), a brilliant survey using archaeological data; KAZIMIERZ MOSZYNSKI, *Pierwotny zasiąg języka prasłowiańskiego* (1957), employs a strictly linguistic approach, and is hardly convincing; K.H. MENGES, *An Outline of the Early History and Migrations of the Slavs* (1953), cites important Oriental sources; MARIJA GIMBUTAS, *The Slavs* (1971), provides an archaeological and linguistic survey of Slavic migrations to Central Europe and the Balkan Peninsula; and S.H. CROSS, *Slavic Civilization Through the Ages* (1948, reprinted 1963), provides some general information. The only reliable Common Slavic etymological dictionary remains the unfinished work of ERICH BERNEKER, *Slavisches etymologisches Wörterbuch,* 2nd ed., vol. 1 (1924). Of essential value for proof of the close links between Slavic and Baltic is REINHOLD TRAUTMANN, *Baltisch-Slavisches Wörterbuch* (1923); on Slavic, Baltic, and Germanic see CHRISTIAN S. STANG, *Lexikalische Sonderüberstimmungen Zurischen dem Slavischen, Baltischen, und Germanischen* (1972); the study of verbal forms is presented in CHRISTIAN S. STANG, *Das slavische und baltische Verbum* (1942); and H. KOLLN, *Oppositions of Voice in Greek, Slavic and Baltic* (1969). For the accentual pattern of Slavic as compared to Baltic, see especially ВЛАДИСЛАВ МАРКОВИЧ ИЛЛИЧ-СВИТЫЧ, *Именная акцентуация в балтийской и славянском* (1963). The only English textbook introduction to concrete data on each language remains R.G.A. DE BRAY, *Guide to the Slavonic Languages,* rev. ed. (1969). Russian phonetics is treated in KÁLMÁN BOLLA, *A Conspectus of Russian Speech Sounds* (1981).

Albanian language. P.L. HORECKY (ed.), *Southeastern Europe: A Guide to Basic Publications,* pp. 102–111 (1969}, includes a compilation by E.P. HAMP of about 50 major annotated items on Albanian scholarship, language, literature, folklore, ethnography, and folk music, with references to other supporting work; E.P. HAMP, "Albanian," in T.A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 9, pp. 1626–92 (1972), a review of work since 1918, with copious bibliography; LEONARD NEWMARK, P. HUBBARD and PETER PRIFTI, *Standard Albanian* (1982), a reference grammar.

Uralic languages. *General works:* The following manuals primarily reflect the views of their authors, but should serve as a basis for further study (especially in the numerous grammars of Uralic languages and articles on problems of Uralic linguistics—few of which are available in English). BJORN COLLINDER, *Fenno-Ugric Vocabulary: An Etymological Dictionary of the Uralic Languages* (1955), presents comparative Uralic word lists; *An Introduction to the Uralic Languages* (1965); and *Survey of the Uralic Languages* (1957), give short sketches of all but a few of the Uralic languages, with the lesser languages receiving only superficial treatment; LAURI HAKULINEN, *The Structure and Development of the Finnish Language* (Eng. trans. from the Finnish, 1961), an excellent presentation of Finnish from its earliest stages; TOIVO VUORELA, *The Finno-Ugric Peoples* (Eng. trans. from the Finnish, 1964), an anthropological survey.

Works dealing with specific languages: PETER HAJDU, *The Samoyed Peoples and Languages* (Eng. trans from the Hungarian, 1963); ROBERT T. HARMS, *Estonian Grammar* (1962), with an appendix that surveys numerous approaches to the problem of quantity in Estonian; THOMAS A. SEBEOK and FRANCES J. INGEMANN, *An Eastern Cheremis Manual* (1961), a clear, concise description of one of the lesser languages; JOHN ATKINSON, *Finnish Grammar* (1956); and ZOLTAN BANHIDI, ZOLTAN JOKAY, and DENES SZABO, *Learn Hungarian* (1965), two basic grammars.

Altaic languages. N. POPPE, *Introduction to Altaic Linguistics* (1965), is the best up-to-date manual, giving a short but complete picture of the Altaic languages, their history and structure, as well as the theory of the relationship of these languages. *Philologiae Turcicae Fundamenta* (1959); *Handbuch der Orientalistik:* vol. 5, pt. 1, *Turkologie* (1963); pt. 2, *Mongolistik* (1964); and pt. 3, *Tungusologie* (1968), give systematic descriptions of the various historical and contemporary languages with much bibliographic data. K.H. MENGES, *The Turkic Languages and Peoples* (1968), is the best recent account of Turkic linguistics, supplementing the bibliographies in the works mentioned above. G.J. RAMSTEDT, *Einführung in die altaische Sprachwissenschaft,* vol. 1, *Lautlehre* (1957), and vol. 2, *Formenlehre* (1952), is the classical summary of the theory of relationship among the Altaic languages. R. LOEWENTHAL, *The Turkic Languages and Literatures of Central Asia* (1956); D. SINOR, *Introduction à l'étude de l'Eurasie Centrale* (1963); and G. HAZAI (ed.), *Sovietico-Turcica* (1960), are useful bibliographies that complement one another.

Dravidian languages. R. CALDWELL, *A Comparative Grammar of the Dravidian or South-Indian Family of Languages,* 3rd ed. by J.L. WYATT and T.R. PILLAI (1913, reprinted 1956 and 1961), the classic work that laid the foundations of Dravidian linguistics; G.A. GRIERSON (ed.), *Linguistic Survey of India,* vol. 4, *Munḍā and Dravidian Languages,* by S. KONOW (1906); T. BURROW and M.B. EMENEAU, *A Dravidian Etymological Dictionary* (1961, reprinted 1966; *Supplement,* 1968), the first etymological dictionary of the family, marking a new era in Dravidian studies (indispensable point of departure for any further work in the field); B. KRISHNAMURTI, *Telugu Verbal Bases: A Comparative and Descriptive Study* (1961), an indispensable study of the phonology and derivational morphology of Dravidian, with a much wider coverage of problems than the title suggests; "Comparative Dravidian Studies," *Linguistics in South Asia,* pp. 309–333, vol. 5 of *Current Trends in Linguistics* (1969), a summary treatment of the latest developments in the field; K. ZVELEBIL, *Comparative Dravidian Phonology* (1970), the first systematic compendium of the comparative phonology of Dravidian; J. BLOCH, *Structure grammaticale des langues dravidiennes* (1946; Eng. trans., *The Grammatical Structure of Dravidian Languages,* 1954), an excellent description of the main morphological and syntactic features of the family that ignores phonology totally; F.B.J. KUIPER, "The Genesis of a Linguistic Area," *Indo-Iranian Journal,* 10:81–102 (1967), a brief and brilliant treatment of the problems of Aryan and Dravidian convergence; M.S. ANDRONOV, *Materials for a Bibliography of Dravidian Linguistics* (1966); M. ISRAEL, "Additional Materials for a Bibliography of Dravidian Languages," *Tamil Culture,* 12:69–74 (1966); S.E. MONTGOMERY, "Supplemental Materials for a Bibliography of Dravidian Linguistics," *Studies in Indian Linguistics,* pp. 234–246 (1968), three bibliographies that provide fairly complete coverage.

Austro-Asiatic languages. H.L. SHORTO, J.M. JACOB, and E.H.S. SIMMONDS (comps.), *Bibliographies of Mon-Khmer and Tai Linguistics* (1963), is an unannotated bibliography of linguistic books and articles from the beginning (1790) to 1960, which does not include the Munda subfamily or the Viet-Muong branch but incorporates the (Austronesian) Cham languages into Mon-Khmer. See W. SCHMIDT, *Grundzüge einer Lautlehre der Mon-Khmer-Sprachen* (1906); *Die Mon-Khmer-Völker: Ein Bindeglied zwischen Völkern Zentralasiens und Austronesiens* (1906; French trans., "Les Peuples Mon-Khmer: trait-d'union entre les peuples de l'Asie Centrale et de l'Austronésie," in *Bulletin de l'École française d'Extrême-Orient,* 7: 213–263 and 8:1–35, 1907–08); and W.W. SKEAT and C.O. BLAGDEN, *Pagan Races of the Malay Peninsula,* 2 vol. (1906). Schmidt's articles for the first time supported the Austro-Asiatic hypothesis with lexical, phonological, and morphological evidence. They remain until today the basic work of Austro-Asiatic studies. Blagden compiled a very large comparative vocabulary but did not attempt any analysis. H.J. PINNOW, *Versuch einer historischen Lautlehre der Kharia-Sprache* (1959), an ambitious project with somewhat uncertain results, contains an analysis and systematic comparison of the phonologies of Munda languages and establishes connections with the rest of the Austro-Asiatic group. The *Mon-Khmer Studies* (Linguistic Circle of Saigon), 4 vol. (1964–), are collections of short technical articles mostly on the Montagnard languages of South Vietnam, with topics varying from basic vocabulary to phonology, morphology, syntax, folk taxonomies, and oral literature.

Sino-Tibetan Languages. *General works:* PAUL BENEDICT, *Sino-Tibetan: A Conspectus* (1971), a comprehensive and original study; G.A. GRIERSON, *Tibeto-Burman Family,* in the *Linguistic Survey of India,* vol. 3 (1909), a wealth of material but of uneven quality; FRANK M. LEBAR, GERALD C. HICKEY, and

JOHN K. MUSGRAVE, *Ethnic Groups of Mainland Southeast Asia* (1964), an excellent reference work; HENRI MASPERO, "Langues de l'Asie du Sud-Est," in ANTOINE MEILLET and MARCEL COHEN (eds.), *Les Langues du Monde,* new ed., pp. 524–644 (1952), the most authoritative concise treatment of Sino-Tibetan to date; C.F. and F.M. VOEGELIN, "Languages of the World: Sino-Tibetan," *Anthropological Linguistics,* vol. 6 and 7 (1964–65), much information of a semitechnical nature; ROBERT SHAFER, *Introduction to Sino-Tibetan,* 3 vol. (1966–), a comprehensive and extensive, but technical, series; and *Bibliography of Sino-Tibetan Languages,* 2 vol. (1957–63), indispensable for further research.

Chinese: NICHOLAS C. BODMAN, "China: Historical Linguistics," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 2, pp. 3–58 (1967); KUN CHANG, "China: Descriptive Linguistics," *ibid.,* pp. 59–90; YUEN-REN CHAO, *Mandarin Primer* (1948), excellent chapters on script and grammar; and with LIEN-SHENG YANG, *Concise Dictionary of Spoken Chinese* (1947), the best dictionary of modern Chinese, with an excellent introduction; JOHN DeFRANCIS, *Nationalism and Language Reform in China* (1950), informative and readable; and "China: Language and Script Reform," in *Current Trends in Linguistics,* vol. 2, pp. 130–150 (1967); SOREN EGEROD, "China: Dialectology," *ibid.,* pp. 91–129, technical in nature; R.A.D. FORREST, *The Chinese Language,* 2nd ed. rev. (1965), a standard reference work that also treats related and contiguous languages; BERNHARD KARLGREN, *Études sur la phonologie chinoise,* 4 pt. (1915–26), an epoch-making work but very technical; *Compendium of Phonetics in Ancient and Archaic Chinese* (1954), also technical; *Grammata Serica: Script and Phonetics in Chinese and Sino-Japanese* (1957), the standard dictionary of Old Chinese characters; *Sound and Symbol in Chinese,* rev. ed. (1962), very readable, but somewhat out of date; *The Chinese Language* (1949), a popular account of phonetic reconstructions; and *Easy Lessons in Chinese Writing* (1958), an interesting account of the etymology of Chinese characters; PAUL KRATOCHVIL, *The Chinese Language Today* (1968), very readable and up to date.

Tibeto-Burman and Karen: KUN CHANG, "China: National Languages," in *Current Trends in Linguistics,* vol. 2, pp. 151–176 (1967), treats minority languages in China; ROBERT B. JONES, *The Burmese Writing System* (1953), offers the best description; *Karen Linguistic Studies* (1961), up to date but technical; ROY A. MILLER, *The Tibetan System of Writing* (1956), offers the best description of the subject; "The Tibeto-Burman Languages of South Asia," *Current Trends in Linguistics,* vol. 5, pp. 431–449 (1969), informative and up to date; STUART N. WOLFENDEN, *Outlines of Tibeto-Burman Linguistic Morphology* (1929), the classic statement.

Tai languages. M.R. HAAS, *Thai-English Student's Dictionary* (1964), a concise dictionary arranged according to the order of the Thai alphabet with a brief description of the phonological and grammatical system, and with H.R. SUBHANKA, *Spoken Thai,* 2 vol. (1947), a textbook for learning spoken Thai step by step; F.M. LABAR, G.C. HICKEY, and J.K. MUSGRAVE, *Ethnic Groups of Mainland Southeast Asia,* pt. 3, pp. 187–244 (1964), a detailed description of different Tai ethnic groups, including their location, demography, and social organization; FANG KUEI LI, "Consonant Clusters in Tai," *Language,* 30:368–379 (1954), an article demonstrating the existence of a large number of initial consonant clusters in the protolanguage and their simplification in the modern dialects; "A Tentative Classification of Tai Dialects," in S. DIAMOND (ed.), *Culture in History,* pp. 951–959 (1960), a classification establishing three groups of Tai languages and dialects according to lexical distribution and phonological development, and "The Tai and the Kam-Sui Languages," *Lingua,* 14:148–179 (1965), a discussion of the relationship of these two language groups; R.B. NOSS, *Thai Reference Grammar* (1964), a detailed descriptive and structural analysis of the Thai language; P.K. BENEDICT, "Thai, Kadai and Indonesian: A New Alignment in Southeastern Asia," *American Anthropologist,* 44:576–601 (1942), a provocative attempt to show the relationship of the Tai languages to the Indonesian languages; TATSUO HOSHINO and RUSSELL MARCUS, *Lao for Beginners* (1981).

Paleo-Siberian languages. ROMAN JAKOBSON, GERTA HUTTLWORTH, and JOHN FRED BEEBE, *Paleosiberian Peoples and Languages: A Bibliographical Guide* (1957, reprinted 1981), very useful, with informative appendix; DEAN WORTH, "Paleosiberian," in T.A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 1, *Soviet and East European Linguistics,* pp. 345–373 (1963), good coverage of Soviet work since World War II. The best and most recent source on Paleosiberian languages is in Russian: П.Я. СКОРИК *et al.* (eds.), *Языки народов С.С.С.Р.,* vol. 5, *Монгольские, тунгусоманьчжурские и лалеоазиатские языки* (1968).

Caucasian languages. *General works of reference:* G. DEETERS, *Armenisch und kaukasische Sprachen* (1963), a general survey and a combined presentation of the structure of the Caucasian languages according to the most characteristic features of phonology, morphology, and syntax, with an extensive bibliography; G.A. KLIMOV, *Die kaukasischen Sprachen* (1969; orig. pub. in Russian, 1965), a brief exposition of the history and structures of the Caucasian languages, with a general characterization of each group, including an extensive bibliography; A.H. KUIPERS, "Caucasian," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 1, *Soviet and East European Linguistics,* pp. 315–344 (1963), a useful brief survey of Caucasian linguistics, with a selected bibliography; A. DIRR, *Einführung in das Studium der kaukasischen Sprachen* (1928), contains a survey of the structure of individual Caucasian languages and their interrelationships; "Jazyki Narodov SSSR," IV. *Iberijsko-Kavkazskie jazyki* (1967), a brief exposition of the structures of all the Caucasian languages, with a selected bibliography.

South Caucasian languages: The following is a selection of the more important special works on the South Caucasian languages: G. DEETERS, *Das kharthwelische Verbum: Vergleichende Darstellung des Verbalbaus der südkaukasischen Sprachen* (1930), a comprehensive comparative study of the verb structure of the Kartvelian languages; A.S. CHIKOBAVA, *Drevnejšaja struktura imennyx osnov v kartvel'skix jazykax* (1942; in Georgian, with a Russian and French summary), a comparative analysis of the ancient structure of nominal stems in the Kartvelian languages, with an interpretation of certain prefixes as the ancient classmarkers; K.H. SCHMIDT, *Studien zur Rekonstruktion des Lautstandes der südkaukasischen Grundsprache* (1962), a detailed analysis of sound correspondences with a reconstruction of the Proto-Kartvelian phonemic system; T.V. GAMKRELIDZE and G.I. MACHAVARIANI, *Sistema sonantov i ablaut v kartvel'skikh iazykakh* (1965; in Georgian with a Russian summary), a detailed comparative analysis of the Kartvelian phonological and morphophonological system, with a reconstruction of resonants and ablaut alternations in Proto-Kartvelian and their typological evaluation. For the grammars of the individual languages see A.G. SHANIDZE, *Osnovy gruzinskoj grammatiki,* vol. 1 (1953), a most comprehensive exposition (in Georgian) of the structure of modern Georgian; and H. VOGT, *Grammaire de la langue géorgienne,* rev. ed. (1971). An account of the Georgian sound system is given in G. AXVLEDIANI, *Osnovy obščej fonetiki* (1949, in Georgian). For a detailed descriptive analysis of the Svan verb system according to dialects, see V.T. TOPURIA, *Svanskij jazyk,* vol. 1, *Glagol* (1931; 2nd ed., 1967), in Georgian, with a Russian summary. Much useful information about Georgian and the history of Georgian (Kartvelian) studies is contained in S.V. DZIDZIGURI, *The Georgian Language* (Eng. trans. 1968). A useful practical guide to Georgian is K. TSCHENKELI, *Einführung in die georgische Sprache,* vol. 1, *Theoretischer Teil,* vol. 2, *Praktischer Teil* (1958). See also DEE ANN HOLISKY, *Aspect and Georgian Medial Verbs* (1981).

North Caucasian languages: P.K. USLAR, *Etnografija Kavkaza. Jazykoznanie,* 6 vol. (1887–96), contains descriptive grammars of the individual North Caucasian languages; A. TSCHIKOBAVA, "Die ibero-kaukasischen Gebirgssprachen und der heutige Stand ihrer Erforschung in Georgien," *Acta Orientalia Academiae Scientiarum Hungaricae,* 9:109–161 (1959), a brief survey of the North Caucasian languages, containing an extensive bibliography; N. TRUBETZKOY, "Nordkaukasische Wortgleichungen," *Wiener Zeitschrift für die Kunde des Morgenlandes,* vol. 37, no. 1–2 (1930), establishes sets of sound correspondences between the West and East Caucasian languages and deals with the history of their consonantism; G. DUMEZIL, *Études comparatives sur les langues caucasiennes du nord-ouest* (*morphologie*) (1932), a comparative analysis of the grammatical structure of the Abkhazo-Adyghian languages; A.H. KUIPERS, *Phoneme and morpheme in Kabardian* (*Eastern Adyghe*) (1960), a detailed analysis of the phonemic structure of morphemes in Kabardian with a typological comparison with other linguistic systems; W.S. ALLEN, "Structure and System in the Abaza Verbal Complex," *Transactions of the Philological Society of London,* pp. 127–176 (1956), a comprehensive analysis of the verb structure in Abaza; A. SOMMERFELT, "*Études comparatives sur le caucasique du nord-ouest, Norsk Tidsskrift for Sprogvidenskap,* vol. 7 and 9 (1934–38), a comparative study of the sound system of the Nakh languages; E.A. BOKAREV, *Vvedenie v sravnitel'no-istoričeskoe izučenie dagestanskix jazykov* (1961), a comparative study of the vocalism and consonantism of the languages of Dagestan; T.E. GUDAVA, *Konsonantizm andijskix jazykov* (1964), a reconstruction of the original consonant system of the Avaro-Ando-Dido languages.

Hamito-Semitic languages. *General works:* There is still no general survey of the field (including bibliography) to replace I.M. DIAKONOFF, *Semitohamitskie iazyki* (1965; Eng. trans., *Semito-Hamitic Languages,* 1965), although the developments in the field after 1965 have been considerable (see the *Hamito-Semitic a,* 1975). Among other general surveys are

G.R. CASTELLINO, *The Akkadian Personal Pronouns and Verbal System in the Light of Semitic and Hamitic* (1962), and T.W. THACKER, *The Relationship of the Semitic and Egyptian Verbal Systems* (1954).

The theoretical problems of the Hamito-Semitic family have, until recently, been studied mostly on the basis of the Semitic branch alone. Especially important in this respect are I.J. GELB, *Sequential Reconstruction of Proto-Akkadian* (1969); JERZY KURYŁOWICZ, *L'Apophonie en sémitique* (1961); FRITHIOF RUNDGREN, *Intensiv und Aspektkorrelation* (1959). Perhaps the most crucial problem of the Proto-Hamito-Semitic linguistic typology is the reconstruction of the verbal system, for which, besides the above-mentioned works, see also MARCEL COHEN, *Le Système verbal sémitique et l'expression du temps* (1924); O. ROESSLER, "Akkadisches und libysches Verbum," *Orientalia,* vol. 20 (1951): A. KLINGENHEBEN, "Die Präfix- und die Suffixkonjugation des Hamitosemitischen," *Mitteilungen des Instituts für Orientforschung,* vol. 2 (1957), and the critical, although certainly not final review of some later ideas in PELIO FRONZAROLI, "Ricostruzione interna del verbo semitico in alcuni studi recenti," *Accademia Toscana "La Colombaria,"* pp. 71–85 (1972). Another theoretical problem is broached in I.M. DIAKONOFF, "Problems of Root Structure in Proto-Semitic," *Archiv Orientální,* 38:453–480 (1970). The best general review of the Semitic branch of Hamito-Semitic is GOTTHELF BERGSTRAESSER, *Einführung in die semitischen Sprachen* (1928, reprinted 1963). See also GEORGIO LEVI DELLA VIDA (ed.), *Semitic Linguistics: Present and Future* (1961); SABATINO MOSCATI, *An Introduction to the Comparative Grammar of the Semitic Languages* (1964); I.M. DIAKONOFF, *Jazyki drevnej Perednej Azii* (1967), on languages of the Ancient Near East, including Semitic and a general survey of Common Hamito-Semitic. Recent developments are evaluated in *Current Trends in Linguistics,* vol. 6: *Linguistics in South West Asia and North Africa,* with a comprehensive bibliography. JOSHUA BLAU, *The Renaissance of Modern Hebrew and Modern Standard Arabic* (1981), is a comparative study.

For treatment of the individual Semitic languages and the other branches of Hamito-Semitic, the following works may be consulted:

Akkadian: WOLFRAM VON SODEN, *Grundriss der akkadischen Grammatik,* 2nd ed. (1969), and *Akkadisches Handwörterbuch* (1959–); ERICA REINER, *A Linguistic Analysis of Akkadian* (1966); *The Assyrian Dictionary,* published by the Oriental Institute, the University of Chicago (1956–).

Northern Central Semitic: C.H. GORDON, *Ugaritic Textbook,* 2nd ed. (1965); JOSEPH AISTLEITNER, *Wörterbuch der ugaritischen Sprache* (1963); I.J. GELB, "La lingua degli Amoriti," *Rendiconti d. Accademia Nazionale dei Lincei,* 13:143–164 (1958); GIOVANNI GARBINI, *Il Semitico di Nord-Ovest* (1960); Z.S. HARRIS, *Development of the Canaanite Dialects* (1939); JOHANNES FRIEDRICH, *Phönizisch-punische Grammatik* (1951); GEORG BEER and RUDOLF MEYER, *Hebräische Grammatik,* 2 vol. (1952–55); WILHELM GESENIUS and GOTTHELF BERGSTRAESSER, *Hebräische Grammatik,* 29th ed., 2 vol. (1918–29); HANS BAUER and PONTUS LEANDER, *Historische Grammatik der Hebräischen Sprache des Alten Testamentes* (1922); H.B. ROSEN, *A Textbook of Israeli Hebrew,* 2nd ed. (1966); LUDWIG KOEHLER and WALTER BAUMGARTNER, *Lexicon in Veteris Testamenti Libros* (1958); E. BEN YEHUDA, *Thesaurus totius Hebraitatis* (1908–58); FRANZ ROSENTHAL, *Die aramaistische Forschung seit Th. Nöldeke's Veröffentlichungen* (1939) and *A Grammar of Biblical Aramaic* (1961); PONTUS LEANDER, *Lautund Formenlehre des Ägyptisch-Aramäischen* (1928, reprinted 1966); JEAN CANTINEAU, *Le Nabatéen,* 2 vol. (1930–32); HARRIS BIRKELAND, *The Language of Jesus* (1954); E.Y. KUTSCHER, *Studies in Galilean Aramaic* (1952); THEODOR NOELDEKE, *Kurzgefasste syrische Grammatik* (1880, reprinted 1966); CARL BROCKELMANN, *Syrische Grammatik,* 8th ed. (1960); RUDOLF MACUCH, *Handbook of Classical and Modern Mandaic* (1965); KONSTANTIN CERETELI, "Abriss der vergleichenden Phonetik der modernen assyrischen Dialekte," in FRANZ ALTHEIM, *Geschichte der Hunnen,* vol. 3, pp. 218–266 (1961); IRENE GARBELL, *The Jewish Neo-Aramaic Dialect of Persian Azerbaijan* (1965); CHARLES F. JEAN and JACOB HOFTIJZER, *Dictionnaire des inscriptions sémitiques de l'Ouest* (1960–65); MARCUS JASTROW, *A Dictionary of the Targumim, The Talmud Babli and Jerushalmi, and the Midrashic Literature,* 2 vol. (1950); ROBERT PAYNE SMITH, *Thesaurus Syriacus,* 2 vol. (1868–1901); E.S. DROWER and RUDOLF MACUCH, *A Mandaic Dictionary* (1963).

Southern Central Semitic, or Arabic: CHAIM RABIN, *Ancient West Arabian* (1951); HENRI FLEISCH, *L'Arabe classique* (1956); JEAN CANTINEAU, *Cours de phonétique arabe* (1960) *and La Dialectologie arabe* (1955); A. SUTCLIFFE, *Grammar of the Maltese Language* (1936); E.W. LANE, *An Arabic-English Lexicon,* 8 vol. (1863–93).

Southern Peripheral Semitic: A.F.L. BEESTON, *A Descriptive Grammar of Epigraphic South Arabian* (1962); MARIA HOEFNER, *Altsüdarabische Grammatik* (1943); EWALD WAGNER, *Syntax der Mehri-Sprache* (1953); WOLF LESLAU, *Lexique soqoṭri* (1938); M. BITTNER, "Charakteristik der Sprache der Insel Soqoṭra," *Anzeiger der Wiener Akademie der Wissenschaften, Ph.-hist. Kl.,* vol. 55 (1918)—the same author has published a number of studies on the important languages Mahrī and Shaḥrī ("Shkhauri") in the *Sitzungsberichte* of the *Wiener Akademie* between 1909 and 1917; CHRISTIAN DILLMANN, *Ethiopic Grammar,* 2nd ed. (1907) and *Lexicon linguae Aethiopicae* (1865); WOLF LESLAU, *Étude descriptive et comparative du Gafat* (1956) and *Etymological Dictionary of Harari* (1963); EDWARD ULLENDORFF, *The Semitic Languages of Ethiopia: A Comparative Phonology* (1955).

Egyptian: ELMAR EDEL, *Altägyptische Grammatik* (1955–64); A.H. GARDINER, *Egyptian Grammar,* 3rd ed. (1957); WALTER TILL, *Koptische Grammatik (Saïdischer Dialekt)* (1955); ADOLF ERMAN and HERMANN GRAPOW (eds.), *Wörterbuch der ägyptischen Sprache,* 6 vol. (1926–31, reprinted 1955); WOLJA ERICHSEN, *Demotisches Glossar* (1954).

Berber: ANDRE BASSET, *Handbook of African Languages,* vol. 1, *La Langue berbère* (1952), and *Articles de dialectologie berbère* (1959). An extensive classified bibliography is provided by JOSEPH R. APPLEGATE as a sequel to his article "The Berber Languages" in *Current Trends in Linguistics,* vol. 6 (1970).

Cushitic: M.L. BENDER, "The Languages of Ethiopia: A New Lexicostatistic Classification and Some Problems of Diffusion," *Anthropological Linguistics,* vol. 13 (1971); A.B. DOLGOPOL'SKIJ, *Sravitel'no-istoricheskaja fonetika kušitskikh jazykov* (1973); on the individual branches and languages of Cushitic, see C.R. BELL, *The Somali Language* (1953), a manual of the Isāq dialect; M.M. MORENO, *Il somalo della Somalia* (1955), devoted to the Benadir, Darod, and Digil dialects; ENRICO CERULLI, *Studi etiopici,* 4 vol. (1936–51), contains grammars and vocabularies of Sidamo, Janjero, some Ometo dialects, and Kafa (Kaficho). For a good survey of the individual branches and languages of Cushitic, see F.R. Palmer, "Cushitic," in *Current Trends in Linguistics,* vol. 6, pp. 571–585.

Chadic: On the general problems of the Chadic branch see PAUL NEWMAN and ROXANA MA, "Comparative Chadic: Phonology and Lexicon," *Journal of African Languages,* vol. 5 (1966); D. WESTERMANN and M. BRYAN, "Languages of West Africa," *Handbook of African Languages,* vol. 2 (1952); JOSEPH H. GREENBERG, "Studies in African Linguistic Classfication, IV, Hamito-Semitic," *SWest. J. Anthrop.,* vol. 6 (1950). Among the studies of individual Chadic languages those devoted to Hausa are the most numerous of all; the following may be especially useful: C.T. HODGE and IBRAHIM UMARU, *Hausa: Basic Course* (1963); CHARLES H. KRAFT, *A Study of Hausa Syntax,* 3 vol. (1963); R.C. ABRAHAM, *The Language of the Hausa People* (1959). Among the grammatical studies of the other Chadic languages are HERRMANN JUNGRAITHMAYR, *Die Ron-Sprachen* (1970); CARL HOFFMAN, *A Grammar of the Margi Language* (1963); and H.D. FOULKES, *Angass Manual* (1915). Of the many vocabularies of Hausa the most important is G.P.A. BARGERY (comp.), *A Hausa-English Dictionary and English-Hausa Vocabulary* (1934).

Korean language. The McCune-Reischauer system of transcription used in this article is the one most widely used; it is described in detail in G.M. MCCUNE and E.O. REISCHAUER, "The Romanization of the Korean Language," *Trans. Korea Brch R. Asiat. Soc.,* vol. 29 (1939), and in an abbreviated account in *ibid.,* vol. 38 (1961). For the history and development of the Korean alphabet, see the English resumé in KI-MOON LEE, *Kugŏ P'yogippŏp ŭi Yŏksajŏk Yŏn'gu* (1963). For Chinese characters in Korea, see HYONG-GYU KIM, "Chinese Characters and Korean Language," *Korea Journal,* vol. 3 (1963). The introductory chapter, "A Short History of Korean Literature," in DOO SOO SUH, *Korean Literary Reader* (1965), points out relations between Korean writing systems and literature. Surveys and bibliographies of studies in Korean linguistics are found in FRED LUKOFF, "The Republic of Korea," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 2 (1967), and in KI-MOON LEE, "Linguistics," in SUNG-NYONG LEE (ed.), *Korean Studies Today* (1970). The basic western work in the field of Korean grammar, and one which has provided much of the inspiration and direction for studies in Korean historical and comparative linguistics, is G.J. RAMSTEDT, "A Korean Grammar," *Mémoires de la Société Finno-Ougrienne,* vol. 82 (1939). SAMUEL E. MARTIN, *Korean Morphophonemics* (1954), is the most comprehensive technical description of sound alternations in Korean phonology. There is no general grammatical description of Korean in English available as yet. A.A. KHOLODOVICH, *Ocherk grammatiki koreiskogo iazyka* (1954), presents a description in Russian of modern Korean phonology, morphology, and syntax. RENE DUPONT and JOSEPH MILLOT, *Grammaire coréene*

(1965), assumes some familiarity with Korean writing and phonology on the part of the student but serves as a good reference work for modern Korean morphological forms and syntax, illustrated by many example sentences. BRUNO LEWIN, *Morphologie des koreanischen Verbs* (1970), contains a concise description of Korean verb forms and an alphabetically arranged list of the endings, with example sentences. The following textbooks constitute complete courses in the elementary spoken language and include explanations of phonology and grammar: FRED LUKOFF, *Spoken Korean,* 2 vol. (1945–47); CHANG HAI PARK, *An Intensive Course in Korean,* 2 vol. (1961–65); B. NAM PARK, *Korean Basic Course,* 2 vol. (1968–69); SAMUEL E. MARTIN and YOUNG SOOK C. LEE, *Beginning Korean* (1969). The method of writing and reading *Han'gŭl* is explained and illustrated in the first chapter of ANTHONY V. VANDESANDE and FRANCIS Y.T. PARK, *Myŏngdo's Korean '68,* pt. 1 and 2 (1967). A comprehensive dictionary of the spoken and written language compiled for western students of Korean is SAMUEL E. MARTIN, YANG HA LEE, and SUNG UN CHANG, *A Korean-English Dictionary* (1967), with the entries in *Han'gŭl* and their pronunciations recorded in transcription.

Japanese language. TAMAKO NIWA and MAYAKO MATSUDA, *Basic Japanese for College Students* (1964); and TEC COMMITTEE ON JAPANESE, *TEC Japanese: The Basic Course,* 2 vol. (1968–70), are both recommended for learning spoken Japanese. The latter is accompanied by magnetic tapes that record all the utterances in the books in very good standard pronunciation. ELEANOR H. JORDEN and H.I. CHAPLIN, *Beginning Japanese,* 2 pt. (1962–63), also has good examples of Japanese sentences. HOWARD HIBBETT and GEN ITASAKA, *Modern Japanese: A Basic Reader,* 2 vol. (1965), is a good book for learning written Japanese. ESTHER M.T. SATO, L.I. SHISHIDO, and M. SAKIHARA, *Japanese Now,* vol.1 (1982), is a modern text.

ROY ANDREW MILLER, *The Japanese Language* (1967), an up-to-date and comprehensive outline of the language with a good selective bibliography, includes chapters on the historical and geographical setting, genetic relationship, written systems, dialects, phonology, loanwords, "special and notable" utterances, and grammar and syntax. THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 2, *Linguistics in East Asia and South East Asia* (1967), includes very good accounts on various aspects of Japanese linguistics, with selective bibliographies.

The following works are in Japanese: SHINKICHI HASHIMOTO, *Chosakushū,* 13 vol. (1946–), includes the most fundamental and comprehensive articles on Japanese and Japanese linguistics. SANKI ICHIKAWA and SHIRO HATTORI (eds.), *Sekai Gengo Gaisetsu,* vol. 2 (1955), contains a good outline of Japanese. HIDEYO ARISAKA, *Kokugo On'inshi no Kenkyū,* rev. ed. (1957), is a collection of excellent papers treating various phonetic and phonological problems in the history of Japanese; his *Jōdai On'inkō* (1955) is a fundamental detailed study on the sounds of 8th-century Japanese. *Nihongo no Rekishi* (1947) contains good concise articles on the various periods of the history of Japanese by several authors. Among the dictionaries, the following are recommended: KOKUGOGAKKAI, *Kokugogaku Jiten* (1955), a dictionary of Japanese linguistics and philology; KYOSUKE KINDAICHI (ed.), *Jikai* (1952), a comprehensive dictionary of Japanese with accent notation; *Jidai-betsu Kokugo Daijitsen: Jōdaihen* (1967) lists exhaustively the vocabulary of Old Japanese; SENKICHI KATSUMATA, *New Japanese–English Dictionary* (1954).

Austronesian languages. The classic comparative work is OTTO DEMPWOLFF, *Vergleichende Lautlehre des austronesischen Wortschatzes,* 3 vol. (1934–37), in German, which followed a series of important earlier essays by RENWARD BRANDSTETTER, some of which appear in *An Introduction to Indonesian Linguistics,* trans. by C.O. BLAGDEN (1916). ISIDORE DYEN, *The Proto-Malayo-Polynesian Laryngeals* (1953), modifies some of Dempwolff's reconstructions, while the same author's *A Lexicostatistical Classification of the Austronesian Languages* (1963), presents a subgrouping of 214 languages. G.W. GRACE, *The Position of the Polynesian Languages Within the Austronesian (Malayo-Polynesian) Language Family* (1959), also contains an excellent discussion of previous comparative work. R.H. CODRINGTON, *The Melanesian Languages* (1885), and S.H. RAY, *A Comparative Study of the Melanesian Island Languages* (1926), are now important chiefly for the many grammatical sketches they contain. A. CAPELL, "Oceanic Linguistics Today," *Current Anthropology* 3:371–396, 422–428 (1962), reviews comparative work up to 1960. *A Pacific Bibliography,* 2nd ed. by C.R.H. TAYLOR (1965), and H.R. KLIENEBERGER (comp.), *Bibliography of Oceanic Linguistics* (1957), contain excellent bibliographies of earlier works. More-up-to-date information may be found in *Oceanic Linguistics* (semi-annual), chief among the several journals that publish work on Austronesian; in T.A. SEBOEK (ed.), *Current Trends in Linguistics,* vol. 8, *Oceania* (1971); and in C.F. and F.M. VOEGELIN, *Languages of the World: Indo-Pacific Fascicles* (1964–). See also OTTO C. DAHL, *Proto-Austronesian,* 2nd rev. ed. (1977), and *Early Phonetic and Phonemic Changes in Austronesian* (1981).

Papuan languages. Most earlier general studies have become obsolete. Comprehensive recent bibliographical and factual information may be found in D.C. LAYCOCK and C.L. VOORHOEVE, "History of Research in Papuan Languages," in *Current Trends in Linguistics,* vol. 8, *Linguistics in Oceania* (1971); and in S.A. WURM, "Papuan Linguistic Situation," *ibid.,* updated in *Papuan Languages of Oceania* (1972). Individual language descriptions are given in H. MCKAUGHAN (ed.), *Languages of the Eastern Family* (1971). Extensive descriptive and comparative materials on Papuan languages appear in the serial publications *Pacific Linguistics.*

Australian Aboriginal languages. A concise discussion of all aspects of the study of Australian Aboriginal languages, with extensive bibliography, is provided in S.A. WURM, *Languages of Australia and Tasmania* (1972); extensive information may be found in the contributions by A. CAPELL, G.N. O'GRADY, and S.A. WURM in T.A. SEBOEK (ed.), *Current Trends in Linguistics,* vol. 8, *Linguistics in Oceania* (1971); and in G.N. O'GRADY and C.F. and F.M. VOEGELIN, *Classification and Index of the World's Languages* (1977). Numerous articles and monographs on Australian languages appear in the serials *Australian Aboriginal Studies,* Canberra; *Pacific Linguistics,* Canberra; *Oceania,* Sydney; *Oceania Linguistic Monographs,* Sydney. See also the overview in BARRY J. BLAKE, *Australian Aboriginal Languages* (1981).

African languages. *General works:* ISTVAN FODOR, *The Problems in the Classification of the African Languages: Methodological and Theoretical Conclusions Concerning the Classification System of Joseph H. Greenberg* (1959); JOSEPH H. GREENBERG, *Studies in African Linguistic Classification* (1955), Greenberg's first classification of African languages, includes articles that originally appeared in the *Southwestern Journal of Anthropology;* "Africa As a Linguistic Area," in W.R. BASCOM and M.J. HERSKOVITS (eds.), *Continuity and Change in African Cultures,* pp. 15–27 (1959); *The Languages of Africa* (1963), a revision of Greenberg (1955), representing the most up-to-date version of the author's classification; SIR HARRY H. JOHNSTON, *A Comparative Study of the Bantu and Semi-Bantu Languages,* 2 vol. (1919–22), a pioneering study showing many significant similarities between Bantu and many West African languages, both in structure and vocabulary; RICHARD LEPSIUS, *Nubische Grammatik, mit einer Einleitung über die Völker und Sprachen Afrika's* (1880), a grammar of Nubian with an extensive introduction containing one of the earliest overall classification of African languages; CARL MEINHOF, *Die Sprachen der Hamiten* (1912), discusses a selection of languages considered by the author to be Hamitic, including Masai, and the linguistic characteristics that link them; FRIEDRICH MUELLER, *Grundriss der Sprachwissenschaft,* 4 vol. (1876–88), contains one of the earliest attempts at an overall classification of African languages; GEORGE PETER MURDOCK, *Africa: Its Peoples and Their Culture History* (1959); THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 7, *Linguistics in Sub-Saharan Africa* (1971); ARCHIBALD N. TUCKER and MARGARET A. BRYAN, *The Non-Bantu Languages of North-Eastern Africa* (1956), the most detailed survey of all the Chari-Nile languages and language groups, plus many others; *Linguistic Analyses: The Non-Bantu Languages of North-Eastern Africa* (1966), linguistic sketches of the languages included in Tucker and Bryan (1956); DIEDRICH WESTERMANN, *Die Sudansprachen, eine sprachvergleichende Studie* (1911), an attempt to demonstrate the interrelationship of five West African languages (Twi, Ga, Ewe, Yoruba, and Efik) and three from East Africa (Kunama, Nubian, and Dinka), thereby seeking to show the unity of all the Sudanic languages; *Die westlichen Sudansprachen und ihre Beziehungen zum Bantu* (1927), weakens the case for the interrelationship of all the Sudanic languages by showing a closer interrelationship between the Western Sudanic and Bantu languages; "Charakter und Einteilung der Sudansprachen," *Africa,* 8:129–148 (1935), revises the author's earlier views on the Sudanic languages; and with MARGARET A. BRYAN, *The Languages of West Africa* (1952), the most detailed survey of the languages of West Africa. Also useful is MELVIN K. HENDRIX, *An International Bibliography of African Lexicons* (1982).

Niger-Congo languages: Current Trends in Linguistics, vol. 7, *Linguistics in Sub-Saharan Africa (op. cit.),* is the most authoritative source on the Niger-Congo languages. This volume includes, among others, articles on West Atlantic, Mande, Voltaic, Kwa, Adamawa-Eastern, Benue-Congo (exclusive of Bantu), and Bantu. Significant light is thrown on the history and current research in these languages, and the "Check-List of African Language and Dialect Names" by WILLIAM E. WELMERS is an indispensable guide to the problem of variant names for languages. A summary of selected facts about the languages may be found in the series called *Handbook of African Languages,* sponsored by the International African Institute, especially the two volumes on Bantu by C.M. DOKE (1945 and 1954), the two on the same subject by MALCOLM GUTHRIE (1948 and 1953), and the one by DIEDRICH WESTERMANN and MARGARET A. BRYAN on the *Languages of West Africa* (*op. cit.*). The most important comparative study of the Bantu languages is presented in MALCOLM GUTHRIE, *Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages,* 4 vol. (1967–70). The following periodicals are devoted entirely to articles on African languages, of which the overwhelming majority are concerned with Niger-Congo: *Bantu Studies, Journal of African Languages, Journal of West African Languages, Afrika und Übersee, Studies in African Linguistics; African Language Studies,* and the *African Language Review.*

Chari-Nile languages: MERVYN W.H. BEECH, *The Suk: Their Language and Folklore,* with introduction by SIR CHARLES ELIOT (1911); MARGARET A. BRYAN, "The T/K Languages: A New Substratum," *Africa,* 29:1–21 (1959); "The *N/*K Languages of Africa," *J. Afr. Lang.,* 7:169–217 (1968); MORRIS GOODMAN, "Some Questions on the Classification of African Languages," *Int. J. Am. Linguistics,* 36:117–122 (1970); OSWIN KOEHLER, "Geschichte der Erforschung der Nilotischen Sprachen," *Afrika und Übersee,* vol. 28 (1955), a history of the investigation of the Nilotic languages, including "Nilo-Hamitic," and suggesting a new subgrouping; G.W. MURRAY, "The Nilotic Languages: A Comparative Survey," *Jl. R. Anthrop. Inst.,* 50:327–368 (1920), an attempt to prove the common ancestry of Nubian, Bari, Masai, and Shilluk; P.L. SHINNIE, *Meroe: A Civilization of the Sudan* (1967), includes a discussion of the Meroitic language; BRUCE G. TRIGGER, "Meroitic and Eastern Sudanic: A Linguistic Relationship?" *Kush,* 12:188–194 (1964); ARCHIBALD N. TUCKER, *The Eastern Sudanic Languages* (1940), compares, among others, the languages renamed Central Sudanic by Greenberg, and concentrates on the Moru-Madi group and Lendu; DIEDRICH WESTERMANN, *The Shilluk People: Their Language and Folklore* (1912), discusses the classification of the Nilotic languages.

Khoisan languages: (*Classification*): E.O.J. WESTPHAL, "A Reclassification of Southern African Non-Bantu Languages," *J. Afr. Lang.,* 1:1–8 (1962): "The Click Languages of Southern and Eastern Africa," *Current Trends in Linguistics,* 7:367–420 (1971); and "The Linguistic Prehistory of Southern Africa: Bush, Kwadi, Hottentot, and Bantu Linguistic Relationships," *Africa,* 33:237–265 (1963), with maps; L.W. LANHAM and D.P. HALLOWES, "Linguistic Relationships and Contacts Expressed in the Vocabulary of Eastern Bushman," *Afr. Stud.,* 15:45–48 (1956); for a comprehensive list of classified languages, see D.F. BLEEK, *A Bushman Dictionary* (1956). (*Distribution*): OSWIN KOEHLER "Die Khoe-sprachigen Buschmänner der Kalahari: Ihre Verbreitung und Gliederung," *Kölner Geogr. Arbeiten. Festschrift Kurt Kayser,* pp. 373–411 (1971), with map. (*Phonology*): C.M. DOKE, "An Outline of the Phonetics of the Language of the Chũ: Bushmen of North-West Kalahari," *Bantu Stud.,* 2:129–166 (1925); D.M. BEACH, *The Phonetics of the Hottentot Language* (1938). (*Grammar*): OSWIN KOEHLER, "Noun Classes and Grammatical Agreement in !Xũ," *Actes du VIII^e Congrès Intern. de Linguistique Africaine,* pp. 489–522 (1971); J.W. SNYMAN, *An Introduction to the !Xu (!Kung) Language* (1970); D.F. BLEEK, "Bushman Grammar: A Grammatical Sketch of the Language of the /xam-ka-!k'e'," *Zeitschrift für Eingeborenen-Sprachen,* vol. 19–20 (1928–30); L.F. MAINGARD, "The ≠Khomani Dialect of Bushman," in J.D.R. JONES and C.M. DOKE (eds.), *Bushmen of the Southern Kalahari* (1937); F. RUST, *Praktische Namagrammatik* (1965), based on material of H. Vedder and J. Olpp. (*Comparative studies*): OSWIN KOEHLER, "Studien zum Genussystem und Verbalbau der zentralen Khoisan-Sprachen," *Anthropos,* 57:529–546 (1962), deals with gender and verb system of Central Khoisan; D.F. BLEEK, "A Short Survey of Bushman Languages," *Zeitschrift für Eingeborenen-Sprachen,* vol. 30 (1939–40). (*Vocabulary*): D.F. BLEEK, *A Bushman Dictionary* (1956), a compilation of all language material available, except Hottentot; J.G. KROENLEIN, *Wortschatz der Khoi-Khoin-Namaqua-Hottentoten* (1889), on Nama. (*Oral literature*): W.H.I. BLEEK and L.C. LLOYD, *Specimens of Bushman Folklore* (1911), texts in /Xam with translation and annotations. (*General information*): I. SCHAPERA, *The Khoisan Peoples of South Africa* (1930), deals also with distribution and classification. (*Eyasian languages*): O. DEMPWOLFF, *Die Sandawe* (1916), with a chapter on language and vocabulary; D.F. BLEEK, "Traces of Former Bushman Occupation in Tanganyika Territory," *S. Afr. J. Sci.,* 28:423–429 (1931).

Eskimo-Aleut languages. For bibliography, demographic data, and literacy programs, see M.E. KRAUSS, "Eskimo-Aleut," in T.A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 10 (in prep.). Supporting evidence is given in: L.L. HAMMERICH, "The Western Eskimo Dialects," *Proc. 32nd Int. Congr. Americanists, Copenhagen 1956,* pp. 632–639 (1958); KNUT BERGSLAND, "The Eskimo Shibboleth *Inuk/Yuk,*" in *To Honor Roman Jakobson,* pp. 203–221 (1967), and "The Eskimo-Uralic Hypothesis," *Journal de la Société Finno-Ougrienne,* vol. 61 (1959). S.P. KLEINSCHMIDT, *Grammatik der grönländischen Sprache* (1851; restated by MORRIS SWADESH in HARRY HOIJER *et al., Linguistic Structures of Native America,* pp. 30–54, 1946), is a classic in the field. See also C.W. SCHULTZ-LORENTZEN, *Dictionary of the West Greenland Eskimo Language* (1927).

North American Indian languages. J.W. POWELL, "Indian Linguistic Families of America North of Mexico," *U.S. Bureau of American Ethnology, 7th Annual Report,* pp. 1–142 (1891), the first comprehensive classification; FRANZ BOAS, *Handbook of American Indian Languages,* 3 pt. (1911, 1922, and 1933–38), a classic introduction, with sketches of sample languages; HARRY HOIJER *et al., Linguistic Structures of Native America* (1946), a summary of work on language classification and sketches of languages; C.F. and F.M. VOEGELIN, *Map of North American Indian Languages,* rev. ed. (1966), presents the classification used in this article; T.A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 10, *North America* (1973), surveys of different aspects of the field, with extensive bibliographies (see esp. the valuable article of JOEL SHERZER, "Areal Linguistics in North America"); M.R. HAAS, *The Prehistory of Languages* (1969), a discussion of the principles in the historical study of American Indian languages; WALLACE CHAFE, "Estimates Regarding the Present Speakers of North American Indian Languages," *International Journal of American Linguistics,* 28:162–171 (1962), and 31:345–346 (1965), gives data used in the present article; EDWARD SAPIR, *Selected Writings in Language, Culture, and Personality* (1949), articles on the relationship of language and culture in aboriginal North America; B.L. WHORF, *Language, Thought, and Reality: Selected Writings* (1956), classic articles on American Indian language and world view.

Meso-American Indian languages. Few books ever appear treating the Meso-American Indian languages as a group, although many dictionaries and grammars for individual languages have been prepared. The most up-to-date overview can be gotten from three articles in the *Handbook of Middle American Indians,* vol. 5, *Linguistics* (1967): MARIA TERESA FERNANDEZ DE MIRANDA, "Inventory of Classificatory Materials," which is an annotated bibliography; MORRIS SWADESH, "Lexicostatistic Classification," in which the author applies glottochronology to the classification of all the Meso-American Indian languages; and ROBERT LONGACRE, "Systemic Comparison and Reconstruction," a review of what has been accomplished in the historical-comparative field to date. For linguistic characteristics of Meso-American languages, see TERRENCE KAUFMAN, "Areal Linguistics and Middle America," *Current Trends in Linguistics,* vol. 11 (1972). For Uto-Aztecan, see CHARLES F. and FLORENCE M. VOEGELIN and KENNETH L. HALE, *Typological and Comparative Grammar of Uto-Aztecan* (1962); and WICK R. MILLER, *Uto-Aztecan Cognate Sets* (1967). For Mayan and "new languages," see TERRENCE KAUFMAN, "Teco—A New Mayan Language," *Int. J. Am. Linguistics,* 35:154–174 (1969). For "external contacts," see RONALD D. OLSON, "Mayan Affinities with Chipaya of Bolivia," *ibid.,* 30:313–324 (1964) and 31:29–38 (1965).

South American Indian languages. CESTMIR LOUKOTKA, *Classification of South American Indian Languages* (1968), contains a full list of languages, an extensive and detailed bibliography, the location and classification of languages, and a map; J.A. MASON, "The Languages of South American Indians," in J.H. STEWARD (ed.), *Handbook of South American Indians* (U.S. Bureau of American Ethnology Bulletin 143), 6:157–317 (1950), has an extensive bibliography, general information on language groups, a discussion of classifications, and a map. See also N.A. MCQUOWN, "The Indigenous Languages of Latin America," *Am. Anthrop.,* 57:501–570 (1955), a critical appraisal of classifications with useful lists of languages and families, the location and classification of languages, and a map; C.F. and F.M. VOEGELIN, "Languages of the World: Native America," *Anthrop. Linguistics,* vol. 6, no. 6 and vol. 7, no. 7 (1964–65), general information on American Indian languages, much information on groups and individual languages, and a discussion of classifications; T.A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 4, *Ibero-American and Caribbean Linguistics,* pt. 2, "Linguistics of Non-Ibero-American Languages" (1968), critical surveys of work done during the last 20 years; M. SWADESH, "Afinidades de las lenguas amerindias," in *Akten des 34. Internationalen Amerikanisten Kongress,* pp. 729–738 (1964), and J.H. GREENBERG, "The General Classification of Central and South American Languages," in *Men and Cultures: Selected Papers of the 5th International Congress of Anthropological and Ethnological Sciences, Philadelphia 1956* (1959), two recent classifications of South American languages.

Sumerian language. ARNO POEBEL, *Grundzüge der sumerischen Grammatik* (1923), partly out of date, but still the only full grammar of Sumerian in all its stages; ADAM FALKEN-

STEIN, *Grammatik der Sprache Gudeas von Lagaš,* 2 vol. (1949–50), a very thorough grammar of the New Sumerian dialect, and *Das Sumerische* (1959), a very brief, but comprehensive survey of the Sumerian language; CYRIL J. GADD, *Sumerian Reading Book* (1924), outdated, but the only grammatical tool in English; SAMUEL N. KRAMER, *The Sumerians* (1963), provides a general introduction to Sumerian civilization.

Etruscan language. The following sources provide accurate and reliable information on Etruscan. M. PALLOTTINO, *Etruscologia,* 6th ed. (1968; Eng. trans., *The Etruscans,* 1955), the standard work; A.J. PFIFFIG, *Die etruskische Sprache* (1969), an excellent and complete statement of what is known about the Etruscan language; MURRAY FOWLER and R.G. WOLFE, *Materials for the Study of the Etruscan Language,* 2 vol. (1965), which contains the inscriptions with indexes of several kinds for easy reference—there is no grammatical information included.

Basque language. RENE LAFON, "La lengua vasca," *Enciclopedia lingüística hispánica,* vol. 1 (1960), perhaps the best short introduction to Basque, both descriptive and historical; HUGO SCHUCHARDT, *Primitiae linguae Vasconum,* 2nd ed. (1968), detailed commentary of an Old Basque text; P. LAFITTE, *Grammaire basque,* 2nd ed. (1962), a standard normative grammar; J. COROMINES, *Estudis de toponímia catalana,* 2 vol. (1965–70), presents new data on the survival of Basque dialects in the Middle Ages; J.M. LACARRA, *Vasconia medieval* (1957), authoritative review by an historian of the linguistic situation in and around the Basque country; LUIS MICHELENA, *Fonética histórica vasca* (1961), essay on the reconstruction of the phonological system of Proto-Basque; LUIS MICHELENA (ed.), *Textos arcaicos vascos* (1964), an annotated collection of documents from antiquity to 1700; RENE LAFON, *Le Système du verbe basque au XVIe siècle,* 2 vol. (1943), the best account of form and function of the Basque verb; A. TOVAR, *La lengua vasca,* 2nd ed. (1954; abridged Eng. trans., *The Basque Language,* 1957), and *The Ancient Languages of Spain and Portugal* (1961), a discussion of the problem of the position of Basque among these now extinct languages.

Pidgin. An extensive discussion and bibliography of the problems of pidgins and creoles is given in R.A. HALL, JR., *Pidgin and Creole Languages* (1966). A specialized grammar study is PIETER MUYSKEN (ed.), *Generative Studies on Creole Languages* (1981). See also HENRI TINELLI, *Creole Phonology* (1981).

The History of Latin America

As the term is generally understood, Latin America comprises the entire continent of South America, as well as Central America and Mexico (called Middle America), and the islands of the Caribbean. "Hispanic America" has often been suggested as a more suitable designation since it specifically indicates the Spanish and Portuguese heritage of the region. However, the Indian and black African heritage, as well as United States, British, and French cultural and colonial influence, nullify any advantages of such a change. Despite territorial contiguity and, for much of the area, ties of a similar culture, history, and aspirations for the future, the physiographic, climatic, economic, political, ethnic, and linguistic differences make the term Latin America as connoting a homogeneous region fall short of a true description. Only in deference to popular usage and for lack of a better term, the area remains Latin America.

This article is divided into the following sections:

The Pre-Columbian period

Men of Mongoloid stock entered the Americas at least 20,000 years ago, probably from Siberia by way of Alaska. They spread over North and Middle America, reached South America by way of the Isthmus of Panama, and, between 3,000 and 5,000 years ago, reached the tip of South America. Evidence of other major migration routes such as a trans-Pacific routes from Asia directly to Peru, Middle America, or British Columbia is, as yet, unsatisfactory. However, certain classical Maya art motifs, such as seated cross-legged figures, hint at the possibility of such an event. Older beliefs concerning land bridges and lost continents are now rejected.

Major Pre-Columbian cultural groups

Until recently the Pre-Columbian Indian population had been estimated at 15,500,000. However, recent researchers claim a population that large in central Mexico alone, and any definitive total must await further investigation. Indian cultures ranged from very primitive to highly developed. They may be divided into six major cultural groups (each composed or several language and political groups) as follows: (1) Pima-Pueblo (southwestern United States and northern Mexico); (2) Nahua-Maya (central Mexico, Yucatán, and Guatemala); (3) circum-Caribbean and semi-marginal (Central America south of Guatemala, Panama, northern and western Colombia, Venezuela west of the Orinoco, the Greater Antilles, and the western headwaters of the Amazon); (4) central Andean (western Ecuador, Peru, and northwestern Bolivia); (5) marginal tribes of South America (southern Chile, most of Argentina, eastern Bolivia, Uruguay, east-central Brazil northward to the Brazilian bulge); and (6) tropical forest and southern Andean tribes (occupying the Lesser Antilles, the Amazon basin, the western half of the Orinoco flood plain, the Guiana Highlands, and much of the east coast of Brazil south to the Uruguayan border, then west to Paraguay and central and northern Chile). Within these groups only the central Mexican, Maya, Chibcha, and Andean Indians had reached a high level of culture. The others varied from Paleolithic to late Neolithic.

Within their general cultural groupings, the numerous tribes into which the Indians were divided continuously fought for the control of better lands. Defeated tribes retreated into less hospitable areas, were gradually exterminated, or paid tribute to their conquerors. In a sense the Spanish and Portuguese conquest was but another phase in the struggle for control of land and labour. The Indian tribes were divided into hierarchies of chiefs, nobles, priests, commoners, servile labourers, and slaves. The idyllic life of the noble savage did not exist on the American continent; the struggle for existence and status was strong.

Diversity marked much of Indian life. In addition to the internecine struggle, there were wide variations in physique and language. Languages derived from many basic roots, and the dialects have been estimated to number 2,000 or more. Even the great empires were not linguistically united. The Incas forced Quechua as a lingua franca upon the people that they subdued. The Aztecs governed peoples with a mélange of dialects. Tribes more than 75 or 100 miles apart spoke different dialects if not different languages. Buildings varied from stick-and-thatch hovels to the magnificent stone palaces of the Incas and the temple-topped pyramids of the Maya. Primitive tribes wore little more than breechcloths to cover their nakedness while the upper classes in the high civilizations wore elaborate costumes, sometimes covered with exquisite featherwork. Advanced groups wove fine cotton and woollen cloth; others contented themselves with hides and bark cloth. Techniques and artistic sensitivity in the making of jewelry, pottery, and statuary covered a wide spectrum of differences. While the Maya possessed a system of writing, a vigesimal number system, and a calendar more accurate than that used by contemporary Europeans, the Incas relied on mnemonic devices to keep historical and statistical accounts. Other groups simply depended upon memory or pictographs.

Agriculture

Agricultural techniques and basic crops also varied greatly. Only corn (maize)—apparently first domesticated in Mexico—was raised throughout the New World, although the cultivation of beans and squash was also widely disseminated. While North American Indians planted mainly seed crops—corn, beans, and squash—the South American and West Indian peoples preferred tubers, such as cassava (manioc) and sweet potatoes. Because of its importance in the diet of the early conquistadors, cassava has been called the bread of the conquest. The white potato, native to the Andes, was raised in large quantities in the higher altitudes there where corn would not grow well. Perhaps the greatest deficiency in the Indian diet was the lack of a plentiful meat supply. Except for the guinea pigs and dogs, which were often raised for food, only in the Andean region, where the llama and alpaca were found, was there a dependable, if small, source of meat. Wild game and wild fowl were fairly abundant in some areas. The practice of cannibalism was widespread. Agricultural techniques ranged from fire agriculture (the milpa, in which corn was planted in burned-over forest land in a hole punched in the ground with a stick) to intensive irrigated, fertilized, and terraced agriculture. At best, Indian agriculture was an intensive hoe culture since there were no draft animals to draw a plow. Through their centuries of occupation of the New World the Indians had found the most fertile and productive areas for their techniques, and there the most populous empires developed. To a great extent, early Spanish conquest and settlement simply followed the lines of available food, labour, and cultivable land.

In addition to the more advanced peoples such as the Aztecs, several groups on a lower cultural level are worth noting. The Araucanians of Chile, a high agricultural group, were only partially conquered by the Incas. Soon after the Spanish conquest began, they acquired horses and used them to good account, tenaciously resisting the white invaders. Mounted Araucanians crossed the Andes into the southern pampas and exacted a heavy toll from the Spaniards. They were not completely pacified until the 19th century. In northern Mexico the Apaches and Comanches presented a similar problem. But the neighbouring Pueblos, except for a few sharp rebellions, were generally peaceful. The Guaraní of central South America, just becoming agriculturalists, proved rather docile. The Spaniards and Portuguese used them as farm labourers and slaves while the Jesuits organized them in missions. Brazilian Indians, such as the Tupí, were culturally among the most backward groups in the New World. Their low culture status undoubtedly contributed to their gradual disappearance, under the strain of working on the sugar plantations. Many Brazilian Indians fought the Portuguese to the bitter end; others aided the Europeans in establishing settlements.

In sum, several Indian peoples capitulated with hardly a struggle, while most resisted the Europeans to the best of their abilities, usually in vain. Their descendants, both pure-blooded and mixed, live scattered over the length and breadth of Latin America.

The colonial period

Spanish and Portuguese expansion into the New World was but a facet of their dynamic national policies. Portugal was a wealthy, expanding, trading nation with a large African and island empire. The leadership of Ferdinand and Isabella, the "Catholic Kings," transformed Spain into a national state marked by growing royal power and centralized administration. They kindled a wave of nationalist and religious fervour that eventually led to the expulsion of the Moors and Jews and carried Spaniards beyond the peninsula. Soon Spanish troops were fighting in Italy, the Netherlands, and central Europe to secure the domains of Charles I (better known to English-speaking people as Charles V, the Holy Roman emperor) and Philip II and to contain Protestantism. Overseas Spain fought to enlarge its domain and to convert the heathen Indian. As with the other nations of western Europe, Spain and Portugal were affected by the commercial revolution, and their merchants and crowns were interested in new trade contacts. An ardent, revived religious and nationalistic spirit, a drive for trade and land, a military caste seeking adventure and rewards, monarchs desirous of expanding and unifying their realms, all contributed to the Iberian conquest of the New World.

To say that "gold, glory, and gospel" were the major motivations of the Iberians is a misleading oversimplification. It is true that many individuals who participated in the incredible adventure of New World conquest sought plunder, and that priests converted the heathen. But to call this the cause of conquest is equivalent to saying that the western United States was settled by glory-mad cavalrymen, gold-hungry forty-niners and mountain men, zealous Calvinists, and Indian-cheating, treaty-breaking, landgrabbing frontiersmen and farmers. Iberians, during their 325 years of dominion, left an indelible impression upon the culture and life of the lands that they occupied. Spain preserved an Indian nobility, and many Spaniards took Indian wives. As soon as the wars of conquest and plunder were over, Spaniards and Portuguese staked out land for agriculture, founded cities, opened trading posts, and prospected for mines, advancing their nations' realms. They established the church of their faith and converted millions to Christianity. The evils of the Iberian conquest cannot be overlooked, but they were the horrors common to all European colonial systems. (Ed.)

THE SPANISH CONQUEST OF AMERICA

Search for a sea route to the East

Extensive European exploration of America was a by-product of European efforts in the 15th century to find a sea route to the East and thereby to end the monopoly of Italian and Levantine middlemen over the lucrative trade in spices and other Oriental products. Portugal took a decisive lead; it had the advantages of a long Atlantic seaboard with excellent harbours, a large number of fishermen and sailors, and an aristocracy that had learned to supplement its meagre revenue from the land with income from trade and shipbuilding. In 1497 a fleet of four Portuguese ships commanded by Vasco da Gama sailed from Lisbon on a voyage that inaugurated the age of European imperialism in Asia. After rounding the Cape of Good Hope, Vasco sailed into the Indian Ocean and up the coast of East Africa and then to Calicut, the great spice-trade centre on the west coast of India.

The voyages of Christopher Columbus. The search for a sea road to the Indies inspired more than one solution. An obscure Italian seafarer, Christopher Columbus, became convinced that it was possible to reach the East from Europe by sailing westward across the Atlantic and that his proposed route was shorter than the route around Africa, a conception that underestimated the size of the earth and overestimated the size and eastward extension of Asia. About 1484 Columbus, who then resided in Lisbon, offered to make a voyage of discovery for John II of Portugal, but the king turned it down. Columbus next turned to Castile, where, after eight years of discouraging delays and negotiations, Queen Isabella agreed to support the "Enterprise of the Indies." The contract made by the Queen with Columbus named him admiral, viceroy, and governor of the lands he should discover and promised him a generous share in the venture's profits.

The "Enterprise of the Indies"

On August 3, 1492, Columbus sailed from Palos with three small ships—the "Pinta," the "Santa María," and the "Niña"—manned not by the jailbirds of legend but by experienced crews under competent officers. The voyage was remarkably prosperous, with fair winds the whole way out. On October 12 they made landfall at an island in the Bahamas, which Columbus named San Salvador. Cruising southward through the Bahamas, Columbus came to the northeastern coast of Cuba, which he mistook for part of Cathay (China). Next he sailed eastward to explore the northern coast of an island that he named Española (Hispaniola), where he lost his flagship, the "Santa María." He then returned to Spain to report his supposed discovery of the Indies.

In response to Portuguese charges of encroachment on an area in the Atlantic reserved to Portugal by a previous treaty with Castile, King Ferdinand and Queen Isabella appealed for help to Pope Alexander VI, himself a Spaniard. The pontiff issued a series of bulls (1493) that assigned to Castile all lands discovered or to be discovered by Columbus and drew a line from north to south 100 leagues (345 miles) west of the Cape Verde Islands. To the west of this line was to be a Spanish sphere of exploration; to the east, Portuguese. To John II this line seemed to threaten Portuguese interests in the south Atlantic and the promising route around Africa to the East. Yielding to Portuguese pressure, Ferdinand and Isabella signed, in 1494, the Treaty of Tordesillas, establishing a boundary 270 leagues (930 miles) farther west.

Columbus returned to Española at the end of 1493 with a fleet of 17 ships carrying 1,200 colonists. The settlers soon gave themselves up to gold hunting and preying on the Indians; and Columbus, a foreigner, lacked the power and the personal qualities to control them. In 1496 he returned to Spain to report his new discoveries and to answer charges sent by disgruntled settlers.

The first two voyages had not paid their way, but the Spanish sovereigns still had faith in Columbus and outfitted a third fleet in 1498. On this voyage he discovered the island of Trinidad and the mouths of the Orinoco. He arrived in Española to find chaos. The Spaniards, disappointed in their hopes of quick wealth, blamed Columbus for their misfortunes and rose in revolt. To appease the rebels Columbus had to issue pardons and grant land and Indian slaves. Meanwhile, a stream of complaints against Columbus had caused the sovereigns to send out an agent, Francisco de Bobadilla, to supersede Columbus and investigate the charges against him. Bobadilla seized Columbus and sent him to Spain in chains. Although Isabella ordered Columbus' release, he never again exercised the functions of viceroy and governor in the New World. His monopoly on New World exploration and colonization ended as privileges were granted to other explorers. Columbus made one more voyage—in 1502–04—an unsuccessful search for a strait that would lead into the Indian Ocean. From Española, where he was not permitted to land, he crossed the Caribbean to the coast of Central America and followed it southward to the Isthmus of Panama. He finally departed for Española but was forced to land on Jamaica, where he and his men were marooned for a year.

Complaints against Columbus

When Nicolás de Ovando arrived in 1502 to serve as governor with an expedition of 2,500 persons, including 73 families, the colony numbered only 300. In a short while more than 1,000 of Ovando's group died. During the six years of Ovando's governorship conditions were stabilized. New cities were founded, gold output expanded, and food production increased under a system of forced Indian labour supplemented by slavery. With a permanent beachhead in the Indies, it was now possible (and the attrition of the Indian population of Española made it necessary) to begin the conquest of other islands in the West Indies as well as portions of the Caribbean coast. Having served its purpose as the advance base of the conquest, Española sank into decadence until its economic renaissance in the late 18th century.

Other early explorations. Other explorers followed and gradually made known the immense extent of the mainland coast of South America. In 1499 Alonso de Ojeda, accompanied by the Florentine Amerigo Vespucci, sailed to the mouths of the Orinoco and explored the coast of Venezuela. Vespucci personally directed another voyage, in 1501–02, under the flag of Portugal; this expedition, sent to follow up the discovery of Brazil by Pedro Álvares Cabral in 1500, explored the Brazilian coast and discovered the Río de la Plata before turning back. To Vespucci it was obvious that the landmass thought by Columbus to be a part of Asia was really a new continent—a "fourth part of the world." Vespucci's letters circulated widely in the early 1500s and gave him the fame of being the first European to set foot on the South American continent. A German geographer, Martin Waldseemüller, honoured Vespucci by assigning on a map the name America to the area of Brazil. The name caught on and presently was applied to the whole of the New World.

The naming of America

In 1500 Vicente Yáñez Pinzón explored the northern coast of Brazil from Cape São Roque to the Guianas, discovering the mouth of the Amazon, while in the same year Diego de Lepe explored the coast of the Guianas and Rodrigo de Bastidas in 1501 and 1502 the coast from Lake Maracaibo to Nombre de Díos, Panama (just south of modern Colón). The discovery of pearls along this coast led to the establishment of a lucrative industry based on exploiting Indian divers. Juan de la Cosa in 1504 explored the Gulf of Urabá (Darién) in the crook between Panama and South America.

Following Columbus' last voyage in 1502–04, which explored from Honduras south to Panama, Pinzón and Juan Díaz de Solís in 1506 explored the east coast of Yucatán and the Gulf of Honduras. Along with the stabilization of Española, these explorations led to a rash of colonizing activity. Juan Ponce de León began the conquest of Puerto Rico in 1508 and went on to discover Florida in 1513. Juan de Esquivel began the settlement of Jamaica in 1509, while in 1511 Diego Velázques de Cuéllar started the conquest of Cuba, founding Havana in 1515. Thus ended the first stage of Spanish exploration, the conquest and settlement of the islands of the Caribbean. Alonso de Pineda completed the mapping of the coast of the Gulf of Mexico in 1519 by sailing west from Florida to about modern Veracruz in Mexico and discovering the mouth of the Mississippi River on the way.

Beginnings of colonization

A growing shortage of Indian labour and a general lack of economic opportunities for new settlers on Española incited Spanish slave hunters and adventurers to explore and conquer Puerto Rico, Jamaica, and Cuba between 1509 and 1511. In the same period efforts to found colonies on the coast of northern Colombia and Panama, led by Alonso de Ojeda and Diego de Nicuesa, failed disastrously, and the remnants of two expeditions were united under the conquistador Vasco Núñez de Balboa to form the settlement of Darién on the Isthmus of Panama. Moved by Indian tales of a great sea, south of which lay a land overflowing with gold, Balboa led an expedition across Panama to the shores of the South Sea (Pacific Ocean). This aroused the jealousy of his father-in-law, Pedrarias Dávila, sent out by Charles V in 1514 as governor of the Isthmus, who contrived charges of treason and desertion; Balboa was tried, condemned, and beheaded in 1519. By this date exploration had pushed to the Pacific coast, and the city of Panama had been founded.

The exploration of the west coast of Central America and Mexico was done by Gaspar de Espinoza (1516–19) and Andrés Niño and Gil González Dávila (1522–23) who sailed north from Panama as far as Nicaragua. Pedrarias then began sending men northward into Central America where they finally encountered expeditions sent south overland from Mexico by Hernán Cortés.

The discovery of the South Sea helped confirm Vespucci's view that the so-called Indies formed no part of Eastern Asia. After 1513 the work of discovery centred on the search for a waterway to the East through or around the American continent. Ferdinand Magellan's circumnavigation of the globe in 1519–22 was too long to have commercial value. The net result was to enhance the value of America in Spanish eyes. Disillusioned with the dream of easy access to the East, Spain turned with concentrated energy to the task of extending its American conquests and to the exploitation of the human and natural resources of the New World.

The conquest of Mexico. In 1517 a slave-hunting expedition outfitted by Gov. Diego Velázquez, first governor of Cuba, explored the Yucatán Peninsula of Mexico, inhabited by Maya Indians whose pyramids, temples, and gold ornaments revealed a native culture far more advanced than any the Spaniards had previously encountered. Encouraged by the gold and other signs of Indian wealth brought back by the expedition, Velázquez outfitted a new venture, which he entrusted to his kinsman Juan de Grijalba. Grijalba coasted down the Yucatán Peninsula and, in June 1518, reached the limits of the Aztec Empire. From near what is now the port of Veracruz, Grijalba sent one ship back to Cuba with the gold that had been gained by barter with the coastal Indians and sailed on westward with three other ships, perhaps as far as the river Pánuco, marking the northern limits of the Aztec Empire. He returned to Cuba in November.

Cortés' expedition to Mexico

Velázquez sent a third expedition, with some 600 men under Hernán Cortés, to conquer the Mexican mainland in February 1519. Because Velázquez had not yet secured from the emperor Charles V an agreement authorizing conquest and settlement of the mainland, Cortés' instructions permitted him only to trade and explore.

Cortés' fleet first touched land at the island of Cozumel, off the coast of Yucatán. In March 1519, Cortés landed on the coast of Tabasco and defeated local Indians. In April he dropped anchor near the site of modern Veracruz, founding the town of Villa Rica de la Vera Cruz and appointing its first officials, into whose hands he surrendered the authority he had received from Velázquez. These officials then conferred on Cortés the title of captain general with authority to conquer and colonize the newly discovered lands (Cortés thus drew on Spanish medieval traditions of municipal autonomy to vest his disobedience with a cloak of legality).

Some days later the Aztec king Montezuma's (Moctezuma) ambassadors appeared in the Spanish camp. Apparently convinced that Cortés was the god Quetzalcóatl, who was returning to reclaim his lost realm, the envoys brought precious gifts—the finery of the great gods Quetzalcóatl, Tlaloc, and Tezcatlipoca; a gold disk in the shape of the sun, as big as a cartwheel; an even larger disk of silver, in the shape of the moon; and a helmet full of small grains of gold—while pleading with Cortés not to seek a meeting with their king. By plying Cortés–Quetzalcóatl with gifts Montezuma hoped to dissuade him from advancing into the interior and reclaiming the god's lost throne. Suavely Cortés informed the ambassadors that he had come a long way to see and speak with Montezuma, and he could not return without doing so.

Cortés, becoming aware of the bitter discontent of tributary towns with Aztec rule, began to play a double game. He encouraged the Totonac Indians of the coast to seize and imprison Montezuma's tax collectors, and he then promptly obtained their release and sent them to the king with expressions of regard and friendship. He took two other steps before beginning the march on the Aztec capital at Tenochtitlán (now Mexico City). First, he sent dispatches to the emperor Charles V seeking approval for his actions by describing the extent and value of discoveries; and second, to stiffen the resolution of his followers by cutting off all avenues of escape, he scuttled and sank all his remaining ships on the pretext that they were not seaworthy. Then, with his small army, he began a march on Tenochtitlán.

Advancing into the sierra, Cortés met and defeated in battle the Tlaxcalan Indians, traditional enemies of the Aztecs; the Tlaxcalans then formed an alliance with the white invaders. Next Cortés marched on Cholula, centre of the cult of Quetzalcóatl, where he claimed that the Cholulans were conspiring to attack him and staged a mass slaughter of the Cholulan nobility and warriors after they had assembled at his bidding in a great courtyard. Montezuma sent new envoys who brought rich gifts to Cortés but urged him to abandon his plan of visiting the Aztec capital; all of Montezuma's stratagems failed, and he welcomed Cortés at the entrance to the capital as a rightful ruler returning to his throne. The Aztec ruler even allowed himself to be kidnapped from his palace and taken to live as a hostage in the Spanish quarters.

Greeting of Cortés by Montezuma

A blunder on the part of his lieutenant Pedro de Alvarado thwarted Cortés' plan to use Montezuma as a puppet ruler. In Cortés' absence Alvarado ordered a massacre of the leading Aztec chiefs and warriors during a religious festival. This caused an uprising that forced the Spaniards to retreat to their own quarters. The tribal council deposed the captive Montezuma and elected a new chief, who launched vigourous attacks on the white invaders. In the midst of the struggle Montezuma died—killed by his own people as he appealed for peace, according to Spanish accounts; strangled by the Spaniards themselves, according to Indian sources. Threatened by a long siege and famine, Cortés evacuated Tenochtitlán at a heavy cost in lives. The surviving Spaniards and their Indian auxiliaries at last reached Tlaxcala.

In December 1520 Cortés, strengthened by Spanish reinforcements from Cuba, again marched on Tenochtitlán. A struggle began in late April 1521. On August 23, after a siege of four months, the last Aztec king, Cuauhtémoc, surrendered; and Cortés took possession of the ruins that had been the city of Tenochtitlán.

Once in control of central Mexico, Cortés dispatched expeditions to explore northward to the Gulf of California and southeast into Guatemala and Honduras. The Gulf of California was discovered in 1532 and Lower California the next year. Not until 1539 did Francisco de Ulloa explore the gulf and prove that Lower California was a peninsula, not an island. Juan Rodríguez Cabrillo, in 1542, explored the coasts of Lower and Upper California to about 40° N latitude.

Conquest of Central America

From the Valley of Mexico the tide of conquest flowed in all directions. Guatemala was reduced by Pedro de Alvarado; Honduras, by Cortés himself. In 1527 Francisco de Montejo began the conquest of Yucatán (in 1542 the Maya Indians rose in a revolt that was crushed with great slaughter). Meanwhile, expeditions from Darién subjugated the Indians of Nicaragua. Pedrarias' and Cortés' men soon met in armed clashes. A royal settlement relieved Pedrarias as governor of Panama in 1526 and named him governor of Nicaragua, a post he held until his death. Later, the entire area north of Panama was attached to Mexico as the captaincy general of Guatemala. Generally speaking, this area was neglected because of its lack of economic opportunities, and, with the exception of Guatemala where a brilliant nucleus of culture flourished, it became one of the backwaters of the Spanish Empire in America. Thus the two streams of Spanish conquest, both originally starting from Española, came together again.

For a brief time Cortés was undisputed master of the old Aztec Empire, renamed the "Kingdom of New Spain." The crown granted him the title of "Marquis of the Valley of Oaxaca" and the tribute and labour services of 23,000 Indian vassals. But the characteristic royal distrust of the great conquerors soon asserted itself. He was removed from his office of governor, and in 1539 he returned to Spain.

Cortés' conquests as well as those of his lieutenants were carried out with a minimum of bloodshed. In contrast, the adventurers who followed, penetrating northern Mexico and governing virtually unchecked, opened a chapter of wholesale pillaging and bloodletting.

Explorations of North America

Much of northern Mexico and the southwestern United States owes its exploration and eventual colonization to the adventures of Álvar Núñez Cabeza de Vaca. In 1536 he and three companions were found wandering in northern Mexico, the last survivors of Pánfilo de Narváez' expedition of 300 men shipwrecked in Florida eight years before. Núñez' stories of his journey from Florida to northwestern Mexico led to the organization of expeditions to explore the territories he had traversed. Hernando de Soto, financed by a fortune accumulated during his participation in the conquest of Peru, landed in Florida in 1539 with about 600 men and until his death in 1542 roamed over what is now the southeastern United States. He discovered no wealth; the survivors sank his body in the Mississippi and straggled back to Mexico. Fray Marcos de Niza and a guide named Esteban who had accompanied Núñez were sent by Viceroy Antonio de Mendoza in 1539 to retrace Núñez' steps in northern Mexico. Fray Marcos reported the existence of Seven Golden Cities, one of which he said was larger than Mexico City. A 300-man expedition led by Francisco de Coronado left Compostela in western Mexico in 1540; for two years they searched in vain for a golden empire, meanwhile wandering over New Mexico, Arizona, Texas, Oklahoma, Colorado, Kansas, and Nebraska. The failures of De Soto and Coronado, coupled with fruitless searches for a sea-to-sea passage via the Colorado River and the futile explorations of Rodríguez Cabrillo, dampened for the time being further interest in expansion in North America. Spanish policy turned toward settling and developing what had proved fertile and productive, exploiting the mines which were being discovered with increasing frequency and preaching the gospel. Only Florida was colonized in the 1560s to protect the Bahama Channel against pirates and European interlopers.

Silver mining in northwestern Mexico supported the establishment of several stable and prosperous communities that late in the 16th century were the scenes of renewed interest in what is now the southwestern United States. Juan de Oñate, nephew of one of the conquistadors of western Mexico, finally managed to plant the first colony in New Mexico, founding San Juan in 1598, and made a series of explorations ranging from the Colorado River east into Kansas. Oñate's hopes of finding great silver mines and using the labour of the sedentary Pueblo Indians proved abortive. New Mexico remained an isolated outpost surrounded by fierce Indians until the spread of missions over the next 200 years added this area to the cattleman's frontier.

Westward across the Pacific

After the voyage of Ferdinand Magellan, it was recognized that the New World itself was not only a barrier to the Far East (instead of being an extension of it) but also that it hid a tremendous ocean. Attempts to discover a strait from the Atlantic to the Pacific resulted only in the exploration of the west coast from Oregon south to Panama. The Spaniards also attempted to find a dependable route across the Pacific. Acting on royal orders to annex the Philippines to the viceroyalty of New Spain (Mexico), Viceroy Luis de Velasco dispatched an expedition in 1564 under Miguel López de Legazpi with Andrés de Urdaneta second in command. Legazpi remained to conquer the Philippines. Urdaneta, discovering the Japan Current, followed it to the California coast and thence south to Acapulco. The so-called Manila galleon soon plied regularly between Manila and Acapulco to trade Mexican silver for Eastern wares, particularly raw silk, which was woven in Mexico. In 1583 an *audiencia* was established in Manila subordinate to the viceroy in Mexico City.

The conquest of Peru. The conquest of Mexico challenged other Spaniards to match the exploits of Cortés and his companions and to discover the golden kingdom rumoured to lie beyond the South Sea. In 1519 Pedrarias founded the town of Panamá on the western side of the isthmus, which became a base for exploration along the Pacific coast. Three years later Pascual de Andagoya crossed the Gulf of San Miguel and returned with more information about a land of gold called Birú (Peru). Pedrarias then entrusted command of a voyage of discovery southward to Francisco Pizarro, who, in turn, recruited two partners—Diego de Almagro, an adventurer of obscure origin, and Hernando de Luque, a priest who acted as financial agent for the trio. Two preliminary expeditions (1524 and 1526) yielded enough finds of gold and silver to confirm the existence of the elusive kingdom. Pizarro then left for Spain to obtain royal sanction for the exploration of Peru. He returned to Panamá with the titles of captain general and *adelantado* (provincial governor) accompanied by his four brothers and other followers.

Pizarro in Peru

In December 1530 Pizarro again sailed southward from Panamá with a force of some 200 men and landed in the spring on the Peruvian coast. Civil war was raging in the Inca Empire. Atahuallpa, son of the late emperor Huayna Capac by a secondary wife, had defeated and imprisoned the lawful heir to the throne, Huáscar, and was moving toward the imperial capital of Cuzco when he received news of the arrival of white strangers. After an exchange of messages and gifts between Pizarro and Atahuallpa, the two armies advanced toward a meeting at the town of Cajamarca, high in the mountains.

Pizarro planned to win a quick, relatively bloodless victory by seizing the emperor as Cortés had done with Montezuma. When Atahuallpa and his escort appeared in the square of Cajamarca, they found it deserted, for Pizarro had concealed his men in some large buildings opening on the square. At a signal from Pizarro his soldiers, supported by cavalry and artillery, rushed forward to kill hundreds of terrified Indians and take the Inca prisoner.

Atahuallpa attempted to gain his freedom by offering to fill his spacious cell with gold higher than a man could reach. Pizarro accepted the offer; but when the room had been filled to the stipulated height, he informed the Inca that he was to remain in "protective custody." Pizarro proposed to use Atahuallpa as a puppet ruler to ensure popular acceptance of the new order, but he became convinced that Atahuallpa was organizing a resistance movement against the Spaniards. After a farcical trial, a Spanish court found the Inca guilty of polygamy, idolatry, and the murder of his brother Huáscar, and it condemned him to burn at the stake—a sentence commuted to strangling when he accepted baptism. Atahuallpa's enormous ransom of gold was divided among the Spaniards, and Hernando Pizarro was sent to Spain with the emperor Charles's share of the plunder. Hernando's arrival in Spain with his load of gold caused feverish excitement, and a new wave of Spanish fortune hunters flowed to the New World.

Diego de Almagro led an expedition to conquer Chile in 1535. Losing many of his force of Spaniard and Indian porters on the trip south, Almagro found a land without gold, peopled by fierce Indians. Dejected, he turned north, crossing the Atacama Desert to reach Peru. Only a handful of his men reached Cuzco, the old Inca capital, which Almagro seized and declared to be his.

Francisco Pizarro, posing as the defender of the legitimate Inca line, now proclaimed Huáscar's brother, Manco, as the new Inca. But a formidable revolt, organized and led by Manco himself, broke out in many parts of the empire. A large Indian army besieged Cuzco for 10 months but failed to take the city. Defeated by superior Spanish weapons and tactics and by food shortages in his army, Manco retreated to a fastness in the Andean Mountains; there he and his successors maintained a kind of Inca government-in-exile until 1572, when a Spanish military expedition entered the mountains, broke up the imperial court, and captured the last Inca, Topa Amaru, who was beheaded in a solemn ceremony at Cuzco.

The siege of Cuzco had barely ended when fighting broke out between one group of the conquerors, headed by the Pizarro brothers, and another, led by Almagro, over possession of the city. Before these struggles ended Francisco Pizarro had been murdered and two Almagros, father and son, had suffered death on the block. A new round of fighting began in 1544 when a new viceroy, Blasco Núñez Vela, arrived in Peru to proclaim the protective Indian legislation known as the "New Laws of the Indies." Led by Gonzalo Pizarro, the desperate conquistadors rose in revolt. The conquistadors defeated the new viceroy, and Núñez Vela was beheaded. The rebellion collapsed after the arrival of a new crown envoy, Father Pedro de la Gasca, who suspended the New Laws and offered pardons and rewards to all repentant rebels. Gonzalo Pizarro, however, resisted to the last and was captured and executed. Peace and order were not solidly established in Peru until the administration of Viceroy Francisco de Toledo, who arrived in 1569.

Conquest of Chile

Pedro de Valdivia, to whom Pizarro had given Almagro's concession, achieved the distinction of conquering Chile. In 1540, after terrible hardships, Valdivia's expedition reached Chile and began the slow conquest of the land from the fierce Araucanian Indians. After supporting Gasca (who reconfirmed his concession), Valdivia returned to Chile to prosecute his campaign. Chile offered neither wealthy civilizations to plunder nor mines to exploit. But the land was fertile, and the Indian population ample, if warlike. Valdivia granted land in large tracts, but often the would-be conquistador had to work with his Indians—and keep his weapons handy. Valdivia founded the cities of Santiago de Chile in 1541, Valparaíso in 1544, and Concepción in 1550. An expedition sent over the Andes into what is now western Argentina founded the cities of Mendoza (1561) and Tucumán in 1565.

Valdivia's most formidable opponents were the chiefs Lautaro and Caupolicán. According to tradition, Lautaro had learned Spanish military tactics while serving the conquerors as a stable boy. Lautaro defeated and captured Valdivia in 1553 and had him executed. In turn, Lautaro and Caupolicán were eventually captured and killed. Lautaro is now regarded as a Chilean national hero, immortalized in the epic poem *La araucana* by Alonso de Ercilla y Zúñiga. After a half century of war, peace was finally agreed upon when the Spaniards recognized the Bío-Bío River as the southern boundary of Chile, permitting the Araucanians to roam free to the south.

Conquest of Ecuador and Colombia

Peru served also as the nucleus for expansion eastward and northward. After the fall of the Inca Empire, Francisco Pizarro commissioned Sebastián de Belalcázar to secure the kingdom of Quito (Ecuador). Belalcázar, with the aid of Almagro, carried out his commission and founded the city of San Francisco de Quito in 1534. In 1540 Gonzalo Pizarro led an expedition from Quito over the Andes into the Amazon Valley in search of cinnamon. Finding themselves in sore straits, they built a number of boats to help forage for food. Francisco de Orellana took the boats and with a group of men sailed down the Amazon to the Atlantic. Whether Orellana was carried away by the current or deserted is still a moot question. Gonzalo returned to Quito with a handful of survivors. Meanwhile, Belalcázar was working north from Quito, incorporating northern Ecuador and southern Colombia into the Spanish domain. He eventually pushed his way to the highlands of Colombia, the homeland of the Chibcha Indians. There he met two other expeditions. One, led by Nicolaus Federmann, a German in the service of the banking house of Welser, had come southwest from Venezuela. The second expedition, however, under Gonzalo Jiménez de Quesada, which had come up the Magdalena River from the Caribbean coast, had reached the area first by a few days. The three men agreed to submit their case to the crown for judgment. Quesada was made a nobleman and alderman of the city of Santa Fe de Bogotá, which he had founded in 1538; Belalcázar was made governor of Popoyán in southern Colombia; Federmann received nothing.

Federmann's presence was the result of an unusual chapter in the early history of the Indies. After the failure of Ojeda's expedition to northern South America, the area was deserted except for pearl fishers, slave hunters, and an abortive attempt to found an ideal Indian state by Fray Bartolomé de Las Casas. In 1528 Charles V mortgaged Venezuela to the banking house of Welser of Augsburg, Germany, as a hereditary fief conditioned upon their developing the area. A complex formula was set up for dividing the produce of the mines, and the Welsers could enslave all the necessary Indian labour. Ambrosius Alfinger arrived in 1528 as governor. Failing to find gold, he set the colony upon an economic base of slave trading. Killed by Indians in 1531, he was succeeded by Georg Hohermut von Speyer as governor and Nicolaus Federmann as captain general. Federmann's expedition to the highlands near Bogotá was one of several fruitless attempts by him and Speyer to recoup the colony's fortunes. Charles V cancelled the contract in 1547. With the discovery of gold in 1560 the colony was placed on a more permanent footing, although it remained basically a peaceful out-of-the-way agricultural area.

Settlement of the Río de la Plata

The final major area of Spanish exploration, conquest, and settlement was the Río de la Plata. Juan Díaz de Solís discoverd the great estuary in 1516; Magellan explored it in 1520; Sebastian Cabot sailing up the Paraná River in 1526 ascertained that it was not a strait to the Pacific. Cabot picked up a number of silver trinkets which had filtered eastward from Peru. Believing himself to be close to their source, he named the region Río de la Plata (River of Silver). A full-scale attempt at colonization was made in 1535 by Pedro de Mendoza, who had received a royal patent to all lands from the Río de la Plata to the Straits of Magellan and then west to the Pacific. Theoretically, Chile had been granted twice: first to Almagro and then to Mendoza. Mendoza's expedition of 11 ships and 1,200 men soon came to grief on the inhospitable pampas occupied by hostile natives. Santa María de Buenos Aires, named after the patroness of the voyage, was the scene of starvation and cannibalism by the whites until it was destroyed by Indian attacks. Mendoza died at sea while returning to Spain for aid. The remnants of the expedition were led northward by Mendoza's lieutenants, Juan de Ayolas and Domingo Martínez de Irala, who founded the city of Asunción. Ayolas was lost in an unsuccessful attempt to reach Peru. Under Irala's guidance and by virtue of the docile nature of the Guaraní Indians, Asunción prospered. Large grants of land and Indians, as well as open polygamy, made Irala quite popular and, except for the years 1542–43 when Álvar Núñez Cabeza de Vaca ruled as governor, he governed until his death in 1556. Buenos Aires was refounded in 1580 by Juan de Garay to serve as a seaport for Asunción.

The first permanent settlements in modern Argentina were made in the northwest of that nation. Silver strikes at Potosí (in modern Bolivia) led to the establishment of ranches and handicraft industries in the area now comprising the states of Salta and Jujuy to supply the mines with horses, burros, meat, leather, and textiles. The settlement of Buenos Aires in 1580 marked the end of the second, or continental, stage of Spanish expansion. Spain now concentrated upon filling in the broad areas staked out during the first century and only slowly expanded the limits of its dominion. The three major areas of new expansion were northward from Mexico toward California and Texas, southward from the central valley of Chile, and into the plains surrounding the Río de la Plata and the Paraná River.

SPAIN'S COLONIAL EMPIRE

Political organization

The political organization of the Spanish Empire in America reflected the centralized, absolutist regime by which Spain itself was governed. In the Indies, as in Spain, there was a frequent contrast between the formal concentration of authority in the hands of royal officials and the actual exercise of supreme power on the local level by the great landowners.

The pattern of Spain's administration of its colonies was formed in the period between 1492 and 1550. The final result reflected the steady growth of centralized rule in Spain and the application of a trial and error method to the problems of colonial government. To Columbus, Cortés, Pizarro and other great expeditionary leaders, the Spanish kings granted sweeping powers that made these men practically sovereign in the territories they had won or proposed to conquer. Once the importance of these conquests had been revealed, however, royal jealousy of the great conquistadors quickly appeared; their authority was soon revoked or strictly limited, and the institutions that had been used in Spain to achieve centralized political control were transferred to the Western Hemisphere for the same end. By the mid-16th century the political organization of the Indies had assumed the definitive form that it was to retain, with slight variations, until late in the 18th century.

The Council of the Indies. The Council of the Indies, chartered in 1524, stood at the head of the Spanish imperial administration almost to the end of the colonial period. Although great nobles and court favourites were appointed to the Council, especially in the 17th century, its membership consisted predominantly of lawyers. Under the king, whose active participation in its work varied from monarch to monarch, it was the supreme legislative, judicial, and executive organ of colonial government. One of its most important functions was the nomination of all high colonial officials to the king. It also framed a vast body of legislation for the Indies—the famous Laws of the Indies, first codified in 1681—and sought to obtain detailed information on the history, geography, resources, and population of the colonies (the *relaciones,* which incorporated this information, represent a rich mine of materials for students of colonial Spanish America). Often staffed by conscientious and capable officials in the early Habsburg period, the quality of the Council's personnel tended to decline under the inept princes of the 17th century.

Principal royal agents

Viceroys, captains general, and audiencias. The principal royal agents in the colonies were the viceroys, the captains general, and the *audiencias* (high courts). Viceroys and captains general had essentially the same functions, differing only in the greater importance and extent of the territory assigned to the jurisdiction of the former; each was the supreme civil and military officer in his territory. At the end of the Habsburg era, in 1700, there were two great American viceroyalties—the viceroyalty of New Spain, with its capital at Mexico City, included all Spanish possessions north of the Isthmus of Panama; that of Peru, with its capital at Lima, embraced all of Spanish South America except the coast of Venezuela. Captains general, theoretically subordinate to the viceroys but in practice virtually independent of them, governed large subdivisions of these vast jurisdictions. Smaller subdivisions, called *presidencias,* were governed by *audiencias,* with the judge-president acting as governor but with military authority usually reserved to the viceroy.

A colonial viceroy enjoyed an immense delegated authority, which was augmented by the distance that separated him from Spain. By background he might be a lawyer or even a priest, but more often he came from one of the great noble and wealthy houses of Spain. A court modelled on that of Castile, a numerous retinue, and the constant display of pomp and circumstance testified to his exalted status. In theory his freedom of action was limited by the laws and instructions issued by the Council of the Indies, but recognition of the need to adapt the laws to existing circumstances gave him a vast discretionary power. The 16th century saw some able and even distinguished viceroys in the New World; in the 17th century, however, the quality of the viceroys declined (in 1695, for example, the viceroyships of Peru and Mexico were, in effect, sold to the highest bidders).

The *audiencia*

A viceroy or captain general was assisted in the performance of his duties by an *audiencia*, which served as his council of state. The joint decisions of viceroy and *audiencia* had the force of law, giving the *audiencia* a legislative character, roughly comparable to that of the Council of the Indies, in relation to the king. Although the viceroy was not obliged to heed the advice of the *audiencia,* its immense prestige and its right to correspond directly with the Council of the Indies made it a potential check on the viceregal authority. The crown thus developed a system of checks and balances that assured ample deliberation and consultation on all important questions but that also encouraged indecision and delay.

Provincial government. Provincial administration in the Indies was entrusted to royal officials, who governed districts of varying size and importance from their chief towns and who usually held the title of *corregidor.* Some were appointed by the viceroy; others, by the crown. They possessed supreme judicial and political authority in their districts and represented the royal interest in the town councils (*cabildos*). If not trained as a lawyer, a corregidor was assisted by a legal counsel (*asesor*) in the trial of judicial cases. Certain civil and criminal cases could be appealed from the municipal magistrates (*alcaldes*) to the *corregidor,* and from him to the *audiencia.*

Corregidores were of two kinds. Some presided over Spanish towns; others, *corregidores de indios,* administered Indian towns (*pueblos*), which paid tribute to the crown. A principal duty of a *corregidor de indios* was to protect the natives against fraud or extortion on the part of whites, but the *corregidor* was himself the worst offender in this respect. Perhaps the worst abuse of this authority arose in connection with the practice of *repartimiento*—the mandatory purchase of goods from the *corregidor* by the Indians of his district. Originally designed to protect the Indians from the frauds of private Spanish traders, the *corregidor's* exclusive right to trade with the Indians became a means for his speedy enrichment at the expense of the natives.

Restrictions on public officials. A series of regulations was designed to ensure good and honest performance on the part of public officials. Viceroys and *oidores* (members of an *audiencia*) were forbidden to engage in trade or own land within their jurisdictions or to accept gifts or fees; even their social life was restricted. All royal officials faced a judicial review (*residencia*) of their conduct at the end of their term of office. This took the form of a public hearing at which all who chose could appear before the "judge of residence" to present charges or testify for or against the official in question. At the end of the process the judge found the official guilty or innocent of all or part of the charges and handed down a sentence, which could be appealed to the Council of the Indies. Another device, the *visita,* was an investigation of official conduct, usually made unannounced, by a *visitador* especially appointed for this purpose by the crown; or, in the case of lesser officials, by the viceroy in consulation with the *audiencia.* As a rule, the *visita* was no more effective than the *residencia* in preventing or punishing official misdeeds.

Municipal government. The only colonial institution that satisfied to some degree local aspirations for self-

Royal control of local officials

rule was the town council, known as the *cabildo* or *ayuntamiento.* At an early date, however, the crown assumed the right to appoint the councilmen (*regidores*) and municipal judges; under Philip II and his successors it became the established practice for the king to sell these posts to the highest bidder, with a right of resale or bequest, on condition that a certain portion of the price be paid to the crown as a tax at each transfer. Throughout the colonial period the municipal governments were self-perpetuating oligarchies of rich landowners, mineowners, and merchants who frequently received no salaries and who used their positions to distribute municipal lands to themselves, to assign themselves Indian labour, and to serve the narrow interests of their class. Vigilantly supervised by the *corregidor,* who frequently intervened in its affairs, the *cabildo* soon lost such autonomy as it may have possessed in the early days.

Other government offices. The officials and agencies described above represented a small part of the apparatus of colonial government. There were large numbers of such officers as secretaries (*escribanos*), police officers, tax collectors of "the royal fifth," and *alcaldes* with special jurisdiction. Under Charles V such offices were often in the gift of high Spanish officials, who sold them to persons who proposed to go to the Indies to exploit their fee-earning possibilities; beginning with Philip II, many of these offices were withdrawn from private patronage and sold directly by the crown, usually to the highest bidder. In the second half of the 17th century, the sale of offices spread from fee-earning positions to higher salaried posts; as a result, in this period corruption became structural in the government of the Indies.

Power of the landowners

Although royal authority—represented by viceroys, *oidores, corregidores,* and other officials—was more or less supreme in the capitals and the surrounding countryside, the same was not true of the more distant regions. In such areas the royal authority was very remote, and the power of the great landowners was virtually absolute. On their large, self-sufficient estates they dispensed justice in the manner of feudal lords, holding courts and imprisoning peons in their own jails, and raising and maintaining their own private armies. Sometimes these powerful individuals combined their de facto military and judicial power with an official title that made them representatives of the crown in their vicinities. The contrast between the nominal concentration of power in the central government and the effective supremacy of great landowners on the local level was a legacy of the colonial period to independent Latin America, and it still remains a characteristic of the political life of many Latin American republics.

SPAIN'S INDIAN POLICY

From the first days of the conquest the Spanish government faced a problem of defining its attitude toward the American natives (wrongly called Indians) and of determining what relations should exist between the conquerors and the conquered. The Indian question had several facets; the first and most urgent was to harmonize the demand of the conquistadors for cheap Indian labour—frequently employed in a wasteful and destructive manner—with the crown's interest in preserving a large tribute-paying Indian population. There was a political issue, too; the Spanish kings were determined to prevent a concentration of land and Indians in the hands of colonists that might lead to the rise of feudal lords independent of royal authority.

The church also had a major interest in the Indian problem. If the Indians died out, there would be no pagan souls to save, and the good name of the church would suffer. Moreover, the church largely relied on Indian labour for the construction and service of its churches and monasteries in the Indies. For these reasons the church, in general, sympathized with and aided crown efforts to protect the Indians against excessive exploitation.

The Spanish dispute over Indian policy quickly assumed the form of a struggle of ideas. Spanish thought of the 16th century was strongly scholastic in character, and jurists and theologians argued over such questions as the nature and cultural level of the Indians, whether they were a subhuman race who might properly be conquered and made to serve the Spaniards, and the rights and obligations the papal donation of America to the Spanish monarchs conferred upon them. Behind these disputations, however, was a struggle between the crown, the church, the colonists, and the Indians themselves over who should control Indian labour and tribute—the foundations of the Spanish Empire in America.

The *encomienda*

Española was the first testing ground of Spain's Indian policy. Eager to prove to the crown the value of his discoveries, Columbus compelled the natives to bring in a daily tribute of gold dust. Later, yielding to the demands of rebellious settlers, Columbus distributed the Indians among them, along with the right to use the forced labour of the natives. This temporary arrangement, formalized in the administration of Gov. Nicolás Ovando and sanctioned by the crown became the *encomienda*, which consisted of the assignment to a colonist of a group of Indians who were to serve him with tribute and labour, while he assumed the obligation of protecting his Indians, paying for the support of a parish priest, and helping to defend the colony. In practice the *encomienda* in the West Indies proved to be a hideous slavery that decimated the Indian population of Española.

The first protests against this state of affairs were made by a group of Dominican friars who arrived in Española in 1510. King Ferdinand responded to their agitation by approving a code of Spanish–Indian relations—the Laws of Burgos (1512–13). These laws contained detailed regulations prescribing good treatment of Indian labourers; but these provisions were not enforced, and they did little more than sanction and regularize the existing situation.

A former *encomendero,* Bartolomé de Las Casas, joined the struggle against enslavement and mistreatment of the Indians. Las Casas argued that the papal grant of America to the crown of Castile had been made solely for the purpose of conversion, and it gave the Spanish king no temporal power or possession in the Indies; the Indians had rightful possession of their lands by natural law and the law of nations; all Spanish wars and conquests in the New World were illegal. Las Casas' mature program called for the suppression of all *encomiendas,* liberation of the Indians from all forms of servitude except a small tribute to the crown in return for its gift of Christianity, and even the restoration of the ancient Indian states and rulers. Over these states the Spanish king would preside as "emperor over many kings" in order to fulfill his sacred mission of bringing the Indians to the Catholic faith and the Christian way of life. Las Casas' proposals seemed radical, but objectively they served the royal aim of preventing the rise of a powerful colonial feudalism in the New World. Not humanitarianism but self-interest explains the partial official support that Las Casas' reform efforts received during the reign of Charles V (ruled 1516–56).

The New Laws of the Indies

The climax of royal intervention came with the proclamation of the New Laws of the Indies (1542), which appeared to doom the *encomienda.* They prohibited the enslavement of Indians, ordered the release of Indian slaves to whom legal title could not be proved, barred compulsory personal service by the Indians, regulated tribute, and declared that existing *encomiendas* were to lapse on the death of the holder.

The New Laws provoked a great revolt in Peru; in New Spain they caused a storm of protest by the *encomenderos* and a large part of the clergy. Under this pressure the crown again retreated; it reaffirmed the laws forbidding Indian slavery and forced labour, but the right of inheritance by the heir of an *encomendero* was recognized and extended by stages to a third, fourth, and, sometimes, even a fifth life. Thereafter, or earlier in the absence of an heir, the *encomienda* reverted to the crown. In the natural course of events, the number of *encomiendas* steadily diminished and that of crown towns increased.

Meanwhile, the economic value of the *encomienda* to the colonists was declining. They had lost the right to demand labour from their tributaries (1549); they had also lost their fight to make the *encomienda* perpetual. The heaviest blow of all, however, was the catastrophic decline of the Indian population in the second half of the 16th century. In New Spain, according to recent calculations, the Indian population dropped from about 25,000,000 in 1519 to slightly over 1,000,000 in 1605. Disease, especially disease of European origin, against which the Indians had no acquired immunity, was the major immediate cause; but overwork, social disorganization, and loss of will to live were largely responsible for the terrible mortality associated with the great epidemics and even with epidemic-free years.

As the number of their tributaries fell and their income declined proportionately, many *encomenderos* and other Spaniards began to engage in the more lucrative pursuits of agriculture, stock raising, and mining. The decline of the Indian population, sharply reducing the flow of foodstuffs to Spanish cities and mining centres, stimulated a rapid growth of Spanish estates (haciendas), producing grain and meat.

The *repartimiento*

A new system, the *repartimiento*—under which all adult male Indians had to give a certain amount of their time in rotation throughout the year to work in Spanish mines and factories, on farms and ranches, and on public works—replaced the *encomienda.* The crown thus sought to regulate the use of an ever-diminishing pool of Indian labour. The Indians received a token wage for their work, but the *repartimiento* was also essentially a disguised slavery. In Peru, where great numbers of Indians were conscripted for labour in the silver mines of Potosí and in the Huancavelica mercury mine, the *repartimiento* (there known as the *mita*) was particularly disastrous.

The *repartimiento* did not provide a dependable and continuing supply of labour, and Spanish *hacendados* turned increasingly to the use of so-called free labour. From the first, such labour was closely associated with the system of debt peonage, which helped to bind and hold workers in a time of rapid population decline. The heavy weight of tribute and *repartimiento* burdens on the ever-diminishing native population, and the contraction of Indian communal lands as a result of Spanish encroachments, caused many Indians to become farm labourers working for wages, mostly paid in kind. An advance of money or goods bound the peon to work for his employer until the debt was paid, a miracle that rarely occurred. Despite its later evil reputation, peonage had some advantages for many Indians; it usually freed an Indian from the recurrent tribute and *repartimiento* burdens, and it often gave him a piece of land that he could work for himself and his family.

If the hacienda offered many Indians a means of escape from intolerable burdens, it aggravated the difficulties of those who remained on their ancestral lands. The hacienda expanded at the expense of the Indian pueblo, absorbing whole towns and leaving others without enough land for its people when the population decline ended in the first half of the 17th century and a slow recovery began. The hacienda lured labourers from the pueblo, making it difficult for the Indian town to meet its tribute and *repartimiento* obligations.

Debt servitude assumed its harshest form in the numerous workshops (*obrajes*), producing cloth and other goods, that developed in many areas in the 16th and 17th centuries. Convict labour, assigned to employers by Spanish judges, was early supplemented by the "free" labour of Indians, who were often lured into these workshops by an offer of liquor or a small sum of money. Once inside the gates, they were never let out again.

Black slavery

Side by side with the disguised slavery of *repartimiento* and debt servitude existed black slavery. For a variety of reasons—including the fact that Spaniards and Portuguese were accustomed to holding Muslim black slaves and the belief that blacks were better able to support the hardships of plantation labour—Spanish defenders of the Indian did not display the same zeal on behalf of the enslaved Africans. The rapid rise of sugarcane agriculture in the West Indies in the early 1500s brought an insistent demand for black slave labour to replace the vanishing Indian. There arose a lucrative slave trade, chiefly carried on by foreigners under a system of contract (*asiento*) between an individual or a company and the Spanish crown. The high cost of slaves limited their use to the more profitable plantation cultures or to domestic servitude in the homes of the wealthy.

A small class of genuinely free, paid workers also came into existence at an early date; it included resident and migrant farm workers, miners, unskilled urban labourers, and skilled workers who practiced their trades in the Spanish towns, sometimes as journeymen, in Spanish-controlled guilds from which Indians, mestizos (persons of mixed Indian and European ancestry) and mulattoes were excluded as masters. Except for skilled workers, wages tended to remain at the subsistence level throughout the colonial period.

SPAIN'S COLONIAL ECONOMY

The conquest disrupted the traditional economy of the Indians and transformed the character and tempo of Indian economic activity. When the frenzied scramble for treasure had exhausted the available gold and silver objects, the *encomienda* became the principal instrument for extracting wealth from the vanquished. The Aztec and Inca peoples were accustomed to paying tribute to their rulers and nobility, but the Spaniards' demands were unlimited. Driven by visions of infinite wealth, the Spaniards exploited the Indians mercilessly.

Spanish colonial agriculture

As noted above, the Indian population decline, causing acute food shortages in the Spanish towns, created new economic opportunities for Spanish farmers and ranchers; moreover, it left vacant large expanses of land, which Spanish colonists occupied for wheat raising or as sheep and cattle ranges. By the end of the 16th century the Spanish-owned haciendas produced the bulk of agricultural commercial production and pressed ever more aggressively on the shrinking Indian sector of the colonial economy. The establishment of an entail (*mayorazgo*) assured perpetuation of the consolidated property in the hands of the owner's descendants.

Spanish colonial agriculture early produced wheat on a large scale for sale in such urban centres as Mexico City, Lima, Veracruz, and Cartagena, and maize, for the large Indian consumers' markets in Mexico City, Lima, and other cities. Sugar was brought from the Canary Islands to Española and became the foundation of the island's prosperity. From the West Indies, sugar quickly spread to Mexico and Peru. Sugar refining, with its large capital outlays for equipment and Negro slaves, was the largest scale enterprise in the Indies.

Wine and olives as well as sugar were produced in quantity in the irrigated coastal valleys of Peru. The silk industry flourished briefly in Mexico, but it soon declined because of labour shortages and competition from Chinese silk brought from the Philippines to the port of Acapulco. Other products cultivated by colonists on a capitalist plantation basis were tobacco, cacao, and indigo. Cochineal, a blood-red dye, was a unique Mexican and Central American export highly valued by the European cloth industry.

Spain enriched American economic life by introducing various domestic animals—chickens, mules, horses, cattle, pigs, and sheep. The mules and horses revolutionized transport, and the cattle and smaller domesticated animals greatly enlarged the continent's food resources, while providing hides for export to Spain and other European centres of leather manufacture, and hides and tallow for the domestic market, especially in mining areas. Sheep raisers found a large wool market in the textile factories that arose in many parts of the colonies.

In populous central Mexico, cattle trampled the Indian crops, causing untold damage, and the close grazing of sheep caused massive erosion by torrential rain on the slopes of valleys. By the end of the 16th century, however, the Mexican cattle industry had become stabilized; the exhaustion of virgin pasture lands, the mass slaughter of cattle for their hides and tallow, and the official efforts to halt grazing on Indian harvest lands, greatly reduced the herds. Gradually, the cattle ranches and sheep herds moved from the densely settled south and central areas to new permanent grazing grounds in the semiarid north.

A rapid increase of horses, mules, and cattle also took place in the empty grasslands (pampas) of the Río de la Plata, modern Argentina. The inhabitants of this area, forbidden to trade directly with the outside world and lacking precious metals or abundant Indian labour, traded illegally with Dutch and other foreign traders, who carried their hides and tallow to Europe; they also sent mules and horses, hides and tallow, to the mining regions of Upper Peru (Bolivia). Another centre of the cattle industry was the West Indies.

Mining

Mining, as the principal source of royal revenue in the form of the royal fifth (*quinto*) of all gold, silver, or other precious metals obtained in the Indies, received the crown's special attention and protection. Silver was the principal mining product. The great silver mine of Potosí in Upper Peru was discovered in 1545 and produced enormous quantities of the metal between 1579 and 1635; the rich Mexican silver mines of Zacatecas and Guanajuato were opened up in 1548 and 1558, respectively. The introduction of the patio process for separating the silver from the ore with mercury (1556) gave a great stimulus to silver mining. In the same period, important gold placers were found in central Chile and in the interior of New Granada (Colombia). The mining industry brought prosperity to a few and failure to the great majority.

Inefficient production methods, lack of capital to finance technological improvements, flooding, and similar problems, led to a sharp decline of silver production from about 1630 to the end of the century. This mining crisis depressed trade with Europe, since silver was the most important export item in Spanish America's balance of trade, and also had an adverse effect on internal commerce. Colonial agriculture and stock raising, which had expanded to satisfy the demand of the mining centres for grain, meat, hides, tallow, and work animals, also suffered.

Throughout the colonial period the majority of the natives continued, as before the Conquest, to supply their own needs for pottery, clothing, and other household goods. In the Spanish towns craft guilds, modelled on those of Spain, arose in response to the high prices for all Spanish imported goods. These guilds maintained control over the quantity and quality of production in luxury industries serving the needs of the colonial upper class. The 17th century saw a rapid growth of *obrajes,* many of which produced cheap cotton and wool goods for popular consumption. Most were privately owned, but some were operated by Indian communities to meet their tribute payments. Other primitive factories produced such items as soap, chinaware, and leather. A 17th-century depression, restricting colonial capacity to purchase imports, promoted this growth of colonial industry.

COMMERCE, SMUGGLING, AND PIRACY

Colonial trade

Control over all colonial trade, under the Royal Council of the Indies, was vested in the Casa de Contratación (House of Trade), established in 1503 in Seville. Commerce with the colonies was restricted until the 18th century to the wealthier merchants of Seville and Cádiz, who were organized in a guild that exercised great influence in all matters relating to colonial trade. Trade was concentrated in the three American ports—Veracruz in New Spain, Cartagena in New Granada, and Nombre de Dios (later Portobelo) on the Isthmus of Panama. The Seville merchant oligarchy and related merchant groups in the Indies (particularly the merchant guilds at Mexico City and Lima) deliberately kept the colonial markets understocked and in general played into each other's hands at the expense of the colonists, who were forced to pay exorbitant prices for all European goods acquired legally. Inevitably, the system generated colonial discontent and stimulated the growth of contraband trade.

To enforce a closed-port policy and protect merchant vessels against foreign attack, an elaborate fleet system was developed in the 16th century to convoy ships between American ports and to and from Spain. In the 17th century, as a result of Spain's economic decline and growing contraband trade, the fleet sailings became increasingly irregular.

Spanish industry, handicapped by its guild organization and technical backwardness, could not supply the colonies with cheap and abundant manufactures in return for colonial foodstuffs and raw materials. Prices to the colonial consumer also were raised by a multitude of taxes. As a result, the more advanced industrial nations of northern Europe sought to break into the large and unsatisfied Spanish-American markets, rejecting Spain's claim of domination over all the Western Hemisphere except the portion that belonged to Portugal. The foreign challenge to Spain's monopoly assumed the forms of smuggling, piracy, and colonization, as well as efforts to seize and occupy portions of the Indies.

Foreign challenge to Spain

England, under Queen Elizabeth, emerged as the principal threat to Spain's empire in America. Sir John Hawkins' slave-trading voyage to the West Indies in 1562 opened the English drive to penetrate the Spanish-American market and culminated in the near destruction of Hawkins' trading fleet by a Spanish naval force at Veracruz in 1568. English voyages of reprisal followed; they included the expedition of Francis Drake (1577), undertaken with the secret sponsorship and support of Queen Elizabeth, whose objects were to seize Spanish treasure ships, ravage Spanish colonial towns, and display English maritime prowess through a second circumnavigation of the globe.

In the 17th century foreign piracy and smuggling were supplemented by efforts to found colonies not only on the American mainland but in the Caribbean. The Dutch, at war with Spain with brief intervals since 1576, launched a military and commercial offensive against the Spanish West Indies; their principal instrument was the Dutch West India Company, organized in 1621. The Dutch captured the whole homebound Veracruz treasure fleet off the coast of Cuba in 1628. Capture of Curaçao, hard off the coast of Venezuela (1634), gave the Dutch an invaluable smuggling base and emboldened the French and English to seize both unoccupied and occupied Spanish islands—Barbados and St. Christopher, Martinique and Guadeloupe. In 1655 an English fleet, defeated in an effort to capture Santo Domingo, easily captured Jamaica. In the same period French corsairs began to settle the northwest corner of Española, virtually abandoned by Spaniards since 1605; by 1655 this region had become the French colony of Saint-Domingue, with a governor appointed by the trading Compagnie des Indes.

In this period piracy in the West Indies became a highly organized and large-scale activity often enjoying the open or covert protection of the English governors of Jamaica and the French governors of Saint-Domingue. Piracy entered a decline following the Treaty of Madrid (1670) between England and Spain, by which the British government agreed to aid in suppression of the corsairs in return for Spanish recognition of its sovereignty over the British West Indian islands. French buccaneers, however, continued activity until the Treaty of Rijswijk (1697), by which Spain formally recognized French possession of Saint-Domingue.

Smuggling

Pirates and privateers, however, inflicted fewer losses on Spain than those caused by foreign smugglers. Contraband trade steadily increased in the 16th and 17th centuries; and the European establishments in Jamaica, Saint-Domingue, and the lesser Antilles became bases for contraband trade with the Spanish colonies. Buenos Aires was another funnel through which Dutch and other foreign traders poured immense quantities of goods that penetrated as far as Peru. Smuggling flourished even at Seville and Cádiz where, by the end of the 17th century, French companies operating behind the facades of Spanish merchant houses dominated the legal trade with the Indies.

Spanish economists of the 17th century understood the principal cause of Spain's plight: its economic weakness. They offered sound criticisms and constructive proposals for reform. But they were powerless to change the course of Spanish policy, dictated by small mercantile and aristocratic cliques whose special interests and privileges were incompatible with the cause of reform.

THE CHURCH IN SPANISH AMERICA

Royal control over church affairs, in both Spain and the Indies, was founded on the institution of royal patronage (the *patronato real*). Under diplomatic pressure from King Ferdinand, in 1508 Pope Julius II accorded to Spain's rulers the exclusive right to nominate all Church officials, collect tithes, and found churches and convents in America, ostensibly to assist them in the work of converting New World pagans.

The missions. Beginning with Columbus' second voyage, one or more clergymen accompanied every expedition that sailed for the Indies; and they came in growing numbers to the conquered territories. The friars who came to America in the first decades after the Conquest were, in general, an elite group. The products of a revival of asceticism and discipline in the medieval church, and especially of a reform movement, this vanguard group frequently combined with missionary zeal a sensitive social conscience and love of learning.

Conversion of Indians

The friars converted prodigious numbers of natives. In Mexico, the Franciscans claimed to have converted more than 1,000,000 by 1531; when persuasion failed, pressures of various kinds, including force, were used to obtain conversions. (To facilitate the missionary effort the friars studied the native languages and wrote grammars and vocabularies. Scholars devoted themselves to preserving and recording the history, religion, and customs of the ancient Indian peoples.) The work of conversion was less than wholly successful. The result of the missionary effort was generally a fusion of pagan and Christian religious ideas. To this day Indians in such lands as Guatemala and Peru continue to perform rites dating from the time of the Maya and the Inca.

The clergy had to battle not only the Indian tendency toward backsliding but also divisions within their own ranks. A serious conflict was waged during the 16th century between the secular and the regular clergy. Finally, a royal decree of 1583 stated the principle that secular clergy were to be preferred over friars in all appointments to parishes. From first to last, however, the colonies were a scene of strife between groups of clergy over their fields of jurisdiction.

A gradual loss of sense of mission and of morale among the regular clergy also contributed to the decline of their intellectual and moral influence. By the late 16th century there were frequent complaints against the church's excessive number of convents and growing wealth. The principal sources of this wealth were legacies and other gifts from rich donors; invested in land and mortgages, it brought in more wealth. The last important order to arrive in Spanish America, the Society of Jesus (1572), had the largest number of rich benefactors and the most efficient administration.

Material wealth

Inevitably, this concern with material wealth weakened the ties between the clergy and the Indian and mixed-blood masses. Hand in hand with a growing materialism went an increasing laxity of morals (except among the Jesuits); concubinage became so common among the clergy of the later colonial period that it seems to have attracted little official notice or rebuke.

The missionary impulse of the first friars survived longest on the frontier, "the rim of Christendom." Franciscans first penetrated the great northern interior of New Spain; they also accompanied the Juan de Oñate expedition of 1598 into what is now New Mexico and dominated the mission field there until the end of the colonial period; and they were also found in such distant outposts of Spanish power as what are now Florida and Georgia. The early Jesuits worked among the Indians of California, were active in converting and pacifying the tough Chichimec Indians of the north central plateau of Mexico and had exclusive charge of the conversion of the Indian tribes of the northwest coast of Mexico. Following the expulsion of the Jesuits from the Indies in 1767, the Franciscans replaced them in directing missionary work in California.

The mission was one of three closely linked institutions—the other two being the presidio, or garrison, and the civil settlement—designed to serve the ends of Spanish imperial expansion and defense on the northern frontier. This three-pronged attack on the frontier was not very successful. Certain tribes on the northern frontier, such as the Apache of Arizona, New Mexico, and Texas and the Comanche of Texas, never were reduced to mission life. The missionaries had greater success among such sedentary tribes as the Pueblo Indians of New Mexico; but even among these peaceful tribes, Indian revolts and desertions were frequent. In 1680 the supposedly Christianized Pueblo Indians revolted, slaughtered the friars, and maintained a decade of tenacious resistance to Spanish efforts at reconquest. The civil settlements were no more successful. By the end of the colonial period there were only a few scattered towns on the northern frontier, and continuous Indian raids made life and property insecure. Ultimately, the whole task of defending and civilizing the frontier fell on a chain of presidios stretching approximately along the present border between the United States and Mexico. In the end Spain was forced to adopt a policy of neutralizing the Apache and Comanche by the periodic distribution of gifts to these warlike tribes. When the outbreak of the Wars of Independence stopped the flow of gifts, the hostile Indians drove through the useless line of presidios into the interior of Mexico.

The Jesuits in Paraguay. The most successful missionary effort, at least from an economic point of view, was that of the Jesuit establishment in Paraguay, which included more than 30 missions or reductions; these formed the principal field of Jesuit activity in America. Their strict discipline, centralized organization, and absolute control over the labour of thousands of docile Indians producing cotton, tobacco, hides, and other products enabled the Jesuits to make their missions a highly profitable business enterprise. Every effort was made to limit contact with the outside world. The life of the Indians was rigidly regimented in dress, housing, and the routines of work, play, and rest. Jesuit mission activity in the colonies ended as a result of a royal decree (1767) ordering the expulsion of the order from the colonies. Motives for this action included the conflict between the nationalistic church policy of the Bourbons and Jesuit emphasis on papal supremacy, suspicion of Jesuit meddling in state affairs, and belief that the Jesuit mission system constituted a state within a state.

Expulsion of the Jesuits

The Inquisition. The Inquisition formally entered the Indies with the establishment by Philip II of tribunals of the Holy Office at Mexico and Lima in 1569. Before that time its functions were performed by clergy who were vested with or assumed inquisitorial powers. Its great privileges, its independence of other courts, and the dread with which Spaniards generally regarded the charge of heresy made the Inquisition an effective check on "dangerous thoughts," whether religious, political, or philosophical. Most of the cases tried by its tribunals, however, dealt with offenses against morality or with such minor deviations from orthodox religious conduct as blasphemy. Like the Spanish Inquisition, the Inquisition in the Indies relied largely on denunciations by informers and employed torture to secure confessions. Indians were originally subject to the jurisdiction of Inquisitors but were later exempted because as recent converts of supposedly limited mental capacity they were not fully responsible for their deviations from the faith.

COLONIAL BRAZIL

Brazil's existence was unknown when the Treaty of Tordesillas of 1494, between Spain and Portugal, assigned to Portugal a large stretch of the South American coastline. In 1500 Pedro Álvares Cabral, who had been sent with a large fleet to India, was driven far off his course and touched on the Brazilian coast. Cabral claimed the land for Portugal and sent a report of his discovery to the king. Portugal's limited resources, already committed to exploitation of the wealth of Africa and the Far East, made impossible full-scale colonization of Brazil. But the presence of a valuable dyewood (brazilwood) attracted merchant capitalists, who obtained concessions to engage in the brazilwood trade with the Indians. A few settlers—some castaways, others *degredados* (criminals exiled from Portugal to distant parts of the empire)—arrived and were often well received by the local Indians.

In 1503 the Portuguese crown awarded Fernão de Noronha, a converted Jewish nobleman and merchant, a concession to cut brazilwood for three years. In time, settlements near the sites of modern Recife and Salvador were founded, an agricultural base was established, and numbers of Jews newly converted to Christianity emigrated to Brazil to escape the Inquisition's surveillance.

Both Noronha and the Portuguese crown profited from the arrangement, but incursions by Spaniards, Englishmen, and particularly Frenchmen dictated the necessity of defense by the mother country. Cristóvão Jaques in 1516 destroyed a French fleet leaving Brazil and established a trading post in what would later be the captaincy (now state) of Pernambuco; this post began the first cultivation of sugarcane in Brazil in 1521. Jaques' settlement was destroyed by the French in 1530. The Portuguese crown was then forced to reconsider its Brazilian policy. Towns and villages were growing in the harbours and islands along the coast; the gestation period was clearly over. Portugal then needed an aggressive policy to secure Brazil and prepare for further expansion.

In 1530 Martim Afonso de Sousa was sent with a fleet to drive out the French, to set up new settlements and administrative systems, and to recommend a form of colonization adapted to Brazil's needs. He carried out his assignment in excellent order: he established Portuguese claims to northern Brazil and founded settlements near Bahia (modern Salvador), Rio de Janeiro, and present-day Santos, the latter destined to become the entryway to the rich province of São Paulo. Brazil's new government, however, was set up before Martim Afonso returned to Lisbon.

Colonization. Portugal's heavy commitments in the spice-rich East forced its kings to assign to private individuals major responsibility for colonization of Brazil. In 1533, John III divided the Brazilian coastline into 15 parallel strips extending inland to the uncertain line of Tordesillas. These strips were granted as hereditary captaincies to a dozen individuals, each of whom agreed to colonize, develop, and defend his captaincy or captaincies at his own expense. The captaincy system represented a fusion of feudal and capitalist elements. The grantee was both a vassal owing allegiance to his suzerain, the king, and an entrepreneur who hoped to derive large profits from his estates and from taxes obtained from colonists to whom he gave land. These were large grants of land to an individual (called a *donatário*) endowed with ample economic and political privileges in return for his pledge to secure colonists and develop his grant. Along with tax exemptions and certain monopolistic privileges, they were given the right to enslave Indians for their estates and sell a certain number in Portugal each year. The *donatários* laboured hard at peopling and developing their grants; many sold or mortgaged their property in Portugal and moved to Brazil. Enormous sums of money were spent in transporting families and opening new agricultural districts. Few of the captaincies proved successful because few grantees possessed the capital and administrative ability required to attract settlers and defend against Indian attacks and foreign intruders. Only the captaincies at Olinda (that of Pernambuco) and São Vicente (near modern Santos) were financial successes, but the primary objective of securing the coast with permanent colonies was achieved. To supply the needed unity, John III named Tomé de Sousa a captain general in 1549 to head a central government. He founded the city of Bahia as his capital and it was an immediate success. Six Jesuits accompanied Sousa and became the forerunners of a most influential group in the new country.

Hereditary captaincies

Peopling Brazil was a difficult task. Forests taxed the strength of the most ambitious and with the stubborn Indian resistance, one *donatário* was moved to write the king: "The land you granted us in leagues we have had to conquer in inches." But the restless Brazilian frontiersmen continued to push north and south, as well as westward. The 17th century witnessed a tremendous burst of activity as the northern edge of the Brazilian bulge was settled to eliminate French and Dutch interlopers and Manaus, a thousand miles up the Amazon, was first settled in 1660. At the same time a wave of settlement moved up the Rio São Francisco as the Indians were driven out and cattle ranchers moved in. A third area of expansion was São Paulo in the south. Cut off by an escarpment from the coast where São Vicente cultivated sugar, São Paulo, on the plateau, found wealth by taking advantage of its position of easy access to the interior. Large slaving expeditions from São Paulo filled the demand for slaves with Indians when the African supply of blacks was insufficient. These expeditions, whose members were called *bandeirantes* (from a word meaning "banner" or "military company"), penetrated as far west as Paraguay, raiding the Jesuit missions and forcing the fathers to move west. Attempts by Portuguese Jesuits to control slavery proved unavailing, and the order in Brazil even had black slaves of its own. It was due to the activities of the *bandeirantes* that Brazil was able to claim its present southwestern boundaries far to the west of the line of demarcation. By the 1740s settlements extended as far south as Rio Grande do Sul, and by the 1780s as far west as the present Brazilian border. Of particular importance was the founding of Tabatinga (1780) on the upper Amazon which fixed Brazil's claims there. Black slaves were of prime importance, and by the end of the colonial period, they numbered about 1,000,000, constituting the backbone of plantation labour. Brazilian blacks were not always docile. Several large revolts and the setting up in the jungle of slave kingdoms called *quilombos* (one had 20,000 people) marked the colonial era. *Bandeirantes* made excellent slave catchers and broke up the runaways' *quilombos*.

Slave hunters. By the mid-16th century sugar had replaced brazilwood as the foundation of the Brazilian economy. Favoured by its soil and climate, the Northeast (now the states of Pernambuco and Bahia) became the seat of a sugarcane civilization featured by the large estates (*fazendas*), monoculture, and slave labour. The problem of labour was first met by raids on Indian villages, but Indian labour was unsatisfactory because the natives worked poorly and offered resistance ranging from attempts at flight to suicide. After 1550, planters turned increasingly to the use of black slave labour imported from Africa; but the supply of black slaves was often cut off or sharply reduced by the activity of Dutch pirates and other foreign foes, and the services of Brazilian slave hunters were in large demand throughout the colonial period. The most celebrated slave hunters were the *bandeirantes* from the upland settlement of São Paulo; often part Indian, they made slave raiding in the interior their principal occupation. As the Indians near the coast dwindled in numbers or fled before the invaders, the *bandeirantes* pushed ever deeper southward and westward, expanding Brazil's frontiers in the process.

Jesuit missionaries were the first to protest the enslavement and mistreatment of the Indians. The Jesuit program for the settlement of their Indian converts in *aldeas* ("villages"), where they would live under the tutelage of the priests, provoked the slave hunters and the planters. The Portuguese crown pursued for two centuries a policy of compromise that satisfied neither Jesuits nor planters. A decisive turn came in the mid-18th century under the Marques de Pombal, the foreign minister and later prime minister of José I, who, in 1759, expelled the Jesuits from Portugal and Brazil and secularized their missions. Pombal forbade Indian raids and enslavement and thus accepted the Jesuit thesis of Indian freedom; but, unlike the Jesuits, his policy did not segregate the Indians from the Portuguese community; it made them available as paid workers by the colonists, and it encouraged mingling between the two races. Meanwhile, the growth of the African slave trade, also encouraged by Pombal, reduced the demand for Indian labour and this brought a greater measure of peace to the Indians.

Rivals to the Portuguese. The dyewood, sugar, and tobacco of Brazil early attracted foreign powers. The French made sporadic efforts to entrench themselves on the coast, and, in 1555, they founded Rio de Janeiro as the capital of what they called La France Antarctique (Antarctic France). But French colonization in Brazil was weakened by Catholic–Huguenot strife at home, and, in 1567, the Portuguese ousted the French and occupied Rio de Janeiro.

French and Dutch incursions

Portugal itself fell under Spanish rule in 1580 when Philip II seized the vacant throne. Brazil was a fine addition to Philip's empire: its 17,000 Europeans controlled about an equal number of Indians and black labourers, who turned out 180,000 arrobas (one arroba = 25.36 pounds) of sugar a year. In Bahia alone there were more than 100 citizens with fortunes of $100,000 in modern purchasing power. Twenty years later the population was estimated at 25,000 whites, 18,000 Indians, and 14,000 black slaves and the capital investment at $22,000,000. Philip left Brazilian affairs in Portuguese hands; the Brazilians used their opportunity to push westward across the line of demarcation and trade with the Spanish colonies.

The Dutch posed a more serious threat to Portuguese sovereignty over Brazil. The Dutch West India Company seized and occupied (1630–54) the richest sugar-growing portions of the Brazilian coast. Ultimately, weakened by tenacious Brazilian resistance and a simultaneous struggle with England, the Dutch withdrew, taking their capital and the lessons they had learned in the production of sugar and tobacco and transferring both to the West Indies. Thus the Caribbean islands were soon competing with Brazilian sugar in the world market, with a resulting fall of prices, and by the end of the 16th century the Brazilian sugar industry was on the point of collapse.

Shift to the south. The discovery of gold (1690) in the southwestern region opened a new economic cycle and began a major shift of Brazil's economic and political centre from north to south. Large numbers of colonists from the Northeast, accompanied by their slaves and servants, swarmed into the mining area, causing an acute shortage of field hands in the older regions, which continued until the gold boom had run its course by the mid-18th century. In 1709 the mining region was elevated to the status of a captaincy, with the name Minas Gerais. A few years later, in 1729, came the discovery that certain stones in the area, hitherto thought to be crystals, were really diamonds; and many adventurers turned from gold to diamond washing.

By 1750 the river gold washings of Minas Gerais were nearly exhausted; the diamond district also suffered a progressive exhaustion of deposits. But the mineral cycle left a permanent mark in the form of new settlements in the southwest, not only in Minas Gerais but in what later became the provinces of Goiás and Mato Grosso—Brazil's far west. The mining decline also spurred efforts to promote the agricultural and pastoral wealth of the region. The southward movement of population and economic activity was formally recognized in 1763 when Rio de Janeiro became the seat of the viceregal capital.

Meanwhile, the Northeast experienced a partial revival based on increasing European demand for sugar, cotton, and other semitropical products. Brazilian cotton production made significant advances between 1750 and 1800, but then declined rapidly because of competition from the more efficient cotton growers of the United States. The beginnings of the coffee industry also date from the late colonial period.

The cattle industry

Cattle raising contributed to the advance of the Brazilian frontier and to the growing importance of the south. The intensive commercial agriculture of the coast and the concentration of population in such coastal cities as Baía and Recife created a demand for meat that gave an initial impulse to cattle raising. Because the expansion of plantation agriculture in the coastal zone did not leave enough land for grazing, the cattle industry had to move inland; and by the second half of the 17th century the penetration of the San Francisco Valley was well under way. Powerful cattlemen, with their herds, their *vaqueiros* ("cowboys"), and their slaves, entered the backcountry; drove out the Indians; and established fortified ranches and villages for their retainers. The cattle industry later expanded to the extreme southern region of Rio Grande do Sul, colonized by the government to defend against Spanish expansionist designs. The cattle industry provided meat for the coastal cities and mining camps, draft animals for the plantations, and hides for export to Europe.

Trade. During the Portuguese union with Spain (1580–1640), Brazil's commerce was restricted to Portuguese nationals and ships. The Dutch, who had been the principal carriers of Brazilian sugar and tobacco to European markets, responded with extensive smuggling and with a direct attack on the sugar-growing northeast. Following its successful revolt against Spain, Portugal made a trade treaty with England whereby British merchants were permitted to trade between Portuguese and Brazilian ports. English ships, however, frequently neglected the formality of touching at Lisbon and plied a direct trade with the colony. Because Portuguese industry was incapable of supplying the colonists with the requisite quantity and quality of manufactured goods, England provided cargoes of textiles and other manufactures. The decree of free trade of 1808, issued by Prince John following the flight of the Portuguese royal family from Portugal to Brazil after French invasion of his country, only confirmed the virtual English monopoly of trade with Brazil.

Government and church. The donatory system of government soon proved unsatisfactory, because few donatories could cope with the tasks of defense and colonization they had assumed. A governmental reform followed. In 1549 Tomé de Sousa was sent out as governor general to head a central colonial administration for Brazil; and Bahía, situated about midway between the flourishing settlements of Recife and São Vicente, became his capital. Gradually the hereditary rights and privileges of the donees were absorbed by governors appointed by the king. As the colony expanded, new captaincies were created. In 1763, as noted above, the governor of Rio de Janeiro replaced his colleague at Bahía as head of the colonial administration with the title of viceroy. In practice, however, his authority over the other governors was negligible.

Colonial administration

During the period of Spanish–Portuguese union their colonial policies were aligned by the creation, in 1604, of the Conselho da India, whose functions resembled those of the Spanish Council of the Indies. In 1736 a newly created ministry of Marinha e Ultramar (marine and overseas) assumed the functions of the Conselho; under the king, this body framed laws for Brazil, appointed governors, and supervised their conduct. The governor or viceroy combined in himself military, administrative, and even some judicial duties. His power tended to be absolute but was tempered by the constant intervention of the home government, which bound him with precise, strict, and detailed instructions; by the counterweights of other authorities, especially the high courts (*relações*), which were both administrative and judicial bodies; and the existence of special administrative organs, such as the intendencies created in the gold and diamond districts, which were completely independent of the governor. His authority was also diminished by the vastness of the country, the scattered population, the lack of social stability, and the existence of enormous landholdings in which the feudal power of great planters and cattle barons was virtually unchallenged.

The Senado de Camara, or municipal council, was the most important institution of local government. Elected either by a restricted property-owning electorate or chosen by the crown, its membership represented the ruling class of planters, merchants, and professional men; its authority was limited by frequent intervention of the royal judge (*ouvidor*), who usually combined his judicial functions with the administrative duties of *corregidor.* Generally speaking, the greater the size and wealth of the city, and the farther it was from the viceregal capital, the greater were its powers.

The evils of Spanish colonial administration—inefficiency, bureaucratic attitudes, slowness, and corruption—were also prominent in the Portuguese colonial system. In vast areas of the colony, however, administration and courts were virtually nonexistent. Outside the few large towns, local government often meant the rule of great landowners, for it was from their ranks that royal governors invariably appointed the *capitães môres,* or district militia officers. With unlimited power to enlist, command, arrest, and punish, the *capitão môr* became a symbol of despotism and oppression. (The feudalism that still dominates the Brazilian backcountry can be traced to these colonial origins.)

Some improvement, at least on the higher levels of administration, took place in the 18th century under Pombal, who abolished the remaining hereditary captaincies, reduced the special privileges of the municipalities, and increased the power of the viceroy. He sought to promote the economic advance of Brazil in order to rehabilitate Portugal, whose state was truly forlorn.

Church affairs

In Brazil, as in the Spanish colonies, church and state were intimately united. By comparison with the Spanish monarchs, however, the Portuguese kings were almost niggardly in their dealings with the church; but their control over its affairs, exercised through the *padroado*—the ecclesiastical patronage granted by the pope to the Portuguese king in his realms—was as absolute. Rome, however, maintained a strong indirect influence through the Jesuits, who were very influential in the Portuguese court until expelled from Portugal and Brazil in 1759.

With some honourable exceptions, notably that of the Jesuits, the tone of clerical morality and conduct in Brazil was low. The clergy provided such educational and humanitarian establishments as existed in the colony; and from its ranks—which were open to talents and even admitted indivduals of mixed blood, despite the formal requirement of a special dispensation—came most of the distinguished names in Brazilian colonial science, learning, and literature.

Masters and slaves. Race mixture played a decisive role in the formation of the Brazilian people. The scarcity of white women, the freedom of the Portuguese from puritanical attitudes, and the despotic power of great planters over Indian and black slave women, all gave impetus to miscegenation. Of the three possible combinations—white–black, white–Indian, black–Indian—the first was most common.

In principle, and to a considerable degree in practice, colour lines were strictly drawn. But the enormous number of mixed unions and the resulting large progeny, some of whom were regarded with affection by white fathers and provided with some education and property, inevitably led to some blurring of colour lines and to a fairly frequent phenomenon of "passing," with a tendency to classify individuals racially, if their colour was not too dark, on the basis of social and economic position rather than by physical appearance.

Slavery in Brazil

Slavery played as important a role in Brazil's organization as did race mixture in its ethnic make-up. Slavery corrupted both master and slave, festered harmful attitudes toward the dignity of labour, and retarded the nation's economic development. The virtual monopoly of labour by slaves sharply limited the number of socially acceptable occupations in which whites or free mixed bloods could engage. This gave rise to a large class of vagrants, beggars, "poor whites," and other degraded elements who would not or could not compete with slaves in agriculture and industry. Given the lack of incentive to work on the part of the slave, the level of efficiency and productivity of his labour was very low.

The 20th-century Brazilian sociologist Gilberto Freyre has emphasized the patriarchal relations between masters and slaves in the sugar plantation society of the Northeast. But the slaves described by Freyre were usually house slaves, who occupied a privileged position and whose situation was different from that of the great majority of slaves, who worked on the sugar and tobacco plantations of Bahía and Pernambuco. A royal dispatch of 1700 denounced the barbarity with which owners of both sexes treated their slaves. The very low rate of reproduction among slaves and frequent suicides speak eloquently about their condition; many slaves ran away and formed *quilombos* (settlements of fugitive slaves) in the bush. The most famous of these was the so-called republic of Palmares, whose destruction required a major military campaign in the 17th century.

The nucleus of Brazilian social and economic organization was the *fazenda,* based on black slavery and centred about the *casa grande* (the "big house"); it constituted a patriarchal community that included the owner and his family, his chaplain and overseers, his slaves, his sharecroppers (*obrigados*), and his *agregados,* or retainers. The system implied relations of mutual aid and a paternalistic interest in the welfare of the landowner's people; but it did not exclude intense exploitation of those people or the display of the most ferocious cruelty if they should dispute his absolute power.

Planter aristocracy

In the sugar-growing northeast, the great planters became a distinct aristocratic class. Most colonial towns were mere appendages of the countryside, dominated politically and socially by the rural magnates. These men often left supervision of their estates to majordomos and overseers, preferring to live in the cities. In the cities lived other social groups that disputed or shared power with the great landowners: high officials of the colonial administration; dignitaries of the church; wealthy professional men, especially lawyers; and merchants, usually European-born, who monopolized the export–import trade and financed the industry of the planters. The conflict between native-born landowners and European-born merchants, aggravated by nationalist resentment of upstart immigrants, sometimes led to armed struggle. An illustration is the War of the Mascates (1710–11) between Olinda, provincial capital of Pernambuco, dominated by the sugar planters, and its neighbouring seaport of Recife, controlled by the merchants.

THE BOURBON REFORMS AND SPANISH AMERICA

Reforms. After the War of the Spanish Succession (1702–13), Spain was left with a more manageable, more truly Spanish, empire consisting of the kingdom of Castile and Aragon, and the Indies, and then turned its attention to implementing a program of reform. The ensuing revival of Spain is associated with three princes of the House of Bourbon—Philip V (ruled 1700–46) and his two sons, Ferdinand VI (ruled 1746–59) and Charles III (ruled 1759–88). The work of national reconstruction reached its climax under Charles III, who attempted to revive Spanish industry and agriculture.

The outbreak of the French Revolution, which followed by a few months the death of Charles III in December 1788, brought the reform era effectively to a halt. Frightened by the overthrow of the French monarchy, Charles IV and his ministers turned sharply to the right; the leading reformers were banished or imprisoned, and the importation of French rationalist and revolutionary literature was forbidden. Yet the clock could not be entirely turned back, either in Spain or its colonies. Under Charles IV, for example, an expedition sailed from Spain (1803) to carry the procedure of vaccination to the Spanish dominions in America and Asia.

With colonial reform the Bourbons moved slowly and cautiously, because of the powerful vested interests identified with the status quo. There was no intention of giving more self-government to the colonists or of permitting them to trade more freely with the non-Spanish world. On the contrary, the Bourbons centralized colonial administration still further, with a view to making it more efficient, and their commercial reforms were designed to diminish smuggling and strengthen the exclusive commercial ties between Spain and its colonies.

Commercial and administrative reforms

Under the first Bourbons, efforts were made to check smuggling in the Caribbean by the use of *guardacostas,* which prowled the main lanes of trade in search of ships loaded with contraband. The depredations of these *guardacostas* led to English demands for compensation and finally to war between England and Spain in 1739.

The first Bourbons made few changes in the administrative structure of colonial government, contenting themselves with efforts to improve the quality of administration by more careful selection of officeholders. One major reorganization was the separation of the northern Andean region (present-day Ecuador, Colombia, and Venezuela) from the viceroyalty of Peru and its elevation to the status of a new viceroyalty, named New Granada, with its capital at Santa Fé (modern Bogotá). This change reflected a desire to provide better protection for the Caribbean coast and especially the fortress of Cartagena; it also reflected the rapid growth of population in the central highlands of Colombia. Within the new viceroyalty, Venezuela was named a captaincy general, with its capital at Caracas and virtually independent of Santa Fé.

Colonial reform, like domestic reform, reached its peak under Charles III. In 1765 commerce with the West Indies was thrown open to seven ports besides Cádiz and Seville; this, coming at a time when Cuban sugar production was beginning to expand, stimulated the island's economy.

This privilege was extended to other regions until, by a decree of free trade of 1778, commerce was permitted between all qualified Spanish ports and all the American provinces except Mexico and Venezuela, which were opened to trade on the same terms in 1789. Restrictions on intercolonial trade were also progressively lifted, but this trade was largely limited to non-European products. A major beneficiary of this was the Río de la Plata area, which, in 1776, was opened to trade with the rest of the Indies. Meanwhile, the Casa de Contratación steadily declined in importance until it closed its doors in 1790. A similar fate overtook the venerable Council of the Indies; most of its duties were entrusted to a powerful colonial minister appointed by the king.

Failure of colonial reforms

The entrance of new trading centres and merchant groups into the Indies trade, the reduction of duties, and the removal of irksome restrictions, had the effect of increasing commerce, reducing prices, and perhaps diminishing contraband. But the Bourbon commercial reform ultimately failed in its aim of reconquering colonial markets for Spain for several reasons—first, the opposition of the still dominant merchant oligarchs of Cádiz, who resisted intercolonial trade and efforts to replace French and English manufactures in the export trade to the colonies with noncompetitive Spanish products; second, Spain's continuing industrial weakness, which the best efforts of the Bourbons could not overcome; and, third, Spain's closely related inability to keep its sea-lanes to America open in time of war with England, when foreign traders again swarmed in Spanish American ports.

Colonial economic growth. Perhaps the most significant result of the Bourbon commercial reforms was the stimulus given to economic activity in Spanish America. To what extent this economic growth should be ascribed to the Bourbon reforms and to what degree it resulted from the economic upsurge in western Europe in the 18th century cannot be stated with certainty. Stimulated by the Bourbon reforms and a growing European demand for sugar, coffee, tobacco, hides, and other staples, production of these products rose sharply. But this increase in agricultural production resulted from more extensive use of land and labour rather than from the use of improved implements or techniques. The inefficient *latifundio* (a large, poorly cultivated estate), using poorly paid peon labour, and the slave plantation accounted for the bulk of agricultural production. Such natural disasters as drought or excessive rains easily upset the precarious balance between food supplies and population, producing frightful famines (as in 1785–87 in central Mexico, when thousands died of hunger or diseases).

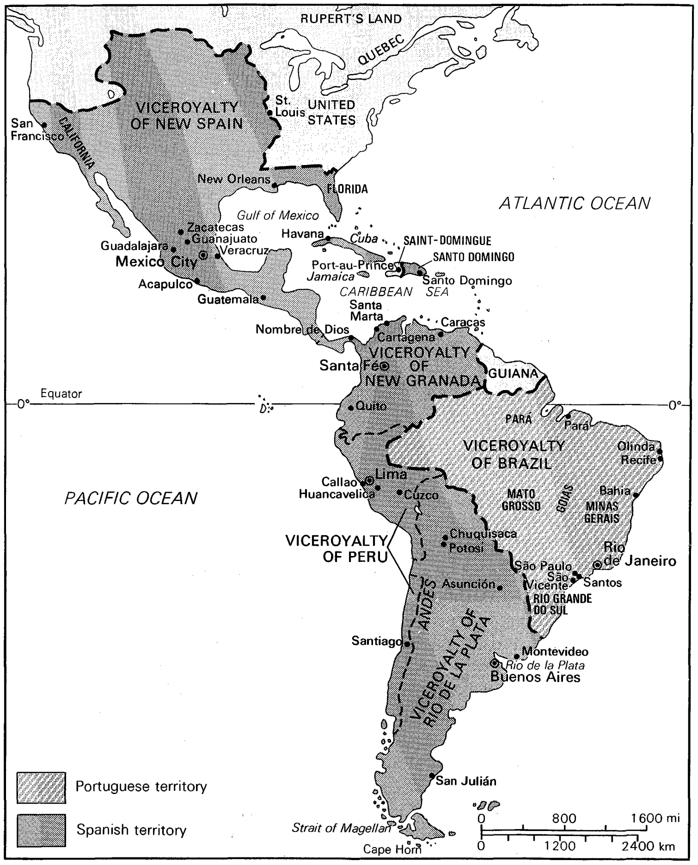
What sugar, cacao, and coffee were for the Caribbean area, hides were for the Río de la Plata. Rising European demand for leather and the opening of direct trade with Spanish ships in 1735 sparked an economic upsurge in the Plata area. By the end of the 18th century the *estancias* (cattle ranches) were often huge, with as many as 80,000 or 100,000 head of cattle. Meat gained in value as a result of the demand for salt beef. Markets for salt beef were found above all in the Caribbean area, especially Cuba, chiefly for feeding the slave population. The growth of cattle raising in La Plata was attended by concentration of land in ever fewer hands and took place at the expense of agriculture, which remained in a depressed state.

Silver mining revival

The 18th century also saw a strong revival of silver mining in the Spanish colonies. Peru and Mexico shared in this advance, but the Mexican mines forged far ahead of their Peruvian rivals. The crown, especially under Charles III, made serious efforts to encourage the industry by such means as tax reductions and the dispatch of foreign and Spanish experts to Mexico and Peru. These efforts were largely frustrated by the traditionalism of the mine owners, lack of capital to finance changes, and mismanagement. Yet the production of silver steadily increased; supplemented by the gold of Brazil, it helped to spark the industrial revolution in northern Europe and stimulated commercial activity on a worldwide scale. More especially, American silver helped the Bourbons meet the enormous expenses of their chronic wars.

Colonial manufacturing, on the other hand, began to decline late in the 18th century, principally because of competition from cheap foreign wares. The textile and wine industries of western Argentina and the textile production of Quito in Ecuador decayed as they lost their markets to lower-priced foreign wines and cloth. In the Mexican city of Puebla, production of chinaware slumped catastrophically between 1793 and 1802. Although Spain adopted mercantilist legislation designed to restrict colonial manufacturing, this legislation seems to have been a small deterrent to the growth of large-scale manufacturing. More important were lack of investment capital, preference for land and mining as fields of investment, and a semiservile system of labour that was harmful equally to the workers and to productivity.

Reprinted by permission of the publisher, from *People and Politics of Latin America,* copyright © 1955 by Ginn, 1945 by Ginn & Co., 1938 by Mary W. Williams, 1930 by Mary W. Williams; all rights reserved



Spanish and Portuguese America in 1784.

Decentralization. Under Charles III the work of territorial reorganization continued. The viceroyalty of Peru was further curtailed by the creation (1776) of the viceroyalty of the Río de la Plata, with its capital at Buenos Aires. In 1783 the establishment of a royal *audiencia* at Buenos Aires completed the liberation of the Plata provinces from the distant rule of Lima. The inclusion of Upper Peru in the new viceroyalty, with the resulting redirection of the flow of Potosí silver from Lima to Buenos Aires, signified a major victory for the landlords and merchants of Buenos Aires over their mercantile rivals of Lima.

The trend toward decentralization reflected not only a struggle against foreign military and commercial penetration but also an awareness of the problems of communication and government posed by the great distances between the various provinces. One indication of this was the greater autonomy and the increased number of the captaincies general in the 18th century. Venezuela was raised to a captaincy general in 1777, Chile, in 1778.

The intendant system

Between 1782 and 1790 the intendant system was introduced to the colonies. The intendants—provincial governors who ruled from the capitals of their provinces—were expected to relieve the overburdened viceroys of many of their duties, especially in financial matters. The offices of *corregidor* and *alcalde mayor,* holders of which were notorious as oppressors of the Indians, were replaced in Indian districts by that of *subdelegado,* nominated by the intendants and confirmed by the viceroys. Many intendants at the height of the reform era were capable, cultivated men who not only worked to increase economic activity and revenue collection but also promoted education and other cultural projects. Most *subdelegados,* however, soon became as notorious as their predecessors for their oppressive practices; a common complaint was that they compelled the Indians to trade with them, although forbidden by the Ordinance of Intendants. The failure of the Indian and mixed-blood masses to profit by the 18th-century economic advance—whose principal beneficiaries were Creole (American-born Spaniards) landowners, mine owners, and merchants—helped to produce popular revolts of 1780–81 (of Topa Amaru in Peru and of the *comuneros* in New Granada), which, although suppressed, shook Spanish power to its foundations in those areas.

Defense of the empire. Increased revenue was a major objective of the Bourbon commercial and political reforms, needed for strengthening the sea and land defenses of the empire. Spanish losses in the Seven Years' War (1756–63), and especially the loss of Havana and Manila to the English (1762), resulted in efforts to improve the defense system of the colonies. Fortifications of important American ports were strengthened and colonial armies were created. Regular units were stationed permanently in the colonies or rotated between peninsular and overseas service, and a colonial militia was filled by volunteers or conscripts. To make military service attractive to the Creole upper class, which provided the officer corps of the new force, the crown granted extensive privileges and exemptions to Creole youth who accepted commissions, thus adding to the lure of prestige and honours protection from civil legal jurisdiction and liability, except for certain specified offenses. This led to the development of a special officer class, which survives in many Latin American countries.

The wars of independence

BACKGROUND FOR THE WARS

The Bourbon reforms combined with the growing European demand for colonial products in the 18th century to bring material prosperity and other benefits to the upper class Creoles. The improvements and refinements introduced by enlightened viceroys and intendants made life in the colonial cities more attractive; educational reforms and opportunities widened the intellectual horizons of Creole youth. These advances, however, only enlarged Creole aspirations and discontents. Increased production strained against the trade barriers maintained by Spanish mercantilism. Awareness of their economic importance made intolerable to the Creoles their virtual exclusion from commanding posts in government and the church.

Creole discontent

The conflict of interests between Spain and its colonies found expression in the cleavage between Creoles and peninsular Spaniards. This cleavage was a major cause of the Spanish American wars of independence.

Enlightenment and revolutionary influences. Enlightenment ideas contributed to Creole restlessness and discontent. The forbidden writings of G.T. Raynal, Montesquieu, Voltaire, and Rousseau undoubtedly were read by educated Creoles; and scientific works based on the premises of Descartes, Leibnitz, and Newton circulated freely in the colonies and helped to spread Enlightenment ideas. By 1800 such influences had partly renovated the intellectual climate of Spanish America and had given a rationalist, pragmatic stamp to Creole thought.

The United States War of Independence contributed to the growth of "dangerous ideas" in the Spanish colonies. Spain was aware of the ideological as well as the political threat to its empire posed by the United States. After 1783 a growing number of U.S. ships touched legally or illegally at Spanish American ports, sometimes introducing such subversive documents as the writings of the revolutionists Thomas Paine and Thomas Jefferson.

The French Revolution exerted greater influence on the Creole mind. The Colombian Antonio Nariño translated and printed the French Declaration of the Rights of Man of 1789; he was sentenced to prison in Africa for 10 years but later led the successful independence movement in Colombia. The French Revolution soon took a radical turn, and the Creole elite became disenchanted with it as a model. The most important direct result of the French Revolution was a slave revolt in the French part of Haiti under Toussaint-Louverture and other black leaders; by January 1, 1804, black revolutionaries in Haiti had established the first liberated territory in Latin America, but their achievement dampened rather than aroused support for independence among the Creole elite of the colonies.

Beginnings of the independence movement. In 1806 the revolutionary Francisco de Miranda landed on the shores of his native Venezuela with a force of some 200 foreign volunteers, but his call for an uprising met with no response, and he hastily withdrew. The colonial independence movement might have remained ineffectual if not for decisions and actions by European powers with very different ends in view.

Spanish–French alliance

A major cause of the revolutionary crisis was Spain's involvement in the European wars unleashed by the French Revolution. Spain became France's ally in 1796, and English sea power promptly drove Spanish shipping from the seas, virtually cutting off communication between Spain and its colonies. The alliance with France had other results. An English fleet sailed with a regiment of soldiers against Buenos Aires (1806). The English soldiers entered Buenos Aires meeting only token resistance, but Creoles and peninsular Spaniards soon rallied to expel their unwanted liberators; a volunteer army attacked and routed the occupation force, capturing the English general and 1,200 of his men. A second British invasion was beaten back with heavy losses. Impressed by the tenacious defense, the British commander agreed to evacuate both Buenos Aires and the town of Montevideo. With this victory, the Creoles of Buenos Aires had tasted power and would not willingly relinquish it again.

In Europe, Napoleon gradually reduced Spain to a helpless satellite. In 1807 Napoleon obtained from Charles IV permission to invade Portugal through Spain; and the Portuguese royal family and court escaped to Brazil in a fleet under British convoy. Popular resentment at the French presence in Spain forced Charles IV and his son Ferdinand to abdicate; Napoleon tried to place his brother Joseph on the Spanish throne.

An insurrection against the French occupation forces began in Madrid and spread across Spain beginning in 1808; it led to the formation of a national Cortes, or parliament, which met in Cádiz from 1810 to 1814 under the protection of English naval guns. The constitution, approved by the Cortes in 1812, provided a limited monarchy and freedom of speech and assembly, and it abolished the Inquisition. But the Cortes made few concessions to the American colonies; it invited American delegates to join its deliberations but made clear that the system of peninsular domination and commercial monopoly would remain essentially intact.

Creole plans for gaining power

In Spanish America events had transformed the remote prospect of independence into a realistic goal. Confident that the French armies would crush all opposition, Creole leaders prepared to take power with the pretext of loyalty to the "beloved Ferdinand." The confusion caused among Spanish officials by the arrival of rival emissaries who proclaimed Ferdinand or Joseph Bonaparte the legitimate king of Spain aided the Creole plans. In the spring of 1810, with the fall of Cádiz seemingly imminent, the Creole revolutionaries moved into action; charging viceroys and other royal officials with doubtful loyalty to Ferdinand, they organized demonstrations in Caracas, Buenos Aires, Santiago, and Bogotá that forced these authorities to surrender control to Creole-dominated local juntas. Their hopes of a peaceful transition to independence, however, were doomed to failure; their claims of loyalty did not deceive the groups truly loyal to Spain, and fighting soon broke out between patriots and loyalists.

THE INDEPENDENCE OF SOUTH AMERICA

The Latin American struggle for independence lacked a unified direction or strategy that was caused not only by the vast distances and other geographical obstacles to unity, but also by the economic and cultural isolation of the various Latin American regions. Moreover, the Latin American movement for independence lacked a strong popular base—the Creole elite, itself part of an exploitative white minority, feared the Indians, blacks, and oppressed castes and usually sought to keep their intervention in the struggle to a minimum. This lack of regional and class unity helps explain why it took Latin America so long to achieve independence.

The struggle for independence had four main centres. In Spanish South America there were two principal theatres of military operations—one flowed southward from Venezuela, the other ran northward from Argentina; these two currents joined at Peru, the last Spanish bastion on the continent. The third centre, Brazil, achieved its own swift and relatively peaceful separation from Portugal. Finally, Mexico had to travel a difficult, devious road toward independence.

Bolívar's involvement in the movement

Spanish America. Simón Bolívar of Venezuela is the symbol and hero of the struggle for independence in northern South America. Soon after his return to Venezuela from a visit to Europe (1804–07), he became involved in conspiratorial activity directed at overthrow of the Spanish regime. In April 1810, the Creole party in Caracas forced the captain general to abdicate, and a Creole-dominated junta took power. In 1811 a Venezuelan congress proclaimed the country's independence and framed a republican constitution that abolished special privileges and Indian tribute, but retained black slavery and made Catholicism the state religion. Meanwhile, the veteran revolutionary Francisco de Miranda had returned from England and assumed command of the patriot army.

Differences soon broke out between Miranda and his young officers, especially Bolívar. Amid these disputes came the earthquake of March 26, 1812, which caused great loss of life and property in Caracas and other patriot territory, but spared the regions under Spanish control. The royalist clergy proclaimed this disaster a divine retribution against the rebels. A series of military reverses completed the discomfiture of the revolutionary cause. With his forces disintegrating, Miranda negotiated a treaty with the royalist commander and then tried to flee the country, taking with him part of the republic's treasury. Bolívar and some of his comrades, regarding Miranda's act a form of treachery, seized him before he could embark and turned him over to the Spaniards. Miranda died in a Spanish prison four years later.

Bolívar was saved from a Spanish reaction by the influence of a family friend. Under a safe conduct, he departed for Colombia, which was still partially under patriot control. Given command of a small force to clear the Magdalena River of enemy troops, he employed a strategy featured by swift movement and aggressive tactics; he also judged his soldiers on merit without regard to social background or colour. His success gained Bolívar the rank of general in the Colombian army and won him the approval of a plan of his for liberation of Venezuela.

In a forced march of three months, Bolívar led a force of 500 men across jungles and swamps toward Caracas; as he approached the capital, the Spanish forces withdrew. He entered Caracas in triumph and received from the city council the title of "liberator"; soon afterward the congress of the restored republic voted to grant him dictatorial powers.

Bolívar's success was short-lived. The fall of Napoleon, in 1814, brought Ferdinand VII to the Spanish throne, released Spanish troops for use in America, and heartened the royalists. Meanwhile, the *llaneros,* or cowboys, of the Venezuelan plains joined the royalist cause. A mass of *llaneros* invaded Caracas, crushing all resistance; and in July 1814 Bolívar hastily abandoned the city and retreated toward Colombia with the remains of his army.

Bolívar found Colombia on the verge of chaos; despite the imminent threat of a Spanish invasion, the provinces quarrelled with each other and defied the authority of the weak central government. Concluding that the situation was hopeless, Bolívar left in May 1815 for the British island of Jamaica. Meanwhile, a Spanish army landed in Venezuela, completed the reconquest of the colony, and then sailed to lay seige to Cartagena; the city surrendered in December, and the rest of Colombia was pacified within a few months. Of all the Spanish American provinces, only Argentina remained in revolt.

Bolívar's "Letter from Jamaica"

Bolívar had an unshakable faith in the triumph of independence. From Jamaica he wrote a famous document—"The Letter from Jamaica"—in which he affirmed his faith and argued that monarchy was foreign to the genius of Latin America; only a republican regime would be accepted by its peoples. Bolívar boldly forecast the destiny of the different regions, taking account of their economic and social structures (Chile, for example, seemed to him to have a democratic future; Peru, on the other hand, was fated to suffer dictatorship because it contained gold and slaves).

From Jamaica, Bolívar went to Haiti, where he received a sympathetic hearing and the offer of some material support. After two efforts to gain a foothold on the Venezuelan coast failed, Bolívar decided to establish a base in the Orinoco River Valley, distant from the centres of Spanish power. Roving patriot bands still operated in this region, and Bolívar hoped to win the allegiance of the *llaneros,* who were becoming disillusioned with their Spanish allies. In September 1816 Bolívar sailed for the Orinoco River Delta and made his headquarters at the town of Angostura (modern Ciudad Bolívar).

The patriot guerrilla bands accepted Bolívar's leadership, and he also gained the support of the *llaneros.* The end of the Napoleonic Wars had idled a large number of British soldiers; and many of these veterans came to Venezuela, forming a British legion that distinguished itself in battle on the patriot side. Bolívar was also helped by English merchants, who made loans that enabled him to secure men and arms, and by the mulish attitude of Ferdinand VII, whose refusal to make any concessions to the colonists caused the British government to lose patience and regard favourably the prospect of Spanish American independence.

Liberation of Colombia and Venezuela

On the eve of the decisive campaign of 1819, Bolívar summoned to Angostura a congress that vested him with dictatorial powers. His strategy for the liberation of Venezuela and Colombia was to strike a blow at Spanish forces from a completely unexpected direction; he advanced with an army of some 2,500 men along the Orinoco and Arauco rivers across the plains and then ascended the Colombian Andes until he reached the plateau where lay Bogotá. On the field of Boyacá the patriot army surprised and defeated the royalists in a short, sharp battle and entered Bogotá. He then prepared for the liberation of Venezuela; in June 1821 patriot troops crushed the last major Spanish force in Venezuela at Carabobo. Save for some coastal towns and forts held by beleaguered royalists, Venezuela was free.

The independence of Spanish America remained precarious as long as the Spaniards held the immense mountain bastion of the central Andes. While Bolívar prepared an offensive from Bogotá against Quito, he sent his lieutenant José Antonio de Sucre by sea from Colombia's Pacific coast to seize the port of Guayaquil. Even before Sucre arrived, the Creoles of Guayaquil revolted, proclaimed independence, and placed the port under Bolívar's protection. Advancing into the Ecuadorian highlands, Sucre defeated a Spanish army on the slopes of Mount Pichincha, near Quito. Royalist resistance crumbled on news of Sucre's victory; and the provinces composing the viceroyalty of New Granada—the future republics of Venezuela, Colombia, Ecuador, and Panama—were free. They were temporarily united into a large state named Great Colombia (Gran Colombia), established at the initiative of Bolívar by the union of New Granada and Venezuela in 1821.

Ever since the defeat of the British invasions of 1806–07, the Creole party, nominally loyal to Spain, had effectively controlled Buenos Aires. In May 1810, when word came that French troops had entered Seville and threatened Cádiz, an open town meeting convened to decide

Liberation of Argentina

the future government of the colony. This first Argentine congress voted to depose the viceroy and establish a junta to govern in the name of Ferdinand. In 1813 a national assembly gave the country the name United Provinces of the Río de la Plata and enacted such reforms as the abolition of *mita, encomienda,* titles of nobility, and the Inquisition. A declaration of independence, however, was delayed until 1816.

The junta promptly attempted to consolidate its control over the vast viceroyalty. The western interior provinces were subdued after sharp fighting. Montevideo, across the Río de la Plata, remained in Spanish hands until 1814, when it fell to an Argentine siege. The junta met even more tenacious resistance from the gauchos of the Uruguayan pampa, led by José Gervasio Artigas, who demanded Uruguayan autonomy in a loose federal connection with Buenos Aires; but Artigas became caught between pressure from Buenos Aires and Portuguese forces who claimed Uruguay for Brazil, and he had to flee to Paraguay (Uruguay did not achieve complete independence until 1828). The Creoles of Paraguay defeated a force sent from Buenos Aires to liberate Asunción and proceeded to depose Spanish officials and proclaim the independence of Paraguay.

The junta's efforts to liberate the mountainous province of Upper Peru failed because of steep terrain, long lines of communication, and the apathy of the Indian population.

Upper Peru in Spanish hands represented a standing threat to the security of the La Plata provinces. The military genius of José de San Martín of Argentina offered a solution to the problem. San Martín was a colonel in the Spanish army when revolution broke out in Buenos Aires. He promptly sailed for La Plata to join the patriot junta and was soon raised to command the army of Upper Peru. San Martín proposed a march across the Andes to liberate Chile, where a Spanish reaction had toppled the revolutionary regime established by Bernardo O'Higgins and other patriot leaders in 1810. Having liberated Chile, the united forces of La Plata and Chile would descend upon Peru from the sea.

San Martín's liberation of Chile and Peru

San Martín obtained an appointment as governor of the province of Cuyo, whose capital, Mendoza, lay at the eastern end of a strategic pass leading across the Andes to Chile. He spent two years recruiting, training, and equipping his Army of the Andes. The army began the crossing of the cordillera in January 1817, and in 21 days it issued on Chilean soil. A decisive victory at Chacabuco in February opened the gates of Santiago to San Martín. Another victory at Maipú (1818) ended the threat to Chile's independence. Rejecting Chilean invitations to become supreme ruler of the republic, a post assumed by O'Higgins, San Martín secured a number of ships in England and the United States and, in August 1820, sailed for Peru in a fleet of seven ships of war and 18 transports. He landed his army about 100 miles (160 kilometres) south of Lima but delayed moving on the Peruvian capital; he hoped to obtain its surrender by economic blockade, propaganda, and direct negotiation with Spanish officials. His strategy was successful; in June 1821 the Spanish army evacuated Lima and retreated toward the Andes. San Martín entered the capital and proclaimed the independence of Peru. But he then had to deal with counterrevolutionary plots and the resistance of Lima's elite to his program of social reform. Meanwhile, a large Spanish army manoeuvred in front of Lima, challenging San Martín to a battle which he dared not join with his much smaller force. San Martín became convinced that only monarchy could bring stability to Spanish America, and he sent a secret mission to Europe to find a prince for the throne of Peru.

San Martín met with Bolívar in Guayaquil (July 26–27, 1822); the proceedings are surrounded with an atmosphere of mystery. Argentine historians hold that San Martín came to Guayaquil in search of military aid but was rebuffed by Bolívar, who was unwilling to share with a rival the glory of bringing the struggle for independence to an end; San Martín then magnanimously decided to leave Peru and allow Bolívar to complete the work he had begun. Venezuelan historians argue that San Martín came to Guayaquil primarily to recover that city for Peru; they deny that he asked Bolívar for more troops and insist that he left Peru for reasons having nothing to do with the conference. Both interpretations tend to diminish the stature and sense of realism of the two liberators—San Martín must have understood that Bolívar alone combined the military, political, and psychological assets needed to solve the factional problems in Peru and to gain final victory over the powerful Spanish army in the sierra; given the situation in Lima, San Martín's presence there could only hinder the performance of those tasks. Viewed in this light, the decision of Bolívar to assume sole direction of the war and of San Martín to withdraw reflected a realistic appraisal of the Peruvian situation and the solution it required.

San Martín returned to Lima to find that in his absence his enemies had usurped his power. In September 1822, before the first Peruvian congress, he announced his resignation as protector and his impending departure. San Martín's departure left Lima and the territory under its control in serious danger of reconquest by the strong Spanish army in the sierra. Bolívar allowed the situation to deteriorate until May 1823, when the Peruvian Congress called on him for help. The scare produced by a brief reoccupation of the capital by the Spanish army prepared the Creole leaders to accept Bolívar's absolute rule.

Bolívar arrived in Peru in September 1823 and needed almost a year to achieve political stability and to weld his army and the different national units under his command into a united force. After a difficult ascent of the sierra, patriot forces won a victory near the lake of Junín (August 6, 1824). To Sucre fell the glory of defeating the Spanish army in the last major engagement of the war, at Ayacucho (December 9, 1824). Only scattered resistance at some points in the highlands and on the coast remained to mop up. The work of continental liberation was achieved.

Brazil. Brazil made a swift, almost bloodless transition to independence. The idea of Brazilian independence first arose in the late 18th century as a Creole reaction to the Portuguese policy of tightening political and economic control over the colony. The first significant conspiracy against Portuguese rule was organized in Minas Gerais, but it was easily crushed. The French invasion of Portugal (1808), followed by the flight of the Portuguese court to Rio de Janeiro, brought large changes and benefits to Brazil. The Portuguese prince regent John opened Brazil's ports to the trade of friendly nations, permitted the rise of local industries, and founded a "Bank of Brazil." Brazilian Creoles took satisfaction in Brazil's new role and the growth of educational and economic opportunities; but this feeling was mixed with resentment at the thousands of Portuguese who came with the court and competed with Brazilians for jobs and favours.

The Portuguese court in Brazil

The revolution of 1820 in Portugal precipitated Brazil's break with the mother country. The Portuguese revolutionaries framed a liberal constitution for the kingdom, but assumed a conservative posture toward Brazil; they demanded the immediate return of Prince John to Lisbon, an end to the system of dual monarchy that he had devised, and the restoration of the Portuguese commercial monopoly. John approved the new constitution and sailed for Portugal, leaving behind his son and heir Dom Pedro as regent and advising him that, in the event the Brazilians demanded independence, he should assume leadership of the movement and set the Brazilian crown on his head.

Soon it became clear that the Portuguese Cortes intended to abrogate all the liberties and concessions won by Brazil since 1808. One of its decrees insisted on the immediate return of Dom Pedro from Brazil in order that he might complete his political education. But Dom Pedro, urged on by Creole advisers who saw a golden opportunity for an orderly transition to independence, rejected the Portuguese demand, issued his famous Fico ("I remain"), and in December 1822, having overcome slight resistance from Portuguese elements, was formally proclaimed constitutional emperor of Brazil.

THE INDEPENDENCE OF MEXICO

In Mexico the movement for independence took an unexpected turn; the Indian and mixed-blood masses joined the struggle and for a time converted it from a quarrel between the Creole and Spanish-born peninsular elites into a social revolution.

A social revolution

In 1810 a Creole plot for independence was being hatched in the important industrial and mining centre of Querétaro; its leaders included Miguel Hidalgo, a priest in the town of Dolores known for his sympathy with the Indians. Informed that their plot had been denounced to Spanish officials, the conspirators decided to launch their revolt before preparations were complete. On September 16, 1810, Hidalgo called on the Indians of his parish to rise against their Spanish rulers; he appealed to the religious fanaticism of the natives by proclaiming the Virgin of Guadalupe the patron of his movement. After his first victories Hidalgo issued decrees abolishing slavery and Indian tribute and ordering the restoration of lands to the Indian communities. These measures gave the Mexican revolution a popular character largely absent from the movement for independence in South America, but they alienated the many Creoles who desired independence without social revolution.

Hidalgo could not weld his Indian horde into a disciplined army or capitalize on his early victories. Within a year he was captured, condemned as a heretic by an Inquisitorial court, and shot. A mestizo priest, José María Morelos, assumed leadership of the revolutionary struggle; he liberated most of southern Mexico, then summoned a congress that proclaimed independence and framed a republican constitution. Morelos extended Hidalgo's social reforms by prohibiting all forced labour and forbidding the use of racial terms. Like Hidalgo, Morelos also had differences with Creole associates that hampered his conduct of the war, and his premature efforts to install a constitutional regime also hindered his military efforts. In 1815 he, too, was captured by the Spaniards and shot.

The revolution then declined into a guerrilla war waged by many rival chiefs. Royalist armies gradually extinguished the remaining centres of resistance. The Spanish revolution of 1820 abruptly changed this state of affairs. Fearing the loss of their privileges, conservative clergy, army officers, and merchants in Mexico schemed to separate from the mother country and establish independence under conservative auspices. Their instrument was the Creole officer Agustín de Iturbide, who had waged implacable war against the patriots. Iturbide offered peace and reconciliation to the principal rebel leader, Vicente Guerrero; his plan combined independence, monarchy, the religious supremacy of the Roman Catholic Church, and the civil equality of Creoles and Spaniards. For the moment Iturbide's program offered advantages to both sides.

Conservative–rebel alliance

The united forces of Iturbide and Guerrero swiftly overcame scattered loyalist resistance. On September 28, 1821, Iturbide proclaimed Mexican independence; and eight months later a congress selected by Iturbide confirmed him as Agustín I, emperor of Mexico. Iturbide's empire included the captaincy general of Guatemala (which included present-day Guatemala, El Salvador, Honduras, Nicaragua, and Costa Rica). Iturbide's empire had no popular base, however, and within a few months he was forced to abdicate and depart for Europe, with a warning never to return. The captaincy general of Guatemala, following Iturbide's departure, declared its independence in 1823, forming the United Provinces of Central America. The constitution adopted by the United Provinces permitted a great amount of individual state autonomy and gave the federal government supreme power only in foreign affairs. This weak central government, although directed by the notable statesman Francisco Morazán, had almost no income and lacked the military force needed to subdue the frequent disorders that arose among the states. Poor communications, jealousy of the power of Guatemala, and inexperience in self-government led to secession and the end of the United Provinces in 1837–38. In 1824 Iturbide landed on the Mexican coast with a small party, was promptly captured by troops of the new republican regime, and was shot.

(B.K.)

Reaction and anarchy, 1825–50

The new republics

For a quarter of a century after 1825 the new republics were plagued by social and economic paralysis and political chaos. Independence had not released the shackles of inherited backwardness. Traditional society had been deprived of its legal framework, but it had not been destroyed. Its legacy of prejudices and customs remained, as did a colonial mentality and an intellectual dependence upon Europe. Latent social disorganization was everywhere in evidence. Different civilizations and stages of culture existed side by side.

The working class, "a melancholy sea of illiterates," had no sense of their stake in good government. African slavery remained in either legal or extralegal form. The large Indian population, accustomed to servitude, showed a strong desire to isolate itself from Europeanized sectors, thereby surrendering any political privileges it might have claimed as citizens. Mestizos and mulattoes, more aggressive than the Indians but distrusted by the Creoles, were essentially a frustrated and unruly lot. Their prototypes were the gauchos (cowboys) of Argentina and *llaneros* (plainsmen-cowboys) of Venezuela who often welcomed the opportunity to join private armies in order to strike out against their supposed oppressors. The privileged groups were the only social elements with permanent standards and values. They were blind to the evils that their systems might bring societies that had not elaborated acceptable substitutes for the displaced Spanish and Portuguese kings and the weakened Roman Catholic Church. The members of the privileged elements (the elites of land and church and the intellectuals) were individualistic to the point of being anarchistic. The laws they wrote were to be applied to others, not to themselves. It is with the failure of the privileged groups to couple liberty with law in mind that the tumultuous quarter-century after independence should be viewed.

POLITICAL GROUPS AND ISSUES

Liberals versus Conservatives

Within the larger framework of extreme individualism two groups vied for power. Between 1825 and 1850 each faction assumed that the primary problem of the new republics was a political one. One group, generally referred to as liberals, was led by the intellectuals of the cities; the other group, commonly designated conservatives, was directed by the aristocrats of the countryside. Those who looked to the intellectuals were in general oriented toward northern Europe and the United States and were anticlerical. They favoured a federal type of government with narrowly delimited executive power, and they were so strongly committed to liberty and the right of individuals that they professed to prefer disintegration to tyranny. Those who accepted the leadership of the landed elites were essentially Spanish in their outlook and encouraged the church to continue as the interpreter of the social value system. They desired efficient, centralized states directed by strong executives to such an extent that they seemed to favour tyranny to disintegration.

Before 1830 the landed elites had established their ability to neutralize the intellectuals. Their victory over the intellectuals was one of the rural areas over the urban centres, or, as the Argentine president Domingo Faustino Sarmiento said, "of thirteenth century feudalism over nineteenth century liberalism." The success of the elitists did not mean either the total defeat of the intellectuals or the resolution of the differences that had separated the contenders for power. The intellectuals retained sufficient influence to keep alive islands of radicalism that prevented the aristocratic elements from achieving completely the social rigidity their system demanded. Arrangements were made and compromises were reached on basic issues. The Conservative political leadership was able to reaffirm the area's attachment to its Spanish-Catholic heritage, but the intellectuals continued to encourage the flow of ideas from abroad, and by 1850 France had become a cultural prop for all of Latin America. By mid-century, too, even the Liberals had generally conceded that, in practice if not in theory, a strong executive was absolutely essential to orderly government; when representatives of the Liberals did attain power it was not unusual for them to disregard democratic forms. It often appeared that Liberalism was invoked in order to impose tyranny. The new states, defeated in their efforts to win for themselves the religious patronage that the Spanish and Portuguese crowns had enjoyed, had accepted the privilege of presentation of lists from which the Vatican selected the high ecclesiastical officials to serve in the republics. The anticlericals intensified their efforts to circumscribe the church's temporal activities—the registration of births and deaths, participation in marriages, control of cemeteries, leadership in education, participation in politics and economic activities—which gave that institution such strength as to make it a threat to the viability of any administration.

The importance of the issues that were delineated and then compromised by the contending power groups should not be discounted, but they alone do not explain the state of Latin America after three decades of independence. There were other considerations of equal or greater significance.

Violence. In the political area, probably the most far-reaching development before 1850 was the elevation of violence to the level of a political fundamental. In order to seize the initiative from the intellectuals, the landholding elites had recourse to the military forces that had been nurtured to maturity during the wars of independence. Their victory, thus, was one of violence, and success through violence engendered more violence. Armies soon came to serve a dual political function. They acted as the final arbiters of political matters, and they provided the guarantee that the privileged groups could fight among themselves over their political, social, and economic preferences without creating a power vacuum into which the masses might rush.

Private armies

Once force was injected into politics the juridical concept of representative democracy was ignored. Facts prevailed against constitutions—sometimes against their letter, usually against their spirit. Differences were settled by the breaking of heads rather than the counting of them. Bullets replaced ballots. Those who rose to power tended to monopolize it to such a degree that quite often their holds could be ended only by the bloody remedy of revolution. There was no place in such political systems for the popular masses, who constituted as much as 90 percent of the population in certain of the republics. They were, in effect, driven into a political wilderness.

Caudillism. Violence paved the way for the caudillos. They were violent men of destiny who broke all the bonds of national and social order. They were individualists who could not be restrained by party or ideology. To them electoral platforms were written to satisfy a formality and not a necessity. In the words of the Argentine sociologist Carlos Bunge, they were "rulers by the will of men without will." Some caudillos were generals from the wars of independence, who, in the course of the 15-year-long struggle for freedom, first lost contact with the people and then developed a proprietary interest in the states that their swords had helped to create. They were the liberators who turned upon the liberated. Some of the caudillos were mere adventurers who profited from the widespread disorder of the day. Many were *hacendados* ("large landowners") who took to the field to redress grievances against the central government or in response to the need of property owners for protection that the states did not provide. Once in possession of an army it was but a short step to using it for the purpose of achieving high public office.

The caudillos bore a striking resemblance to Max Weber's "charismatic leader." They regarded themselves as indispensable, and in office they exercised personal authority regardless of the representation of collective interests. Since the objectives of the caudillos were primarily personal they were forced to rely upon their personal magnetism to win followings. They had to be *muy hombre* ("very much a man"). Personalism (*personalismo*) in politics, consequently, became a fetish. Political parties were little more than ad hoc associations of friends, strong in victory, only to disintegrate in defeat. Prior to 1850 Chile and Colombia offered the only exceptions to the rule. If personalism was one of the strengths of the caudillos it was also one of their major weaknesses. Because their regimes were built on personalism, they were seldom able to consolidate their power sufficiently to pass it on to chosen successors. The opposition understood this phenomenon and was secure in the knowledge that sooner or later power might be expected to revert to it.

Personalism in politics

An often overlooked characteristic of the caudillos was their ideological and philosophical sterility. Although acts of brutality and destruction of property usually went along with the seizure and consolidation of power by caudillos, no caudillo before 1850 ever threatened a basic principle of the elite system of values. Except in Haiti and Brazil, they invariably paid lip service to the republican system of government. All spoke of government by the people, and no one of them felt secure until he legalized his administration by holding an election. They were no more successful than were regularly elected officials in bringing order out of political chaos. They added nothing new to the church–state issue. They were economic traditionalists at least to as great an extent as were those who opposed them. By the time they achieved power they were generally rich in land, and the constitutions they wrote reflected their property-owner mentality. Thus it was that in every major political, social, or economic area the caudillos tempered their actions with enough discretion always to make themselves appear safe. Each one stopped short of revolutionizing the existing social and economic systems.

Not one of the republics of Latin America was spared the price of political tumult born of caudillism. Argentina fell into the hands of "bloody" Juan Manuel de Rosas, who was raised to power by a combination of reactionary churchmen, wealthy landowners, and the gauchos of the Argentine pampas. Brazil experienced more than a decade of civil war after 1835 led by caudillos who threatened to divide Brazil as the Spanish area had been splintered earlier. Mexico barely survived the machinations of the superficially brilliant Antonio López de Santa Anna, who tempered tyranny with treason. The Central American Federation collapsed in a fit of quarrels kept alive by petty tyrants. Colombia, Bolivia, Peru, Ecuador, and Uruguay were surrendered to war lords and adventurers. Paraguay and Venezuela enjoyed rather prolonged periods of paternalistic dictatorship punctuated by blood baths perpetrated upon the citizenry by power-hungry politicians in military uniforms. Haiti, neglected by Europe and shunned by a race-conscious United States, stumbled from one despot to another, and the Dominican Republic was shaken by the onslaught of "the Haitians," the aggressiveness of its own strong men, and the ambitions of Spain. Only Chile, in the 1820s, after a period of extreme anarchy, evolved a political system that fettered the man on horseback. There the landed oligarchy, comprised of not more than a few hundred families, seized power and through a combination of aristocratic supervision and autocratic administration guaranteed its control. The military forces were brought under civilian leadership, and the president was given extensive control over a highly centralized government. The church was given full partnership with the state. Property and literacy qualifications limited the franchise to the point where not more than 10 percent of the population could qualify for the vote. The requisites for office holding eliminated all but a few thousand.

THE ECONOMICS OF ANARCHY

Although violence and caudillism were essentially political phenomena, no major area of activity was spared the pervasive anarchy they generated. Particularly did the disorderliness of the 1825–50 period add to the economic problems inherited from the mother countries and the wars of independence. The crucial economic problem of the new republics was the system of land tenure, the peculiar feature of which was the large estate or the latifundio. Rooted in the colonial period, the latifundio was consolidated during the years after independence as the republics recklessly distributed their public domains to political favourites, and then protected the new holdings with tax systems that taxed the product of the land but not the land itself, and legal systems that provided for primogeniture and debt, slavery, or peonage. Once in debt

The landowners and peasants

the Indian was obligated to the *patrón* until the money was repaid. Since he ordinarily found it impossible to regain solvency, if he had ever enjoyed such a blissful state, his labour obligations tended to become permanent. There was no place in such a tenure system for the growth of a property-owning middle class such as appeared in the United States before 1850.

To the heavy burden of the latifundios and peonage were added other economic problems of somewhat lesser long-range consequence but immediately important to the struggling nations. First Spain's colonial policy, and after independence the fact that the new republics were outside the well-established international trade routes, served to isolate Latin America from Europe and thus to prolong the time needed for European traders and financiers to acquire a somewhat realistic working knowledge of the republics. Violence also delayed the rehabilitation of many of the mines that suffered from destruction and neglect during the war years—mines that might have provided the means of earning an important part of the foreign exchange that Bolivia, Peru, Colombia, and Mexico needed so desperately. Finally, the recurrent political plots and uprisings strengthened the well-developed propensity that domestic capitalists had for investing in rural properties because of the indestructibility of land and the prestige associated with it in agricultural communities.

LEARNING AND THE ARTS

The literature of politics

The ravages of civil war left little time, money, or inclination for developing learning or the arts. Antagonism of the new governments to the clergy resulted in the neglect and disappearance of many colonial educational institutions and the loss of educational and professional services of the religious teaching orders. Meanwhile, the new republics, concerned with other problems and lacking in financial resources, did little to promote public education. Only Chile, which in the 1840s combined stability and wealth, was able to give impetus to learning at all levels. The arts suffered from the general decline of wealth. Almost no new churches and very few public buildings were constructed. There was little demand for art other than religious art. Literature fared somewhat better because it was made to serve political ends. Men of letters were generally on the side of justice, by which was meant justice for the privileged groups. Often from homes in exile—Chile and Uruguay were the favourite refuges from Argentina—writers carried on their campaigns against tyranny. The schoolteacher president of Argentina, Domingo Faustino Sarmiento, spent many of his most productive years fighting the tyrant Rosas from across the Andes in Chile. Esteban Echeverría described the crimes of Rosas and articulated the needs of his homeland from Uruguay. Through the efforts of Echeverría, the Chilean, José Victorino Lastarria, and others, Romanticism was accepted in poetry and then spread to drama and the novel.

Era of material progress, 1850–1914

For a quarter of a century after independence was achieved the conviction persisted that the problems of the inchoate republics could be resolved if only workable political systems could be found. The small politically articulate element of society threw vast amounts of efforts into what proved to be a fruitless search for exclusively political solutions to the dilemmas of the area. An enormous amount of property was destroyed as opposing elements uncompromisingly supported either arbitrary or rational politics as the most direct means of attaining their objectives through constitutionalism. The price was so great, the wounds so deep, that many of the republics still have not fully recovered. But as early as the 1840s in Chile and by the 1850s and 1860s in Argentina, Brazil, and Mexico, many within the responsible elements of society had agreed that the purely political approach had failed. They looked to Great Britain and the U.S. for alternatives, and the experiences of those countries taught that the successful functioning of democratic representative government, the ultimate goal of the Latin-American republics, was closely related to economic progress and that order was requisite to both. Economic development became an obsession in several of the republics. It dominated the era from 1850 to World War I as completely as politics had dominated the antecedent era.

Foundations of mining and agriculture

Latin America turned to economic solutions to its problems just as the world economy under European leadership was fostering a high degree of regional and international specialization. The basic effect of this development was to divide the world into industrialized nations on the one hand and nonindustrialized countries on the other. Latin America fell into the latter category. Mining, traditionally oriented toward export, was now greatly expanded. Agriculture was commercialized and given the same orientation. Enormous acreages were put to the plow for the first time in Argentina, Brazil, and Uruguay. Barbed-wire fences, blooded bulls, and refrigeration combined to transform cattle raising into a highly complex, modern enterprise. At the end of the century coffee became the economic backbone of several republics and bananas of several others.

Prolonged cycles of good markets enriched landowners to the point where many of them could afford to live in the capital cities or in Europe. Absentee landlordism, with all of its abuses, became prevalent in such countries as Argentina, Brazil, Chile, and Uruguay. As the tempo of international trade increased nations became the slaves of their mines and fields as minerals, livestock, cereals, and coffee provided the means of payment for imports of machinery and tools and the soft goods to which an expanding consumer class had become accustomed. Prior to World War I the dependence of the republics upon export markets did not often give rise to serious problems because the consumption of primary products by the manufacturing nations showed a strong tendency to expand and Latin America was ordinarily able to dispose of all it produced for export.

FOREIGN CAPITAL AND TECHNICIANS

The orientation of their economies to overseas markets made the republics more economically dependent upon Europe and the United States. The shift to large-scale commercial agriculture was financed through foreign-controlled banking institutions. The modernization of mines, beyond the economic capacity of domestic capitalists to carry out, fell to foreign investors, although in Chile domestic capital was heavily invested in the mineral economy. Commercial agriculture and modern mining involved the expansion of transportation and public utilities whose development was assumed by foreign capitalists and foreign technicians.

Foreign capital entered the republics without serious restraints of any kind, and the foreign investor showed a distinct preference for those states where the leadership displayed a reasonable capacity to maintain order. Thus it was that foreign venture capital went primarily into Brazil, where a respected monarch (Pedro II) moderated the differences of disparate groups; into Argentina, where a landowning aristocracy firmly controlled the government; into Chile, where power was shared by the landowners and the entrepreneurs of commerce and mining; and into Mexico, where Porfirio Díaz' policy of "bread or the stick" had brought order and stability out of extreme chaos.

Investment by Great Britain and the United States

Great Britain was easily the largest supplier of funds in South America as its capital went into commerce, utilities, railroads, shipping, mines, and agriculture. By 1914 Argentina was an economic colony of Great Britain. "Take Canada from us, but not Argentina," cried one Englishman. The United States did not become a creditor nation until World War I, but before that time United States private capital began flowing into the Caribbean and Central American agriculture and Mexican mines, railroads, and petroleum fields. By 1914 the United States was the largest foreign investor in Middle America. Foreigners had approximately $8,500,000,000 invested in Latin America when World War I broke out; one-third in Argentina, one-fourth in Brazil, and one-fourth in Mexico. In many enterprises, including railroads, mining, and bananas, the investment of foreigners greatly exceeded that of the nationals.

Technicians and managers, ordinarily from Great Britain and the United States, poured into the republics to fill a scientific and technological void left by the humanistic training that the institutions of higher learning continued to impart. In general the teaching of the sciences fell to the military academies. Except in Brazil, where army engineers were active, foreign engineers were primarily responsible for the construction of railroads and ports, and they supervised the work of thousands of native labourers.

IMMIGRATION

The economic expansion induced by foreign capital and technicians provided job opportunities in field and factory for millions of unskilled and semiskilled labourers from Europe. Steamships disgorged most of their human cargo at Rio de Janeiro and Santos (the port for São Paulo) in Brazil, Buenos Aires in Argentina, and Montevideo in Uruguay.

Underpopulation had been a recognized problem in Latin America ever since the independence era. In 1823 Brazil, comprising approximately half the South American continent, had only 4,700,000 inhabitants. Uruguay had a population of only 60,000 when it began its struggle for freedom from Spain. As late as 1852 Argentina's population was about 1,200,000. In a bid for immigrants, several states rewrote their constitutions to provide for freedom of worship for non-Catholics. The Argentine theorist Juan Bautista Alberdi at mid-century declared that "to govern is to populate." The Argentine constitution of 1853, which he fathered, enjoined upon the government the responsibility of promoting European immigration. What Alberdi had in mind was the filling in of the great expanses of his country with landowners, not farm hands. At the time he wrote there was no real shortage of workers in the republic because it had not yet shifted from subsistence to commercial agriculture.

A new dimension was given to the problem of underpopulation as a result of the course that economic development took after 1850. With ready markets in Europe for their products, landowners not only were discouraged from parting with their holdings but bid land prices up in enlarging them. Those who looked to trade with Europe wanted farm hands, railroads, better port facilities, and more public utilities. Also, the growth of retail trading and the beginning of the factory system required many more workers than could be provided locally.

Although several of the republics desperately sought immigrants after 1850, they did not at first find it easy to attract the restless from Europe. Most of those people preferred the United States, with its abundance of cheap land and its rapidly growing industrial system. Also, Latin America's reputation for political and economic instability was disturbing to potential settlers from abroad. Latifundism tended to exclude the newcomers from the ownership of land. Much of the area where lands were available was in the tropics or distant from commercial centres and ports. There was little industrial growth anywhere. Transportation and communications were poor; educational and medical facilities, primitive. Brazil tolerated black slavery until 1888, and Europeans were reluctant to work alongside slave labour. As late as the 1870s, in both Argentina and Chile, marauding Indians periodically encroached upon sparsely settled areas.

There were countervailing forces. Political discontent, famine, and grinding poverty in Europe drove many to take refuge wherever they could find it. In the republics there was some house cleaning as groups determined to promote material progress gained ascendancy and demanded order. So Europeans emigrated, at first in small numbers as traders and miners from the British Isles, and as farmers, mostly from Switzerland and the German principalities, and later in swarms from Italy, Spain, and Portugal.

Latin-American statistics on immigration are notoriously unreliable. According to those that appear most trustworthy, the swarming to Argentina began in the 1880s. Between 1857 and 1900 Argentina had a net inflow of 1,200,000 (equal to the total population of the nation in 1852). The peak year in Argentina was 1913 when 302,000 entered and 157,000 departed. By 1914 Argentina had received 4,660,500 immigrants, for a net gain of 2,640,000. Italians numbered about 1,200,000 and Spaniards about 1,000,000.

Argentine immigration

The impact of the immigrants upon Argentina was profound. The nation's population expanded to 10,215,787 by 1924. The population of the rich pampas increased 625 percent, and the republic's total agricultural acreage increased 60 times between 1865 and 1914. The immigrants helped to introduce farm machinery and mechanical arts and to hold the frontier against the Indians until the latter were finally subdued by Gen. Julio Roca in the late 1870s. They also added to Argentina's contacts with Europe. Their labour raised land values and made railroad building feasible. They popularized dairy products and vegetables in the Argentine diet and mongrelized the Argentine language. The Italian immigrants, who ran the country stores and victimized the gauchos, became the whipping boys of Argentine literature, while the shiftless gaucho assumed the role of a national hero. The flood of immigrants made Argentina more racially European than any other republic of Latin America with the possible exception of Uruguay. But, in the final analysis, the immigrants' greatest influence was probably felt in the city of Buenos Aires. Unable to buy land and unwilling permanently to accept the hard terms of rural employment and tenancy, they flowed into the capital city. There they entrenched themselves in the retail trade, small factories, the building trades, and in the maintenance department of public utility companies. In the burgeoning capital the rootless thousands, caught between the Old World and the New World, quickly developed social types that were new to Argentina and added to the European content of its way of life.

Those who emigrated to Uruguay came from the same places and at about the same time as did those who made their way to Argentina, and the contributions of the two groups were basically the same. By 1889 Montevideo with a total population of 214,000 contained 100,000 foreign born. By 1900 the newcomers possessed more than half of the declared national wealth.

Chile did not have a shortage of labour but actually exported labour throughout the 19th century. The only land the republic had to give away was heavily wooded and in areas where the rainfall was far greater than advisable for profitable agriculture. But the stability of the republic early made it attractive to German settlers. Moving into the forested regions around Valdivia and Puerto Montt as early as the 1840s, and later occupying part of the island of Chiloé, the Germans, although always constituting a relatively small portion of the total population, were able to put their imprint on much of the area south of the Bío-Bío River. By the middle of the 20th century the Germans of the area played an economic, educational, and military role in the region out of all proportion to their numbers.

Brazilian immigration

The most dependable figures available indicate that between 1851 and 1888 an average of about 10,000 newcomers a year entered Brazil and that the flow of immigrants reached its peak during the decade following the abolition of slavery. Approximately 2,500,000 émigrés from Europe entered Brazil between 1888 and 1914. More than 50 percent went to the state of São Paulo; the states of Rio de Janeiro and Rio Grande do Sul were other important recipients. As in the case of Argentina, Italy provided the largest single bloc of newcomers, but in Brazil the Portuguese rather than the Spaniards held second place. Although the Germans never entered Brazil in large numbers—probably not more than 100,000 by 1914—they prospered and had large families, and by 1924 numbered 360,000. Their influence in the state of Rio Grande do Sul was and is as pronounced as that of the Germans in southern Chile. In general, they needed to adapt less than did the Germans who came to the United States because they were concentrated in remote and thinly populated areas and were left to organize their own economic, social and educational life. Japanese emigration to Brazil began in 1908. Only a few thousand entered the republic before World War I but between 1920 and 1940 approximately 200,000 followed; they engaged in all types of agriculture and trade.

Overall, the role of the immigrants in Brazil was similar to that of their counterparts in Argentina and Uruguay. Their labour made possible large-scale commercial agriculture—in Brazil it was coffee. They became the principal source of labour in the cities of São Paulo and Rio de Janeiro. It may be said without qualification that those areas of Brazil where the immigrant went have prospered much more than have those regions where he refused to settle. The city of São Paulo, the Chicago of Latin America, is the handiwork of the newcomers and their descendants, and the greatest monument to the immigrant in Brazil.

Migration to the other republics, when of numerical significance at all, was for the most part non-European. By 1875 more than 74,000 Chinese had entered Peru to work on the sugar plantations, in the guano deposits, and on railroad construction projects. Lima still has a large Chinatown. In 1861 there were nearly 35,000 Chinese in Cuba. By 1899 the number of Chinese in Cuba had declined to less than 15,000. Blacks from the British West Indies entered Panama in large numbers during the construction of the Panama Canal. They were also brought into the coastal areas of Honduras, Nicaragua, Guatemala, and Costa Rica when the banana was commercialized, about 1900.

Haiti, with one of the densest populations in the Western world, traditionally has exported labour, principally to the Dominican Republic and Cuba. In periods of prosperity the Haitian workers have been welcomed; in depressions they have often been made to bear the brunt of attacks of Dominican and Cuban demagogues beating the drums of nationalism. Mexicans like Haitians, have emigrated. The ability of the U.S. to absorb Mexican farm labour has served as a safety valve for Mexico as it seeks to industrialize in order to provide jobs for its ever growing labour force.

TECHNOLOGICAL CHANGE AND ECONOMIC DEVELOPMENT

Growth of transport and communications

Aided by capital, technicians, and labour from abroad the economic and technological advances of the 1850–1914 period were spectacular given the area's inheritance from the colonial era and the anarchic quarter-century following independence. Railroads were laid from mines and farming centres to the ports. In all, 60,000 miles of railroads were built by 1914, including 21,800 (1914) in Argentina; 15,800 (1913) in Mexico; 15,445 (1913) in Brazil; 5,000 (1913) in Chile; and 1,656 (1913) in Peru. River steamboats were put on the Magdalena, Orinoco, Amazon, São Francisco, Jacuí, and La Plata-Paraná river systems. Old port facilities were renovated, as in Rio de Janeiro and Veracruz, and new ones were built. The port of Buenos Aires was dug out of the mud of the Río de la Plata. Rosario, Argentina, became one of the greatest grain-shipping ports in the world when a port was completed there after 1900. Steamship lines were subsidized in an effort to increase contacts with Europe and the United States. Before the end of the century hundreds of vessels flying British, French, Italian, German, and United States flags were carrying raw materials, finished products, and immigrants to all parts of the area but particularly to Argentina and Brazil. Private and public capital combined to put more than 100,000 miles of telegraph lines in operation by 1914. The area had submarine cable connections with Europe by 1876 and with the United States by 1894. The first telephone systems were introduced in the 1880s, and the first modern water and sewer systems were begun in the 1850s. Gas began to replace whale oil for public lighting in the 1850s and by the 1890s electricity in turn had began to supplant gas. During the era public transportation in the larger cities passed through three phases. Beginning with horse-drawn coaches it progressed to horse-drawn streetcars and to electric trolleys. Before 1910, Buenos Aires had in operation an excellent crosstown subway.

As a result of economic development during the era, Argentina became one of the world's major breadbaskets. Brazil strengthened its claim to the title of "coffee king." Chile became the leading exporter of natural nitrates. Mexico became one of the world's largest producers of petroleum by 1911. The total international trade of the area rose from slightly more than $1,000,000,000 annually in the early 1890s to $1,800,000,000 in 1907, and $3,100,000,000 in 1913. In the latter year Latin America accounted for 6.5 percent of the world's imports and 7.9 percent of its exports. The growth of trade and commerce in the more advanced countries gave a strong impetus to banking and insurance, a development which did not come about until after World War I in such countries as Colombia and Venezuela and the Central American republics.

Urban growth

Consequences of economic change. In the republics that had by 1914 established their leadership in the technological fields, the economic transformation brought nearer into balance the wealth, income, and population of the rural areas and the urban centres. The added importance given to land by the commercialization of agriculture was counterbalanced by the fact that the cities were the first to feel the effects of economic expansion because the industrial and commercial sectors, centred in the cities, received the greatest share of the capital and skills from abroad. The new and relatively well-paid management element associated with transportation and finance was essentially urban. The urban centres were the chief benefactors of the growth of both internal commerce and international trade. Villagers and those newly arrived from Europe, discontented with the terms of employment and tenancy in the rural areas, took advantage of cheap modern transportation to move into the cities, and Latin America at this time entered a sustained period of rapid urbanization. Buenos Aires trebled its population between 1895 and 1914, as did Montevideo between 1887 and 1914. São Paulo's population spurted from 65,000 to 350,000 between 1890 and 1910. During the same years Pôrto Alegre grew from 52,-000 to 130,000. Chile's population was 43 percent urban by 1910. On the other hand, the Indian countries, such as Paraguay, Bolivia, Ecuador, Nicaragua, and Honduras remained overwhelmingly rural, as did Brazil, where in 1914 less than 15 percent of the population was urban.

Two new socioeconomic elements emerged in the cities. They were the entrepreneurs, managers, and technicians of industry and commerce, who were added to the middle groups of society, and the urban industrial working class.

New alliance of intellectuals and the business community

The politics of progress. By the outbreak of war in Europe in 1914, the economic transformation and the growth of cities had had a significant impact upon the political life of the republics. In Uruguay, Argentina, Chile, and Mexico the new socioeconomic groups of the cities were making common cause with dissatisfied intellectuals and professionals against the political hegemony of the old elites. The intellectuals had good reason to welcome the new groups as political allies. The alliance of the intellectuals with the elites originally had been dictated by economic necessity rather than ideological compatibility, and consequently when their earnings eroded rapidly under the inflationary pressures that accompanied heavy investment in long-range enterprises, there was little left to keep the old working relationship alive. On the other hand, the intellectuals had much in common with the new business elements. Both groups were centred in the cities, were oriented overseas, supported rapid technological development, favoured public education as a means of limiting the influence of the church, and opposed the use of force in politics.

The urban labouring groups provided the popular support for the new middle sector politicians who opposed the entrenched leadership. Politically immature and organizationally weak, urban labour came under the influence of the middle sectors after a brief and stormy association with leaders who had been brought up in the traditions of anarcho-syndicalism then prominent in southern Europe. Political affiliations with the middle groups gave labour a certain respectability; in turn, worker militancy and intractability gave the new political forces an aggressiveness they would not otherwise have had.

The leaders of the new urban political alliance lacked practical experience in the formulation of policies but this did not deter them. They promised economic progress and social democracy. To the political and moral abstractions which the intellectuals had fought for during the independence wars, the new politicians, in their search for votes, added the demand that a greater share of the material and cultural benefits of 20th-century technology be made accessible to a greater part of the population.

Nationalism and a new type of militarism. Political nationalism and a more professionalized military class were important by-products of the political thinking that resulted from the influx of tens of thousands of hardworking immigrants, the construction of railroads, and investment in industry. Politicians began to envisage the day when their countries would teem with productive citizens and the empty spaces and corners of the republics would be needed to provide living space for growing populations. National boundaries, consequently, had to be guaranteed and, when possible, expanded. Although most international disputes were settled without resort to force, military establishments were strengthened in almost every republic in order to back up the claims of politicians.

Power of the army

The new emphasis upon the army as an instrument of national policy had at least two important consequences. A costly arms race developed, and the republics, feeling that the stakes were rising, turned increasingly to military professionalization. Also, the rise of modern armies equipped with costly armament marked the death knell of the provincial caudillo who had only his personal fortune with which to outfit his private army and keep it in the field.

CULTURAL CHANGES

Catholicism and anticlericalism. Between 1850 and 1914 the position of the Roman Catholic Church in Latin America became increasingly unstable. Severe anticlerical laws were put in force, for example, in Mexico, Venezuela, Chile, and El Salvador. In several countries church and state were separated and at the same time the church was burdened with strict state control. Although retaliatory vengeance for clerical political activities continued to be a cause for assault upon the church, a greater consideration as the century grew to a close was the emergence of socioeconomic groups in the cities whose members attacked the church as an obstacle to economic and social progress. Under the new leadership, anticlericalism to a degree became a political philosophy and a plan of action pointing the way to the modern world. The great mass of the people were unaffected by the attacks upon the church. Communities went along as before. They retained their patron saints, whose names were sometimes added to the names of the towns. Families had saints which they honoured. Other saints were expected to protect certain occupations or guard specialized groups.

Education and the arts. In the rush to modernize, learning was left behind. There were numerous exponents of educational reform, but few projects got beyond the planning stage because leaders in education seldom possessed the capacity for political leadership that would have enabled them to translate their theories into practice. But in any event to incorporate into society a large mass of indigenous peoples through education was beyond the comprehension of those who led the republics. The Catholic Church continued to dominate primary and secondary education, despite steadily growing support in favour of public instruction. Between 1850 and 1914 the Catholic University of Chile, in Santiago, founded in 1888, was probably the outstanding addition to the church's list of institutions of higher learning.

Only Argentina, Chile, and Uruguay, each of which had homogeneous, largely European populations, made major achievements in public education between 1850 and 1914. In response to the demands for a more responsible citizenry and the requirements of the new technical age, but above all because of the driving determination of Pres. Domingo Faustino Sarmiento to create a literate nation, Argentina moved ahead rapidly to become the educational leader of Latin America. Matriculation in Argentine normal schools increased from less than 1,300 to more than 7,200 between 1890 and 1912 and in the same period enrollment in public and private elementary schools rose from about 300,000 to 720,000. Enrollment in the University of Buenos Aires reached 4,600 in 1915. The University of La Plata, founded at the turn of the century, became one of the great institutions of the area before Juan Perón dispersed its faculty during the decade after 1946. Although Chile failed to maintain the very fine record it had set in the 1840s, it ranked in the prewar period as a leader in the field of public instruction. By 1908 enrollment in public elementary schools stood at 240,000 and 18,700 in public secondary schools by 1910. Uruguay placed the emphasis upon elementary instruction while establishing their principle of free and obligatory education under secular control at all levels.

Culture of the newly rich

Although important changes did take place, the economic-technological transformation stopped considerably short of generating a social revolution. Inequities became more pronounced as industry and commerce gave rise to more and greater fortunes than farms and mines had ever supported. Gaudy homes went up in exclusive residential districts. Conspicuous consumption was appreciably more evident in the cities than it had been on the haciendas. Novelists geared their writings to the new rich who wished to read of themselves in glittering palaces and travelling abroad. Portrait painting was popularized as the number of those who could afford such luxury multiplied. Academies of arts were founded and national museums were established. History societies flourished, as the new nationalism aroused interest in the past, and archives were searched for evidence to support boundary claims. A few cities, notably Rio de Janeiro and Buenos Aires, supported operas. The demand for newspapers grew as learning slowly filtered down to the urban working elements and immigrants bought foreign-language newspapers in order to keep abreast of developments in Europe. But the rigid system of social classes remained largely intact. "The darker one's skin the lower one's social position" remained the rule. Although gradations existed, the essential division in the social hierarchy was between the dominating privileged classes and the people. The interpreters of society either ignored the masses or treated them in a residual manner. Many Brazilians wrote against slavery, but they ordinarily attacked the institution on moral rather than social grounds.

THE END OF AN ERA

Well before World War I there were many indications that large parts of Latin America were breaking with their Spanish and Portuguese pasts and that several republics were taking their first steps toward becoming modern nations. Slavery as a legalized institution in the Western Hemisphere was finally ended in 1888 when Brazil decreed the emancipation of approximately 850,000 blacks then held in bondage within its borders. The following year Brazil forswore monarchism, which had become an anachronism in the New World, in favour of republicanism, thus bringing its political system into line with the other hemisphere nations. Political leaders especially in Argentina, Uruguay, and Chile increasingly involved the urban working groups in national affairs.

The republics established an enviable record for the peaceful solution for controversies, expecially in the settlement of boundary disputes. The more recalcitrant cases were submitted to arbitration, and before World War I arbitration had for practical purposes become a part of international law as it applied in Latin America.

The modern political limits of the area were drawn when Cuba won its freedom from Spain (1899) and became an independent republic in 1902, and the number of nations was raised to 20 when Panama separated from Colombia and achieved sovereign status in 1903.

All the republics were invited to the second Hague Peace Conference held in 1907. Although the position taken by the Latin-American delegates against armed intervention for the collection of debts was accepted only after significant modification, the conference was important because it was the first of worldwide scope in which the Latin-American republics participated.

Influence of the United States

As the 20th century began, the republics were deeply conscious of the fact that their future was clouded by an ominous shadow cast over them by the United States. The "Colossus of the North" controlled Cuba through the

Platt Amendment. It had created Panama. Pres. Theodore Roosevelt's big stick policy had not only sanctioned intervention in Latin America but had claimed for the United States a monopoly of the right to engage in it, thereby making the United States not only the policeman of the hemisphere but its judge as well. President William Howard Taft's dollar diplomacy was followed by the sending of Marines into the Caribbean and Central America. President Woodrow Wilson did little to quiet the fears of the Latin-American leaders when he intervened in the internal affairs of Mexico by more or less openly calling for the overthrow of Dictator-Pres. Victoriano Huerta. By 1914 the United States furnished 32 percent of the imports of the entire area and took 36 percent of the exports of the republics.

Distrust of the United States had much to do with the official attitudes of the Latin-American republics in World War I. Despite great sympathy on the part of the privileged groups for France, only eight of the countries declared war against Germany, and only two, Cuba and Brazil, offered military aid. Seven of the eight republics that declared war were dominated financially by either the United States or Great Britain. Only Brazil made its choice with comparative freedom. Five of the republics broke relations, and seven, including Argentina, Chile, Colombia, and Mexico, remained neutral. Mexico was actually pro-German for at least part of the struggle. Early in the war international trade was brought to a near halt, but by the end of the conflict the republics were supplying food and strategic materials in abundance. The foreign commerce of the 20 republics rose to $3,900,000,000 in 1918 after having fallen to $2,200,000,000 in 1914. The fact that under the stress of war the Allies failed to provide the manufactured goods to which the population had become accustomed gave a fillip to domestic manufacture of consumer goods.

Between the wars, 1918–39

By the end of World War I it was quite apparent that the 20 republics were in widely varying states of development and that the differences among them were actually increasing. Bolivia, Ecuador, Paraguay, Peru, Venezuela, and the Caribbean and Central American republics, except Costa Rica, lived in the mid-19th century. They were politically immature, plagued by meaningless revolutions, economically backward and burdened with foreign debts, and generally lacking in social awareness.

Colombia was in a category by itself. It was dominated by a narrow oligarchy that at times had shown unusual administrative skill, but the structure of the country retained most of its colonial features and the nation was just beginning to enter international markets. With tremendous potential compared to many of its neighbours, it was nearly prostrate from civil strife that bordered on anarchy.

By 1920 Argentina, Chile, Costa Rica, Uruguay, and Mexico were clearly in the vanguard, and Brazil joined them in the 1930s. It is with these six republics, containing two-thirds of the land area and two-thirds of the population, and annually producing approximately two-thirds of the gross product of the 20 republics that the remainder of this article is primarily concerned. Between 1918 and 1968 it was in these republics—joined by Venezuela after 1960—that one found most of the region's dynamism, its new self-confidence, its desire to create and its very conscious groping for self-expression. It was to the seven republics that, for one reason or another, the other republics looked, when they looked at all, for leadership.

THE NEW POLITY

Political recognition of the popular masses. During the century after independence the leaders of Latin America had at first sought essentially political and then economic solutions to the problems of the republics. Between World War I and World War II the leadership in the more progressive republics not only merged politics and economics but added a social ingredient to their ideology. Socioeconomic problems were made the fundamental political problems. Democracy, which earlier had been thought of primarily as a theory of government and which in practice was more nominal than real, was expanded to include economic and social democracy. The exaggerated individualism of the 19th century gave way to an emphasis upon social solidarity. The concept of the state as a passive organism, limited largely to maintaining order and collecting taxes, was supplanted by a concept of the vital state that would provide leadership in the economic and social fields as well as in the political. Politicians spoke more of social justice, less of legal justice; more of social equality, less of political equality; and masses came to prefer good government to self-government as they knew it.

Movements toward social reform

There was a twofold explanation for the new concern with the issue brought to the surface by the transformation of the pre-World War I era. The urban working groups, aroused by the knowledge that the stirrings of European workers immediately after the war had won official recognition of their problems, demanded action on their behalf. The middle sector politicians, in their struggle for power with the old elites, had been forced to go to the workers for electoral support. Once the workers were brought into the political arena their social anxieties, frustrations, and discontents could no longer go unheeded.

Under the influence of the new political philosophy the traditional concept of the role of private property in society was challenged. The Mexican constitution of 1917 defined private property in terms of social function. The Chilean constitution of 1925 established the superiority of the state over private property.

Workers' welfare

By 1930 there were numerous indications of the growing concern over working conditions and the welfare of workers. Article 123, often called the Magna Carta of Mexican labour, of the constitution of 1917 not only enjoined the state to foster a strong Mexican labour movement but recognized labour as a status, a way of life, for which the minimum essentials must be guaranteed. By the end of the 1920s Uruguay had adopted legislation providing for the eight-hour day, minimum wages, equal pay for equal work, protection on the job, and fixing the terms under which minors might be employed and workers might be separated from their jobs. Article 56 of the Uruguayan constitution of 1934 made the state responsible for promoting the organization of trade unions, enacting standards for the recognition of their juridical personality, and promoting the creation of tribunals of conciliation and arbitration. The Costa Rican constitution of 1925 contained elaborate guarantees to both urban and rural labour. The Chilean labour code of 1931 embodied radical principles and general practices regarding labour. In Brazil, Pres. Getulio Vargas put many social measures into law: a minimum wage, old-age pensions, accident insurance, workers' savings banks, vacations with pay, cheap housing. As early as 1925 the Chilean government wrote an imaginative program of social security, including an elaborate structure of social security funds to which workers, employers, and the state contributed. Uruguay carried on public works programs as insurance against unemployment. In Mexico, Pres. Lázaro Cárdenas spoke of "land for those who worked it."

Material gains and social welfare for the workers were often delayed and seldom measured up to expectations. A vast share of all social legislation remained as little more than statement of aspirations. Still, the self-respect and the sense of belonging to the nation that the new official attitude gave the labourers in Mexico, Uruguay, and Chile was immediate and profound. On the other hand, the failure of Argentina's "don't-touch-agriculture" leadership of the 1920s and the 1930s to give recognition to the workers prepared them for the acceptance of Perón's quasi-Fascist administration in the 1940s.

State interventionism. In the 1930s the national governments in several of the more advanced states began to reject the laissez-faire doctrines of the 19th century. They broadened their economic responsibilities in search of faster and better ways of fulfilling the social obligations they had assumed. The governments justified their action on three grounds: (1) Industry could not survive without protection from outside competition and only the state could provide that protection. (2) Since the accrual of domestic private capital was slow, the state, with its ability to accumulate capital relatively rapidly through taxation and foreign loans, must intercede in the industrial sphere. (3) Solicitude for the labouring classes required that the state exercise some control over the prices of necessaries.

The economies that the governments sought to vitalize were reeling from repeated blows. The international economy had been severely jolted by World War I and had not been reconstructed in the uncertain political and economic environment of the 1920s. Economic nationalism in Europe had placed at a disadvantage countries, such as those of Latin America, that depended upon the export of primary products. The fact that each republic continued to rely on one or two basic exports had weakened further their bargaining position in the new world economy. The economic dislocations and problems of the 1920s were heightened by the Depression of the 1930s, which saw the drying up of foreign loans, the complete collapse of the world economy, and widespread defaulting on debts. Chile's and Bolivia's exports dropped 80 percent between 1929 and 1932.

The Great Depression

There were many instances in the 1930s of states investing directly in economic enterprises, as, for example, in Uruguay and in Mexico following the nationalization of the railways and the expropriation of the foreign-owned oil properties. But before World War II the general practice was for governments to limit their intervention to placing restrictions on imports, using exchange controls, licensing arrangements, imposing embargoes in order to promote domestic manufactures, and creating autonomous and semiautonomous institutes to assist designated sectors of the economy. These measures mainly represented attempts by nations to protect their foreign exchange earnings. Economic nationalism was at best a secondary consideration.

The manufacturing sector of the economies made significant advances in the 1930s, probably as much because of the impetus given by the Depression, when the foreign exchange situation left no alternative to domestic production of a wide range of manufactures, as to state promotion and planning. While manufacturing made appreciable gains there was a marked decline in capital invested in transportation as revenues dropped and funds for upkeep and replacements were not forthcoming from other sources. Agriculture, still the principal source of foreign exchange for a majority of the republics, was neglected.

THE NEW SOCIETY

Social mobility. The interwar years produced social mobility and social ferment on an unprecedented scale. Increased opportunities for learning, coupled with the growing demand for intellectual and technical skills in the public and private bureaucracies and in education, permitted many persons from the lower groups to achieve middle-sector status. Employment opportunities in factories, commercial establishments, public utilities, and the building trades attracted hundreds of thousands from the rural villages into the burgeoning cities.

In Uruguay, Argentina, Chile, and Costa Rica, where the populations were overwhelmingly European, there was little change in the ethnic composition of the social groups or the cities. But in Mexico the middle classes, the skilled working classes, and the cities all experienced a heavy Indian infiltration, and in Brazil there was a noticeable Africanization of the major industrial and commercial centres. Labourers who had become members of the middle class and farm hands who had become industrial workers experienced the frustrations inherent in passing from one socioeconomic group to another. The new social milieu placed those who knew the lower levels of society from first-hand experience beside those who had only a paternalistic interest in or a theoretical understanding of the working element. Those who depended upon labour leaders to defend their interests often worked for individuals who had an abhorrence for the labour movement. Those who had never owned property and were little concerned with property rights or infringements upon private enterprise were thrown together with persons strongly committed to the defense of private property and personal initiative. The clashes that resulted from the coming together of two essentially different cultures were often bitter, and there was much pulling and hauling as groups fought for advantage. But there was also a sharpening of the focus on the goals and aspirations of the people as they struggled with the crisis of growth, and there was no turning back of the clock as the workers retained, at least in principle, the gains they won.

Rise of Indians, blacks, and other groups

The family in a changing society. Under the pressures of modern life the family, which traditionally had served a political as well as a social role, began to weaken. This was particularly true in the cities where private dwellings began to give way to cramped apartments. Economic pressures forced many middle-class women to find employment outside the home; increased education and job opportunities for girls gave them a new freedom. Modern means of transportation encouraged young people to travel and make associations outside the family. Cinemas, clubs, and public parks bid against the family for the leisure time of its members. Youths left their homes to go into education, business, and the professions, and returned home with information that permitted them to reach decisions independently of the older members of the household. The new independence helped to weaken allegiance to the family as a political unit, and there was a strong tendency to transfer this allegiance to political groups. Political parties provided a common ground for those who had similar objectives based on education and occupational interest and social relationships outside the home.

THE NEW MILITARISM

The role of the military establishment underwent basic modifications between the two wars. The young men who had entered the armies at the turn of the century had now come to power. They were from the same socioeconomic groups as the new civilian leadership; as the economic and political influence of the middle sectors rose, the officers' inclination to associate themselves with the old elites declined. The officers, meanwhile, had learned to accept the working elements, if not always to trust them. Like the middle-sector leadership, the officers of the armed forces supported nationalism as a means of achieving rapid industrialization.

Professionalization of the military forces did not have the effect of keeping them out of politics, but it did lead to a new concept of the military's role in government and society. Under the influence of the new concept the officers tended to discard their historical subservience to an all-powerful strong man. Instead, they organized in such a way that when intervention took place it was in the name of the armed forces rather than of an individual, and government was by juntas in which all branches of the armed forces were represented. While the earlier military leaders had been content to declare for the general will, the new military statesmen felt compelled to define the content of the will. Several of the Caesar-statesmen who won recognition during the 1930s showed a distinct preference for Nazism and Fascism.

The army in politics

NATIONALISM

As a result of both external and internal developments during the 1920s and early 1930s, cultural nationalism was given wide play in the Latin-American area. France, to which Latin America traditionally had looked for cultural inspiration, failed after the war to regain its cultural hegemony in the Western world. The stream of immigrants from Italy, Spain, and Portugal diminished under the impact of the Depression. This development served to loosen the area's ties with the Mediterranean region, and the separation became more complete as those who had migrated before 1920 adapted themselves to their new surroundings. Simultaneously, financial leadership in the world passed from Great Britain to the United States, whose people were considered cultural barbarians by the Latin-American elites. These conditions were made-to-order for the new middle-sector political leaders who were bidding for power against the old elites. They argued that since Europe was losing its cultural position and Latin-American ties with the area were weakening, the old elites, "who lived with their feet in America and their heads in Europe," should be repudiated. They also contended that

Growing cultural independence of Latin America the decadent neo-European culture the elites had imposed upon Latin America could not withstand the increased pressures of an economically dominant United States. Finally they declared that since in the new order each social sector was worthy of representation in government it must be assumed that each group had something worthwhile to contribute to the national personality. It followed from these arguments that Latin America was obliged to look inward and search out the best that each of its social groups had to offer, and then to integrate the whole into a culture that would be truly representative of the national essence.

In the 1920s intellectuals had taken the lead in promoting the nationalization of their culture, and they had directed their appeals to the learned components of society. With the onset of the Great Depression, the states, at the same time that they expanded their social and economic responsibilities, seized the leadership from the intellectuals and added an embryonic economic nationalism to the earlier cultural nationalism. This integrated nationalism was brought down to the masses in concrete and politically charged form. As such it was soon raised to the level of a major political ideology and there it has remained. The glorification of a narrow nationalism was used to unite all elements in the swelling resentment against the exactions of economic imperialism and to cultivate the national spirit while undermining the atomistic assumptions of liberalism. It served as a vehicle for the adoration of collective power. In its more virulent form the new nationalism became anti-imperialism and more specifically anti-United States. At first it was directed against United States political intervention and cultural infiltration of the republics. Later it was manifested in attacks upon U.S. investors, who, it was generally charged, were depleting the area of its irreplenishable natural resources and discriminating against national employees.

World War II

The Good Neighbor Policy had helped to improve understanding between Latin America and the United States. When the second global war erupted, all of the republics except Chile and Argentina were prompt to throw their support to the Allies. By the end of January 1942, nine of the nations had declared war on the members of the Axis and nine others had severed relations with them. The war declarations became unanimous when the Fascist-oriented leadership in Argentina finally declared war in March 1945 and thereby qualified for a seat in the United Nations.

As in World War I, Latin America's contribution to the Allied cause was largely in the forms of foodstuffs and strategic materials. Brazil, however, sent an expeditionary force to the Mediterranean, and a Mexican air force squadron fought in the Pacific theatre of operations.

The Allied demands for foodstuffs and minerals resulted in nearly full employment in Latin America during the war years and lifted the gross product of the area to unprecedented heights. Increased industrial production for domestic consumption was achieved by placing added burdens on existing plants and equipment because the wartime economies of the Allies did not provide for the export of capital goods, except in most unusual circumstances. There was a moratorium on strikes. Labour–employer issues and social welfare demands were reduced to a secondary position as politicians (in Mexico, Chile, and Brazil, for example), and Communists and Communist sympathizers everywhere called for an end to class conflicts in the face of the grave threats from abroad. The United States, whose overriding consideration was strategic, helped to maintain stability in the area by providing incumbent regimes with military and economic aid.

The postwar years

THE ECONOMY

Industrialization was a principal obsession of the peoples of Latin America after World War II and heavy industry became the symbol of national progress. Figuratively speaking, the fiery furnaces of real and anticipated iron and steel plants fused the goals of the working elements and the hopes of the privileged groups into a national will, the expression of which was reflected in all major areas of human activity.

Latin America had been formulating its new views on industrialization for some time before they found expression after the defeat of Germany. The extractive and processing industries that had been promoted so assiduously before World War I and during the 1920s had not lived up to expectations. They drained off natural resources without, it most cases, providing real impetus to other sectors of the industrial economy. Industrial production, even at the expanded rate of the war years, had not brought the hoped-for economic emancipation, since the machines, and many of the raw materials and the fuels needed for the production of semidurable goods, had to be imported. World War II had shown the inconvenience of dependence upon the outside world for replacement parts and machine tools. Finally, it had become evident that the great powers were, in fact, great manufacturers. This meant that if the republics were to gain worldwide recognition and respect they must, among other things, have steel plants and make machines.

The new industrialism

Much of Latin America's industrial expansion has been and continues to be financed through inflationary borrowing. An intermittent phenomenon in Latin America since the late 19th century, inflation surged upward in the late 1930s and after World War II spiralled in several countries, notably Argentina, Bolivia, Brazil, and Chile. Borrowing has also substantially increased the external debt of most Latin-American countries and caused greater portions of national incomes to be devoted to repayment. Rampant inflation has tended to direct capital toward short-run speculative ventures, rather than basic industry requiring relatively long maturation periods. Industrial development, far from contributing to internal equilibrium, has in fact more typically distorted the economies by contributing to regional concentration of income. This is because industry has invariably grown up in the main urban centres where a steady demand for consumer goods exists and where skilled labour is most readily available.

As it has turned out, a large part of Latin America's commitment to industry has been made at the sacrifice of the agricultural sector. After 1945, farm production barely kept pace with population growth. In order to feed their people adequately some countries shifted to subsistence agriculture from commercial agriculture, the principal source of foreign exchange for all the republics except Bolivia (tin), Chile (copper), Mexico (tourism), Peru (industrial ores), and Venezuela (petroleum). As of the mid-1960s the agricultural sector contributed less than 20 percent of the gross product of the republics. Agriculture's contribution decreased further the next decade.

Land reform

Little progress has been made in the redistribution of land. The governments of Mexico, Bolivia, and Cuba have seized and divided land with little regard for the original owners. In Venezuela, the government has had the means to pay for land that it has acquired for redistribution to peasants. Elsewhere, nations and international technical experts and legislators search for legal means to penetrate on a broad front the tight ranks of landholders, who in Latin America exercise political power out of proportion to their economic contribution. In the late 1960s Chile, where Christian Democratic Pres. Eduardo Frei Montalva fought for agrarian legislation, provided the most interesting test as to whether effective agrarian reform could be instituted within a constitutional context. During the 1970s, agrarian reform became an important part of leftist movements, such as the Sandinistas in Nicaragua, in Latin America. This caused some governments, El Salvador for example, to initiate limited land reform policies.

Land redistribution tends to capture the imagination of the public, but as a short-range goal increased production may be equally as important and as difficult of solution. In many areas agricultural techniques are incredibly outdated. Landholders in general seem unconcerned about modernizing their methods and equipment. Many do not have the means to modernize should they choose to do so, while others are reluctant because of the threat of expropriation which is ever present.

State intervention. The clamour for rapid industrialization invited increased state intervention in economic life. Domestic private investors could provide only a small part of the necessary capital. Foreign investors were reluctant to enter the "politically charged" power and transportation fields, particularly when the threat of expropriation hung over them. Nationalists successfully prevented the investment of foreign capital in a large number of activities closely related to natural resources. The national governments, on the other hand, could and did raise capital through taxation and by borrowing abroad.

Government ownership of industry

Examples of large-scale government participation in industrial enterprises are numerous. In Brazil, Volta Redonda, one of the largest iron and steel plants in Latin America, is government controlled. The Brazilian government is the owner of numerous hydroelectric facilities, and Petrobras is a publicly-controlled petroleum monopoly. In Argentina the government controls the railroads and the petroleum industry. In Chile the central government has a direct interest in the production and sale of natural nitrates and has formed a partnership with the copper companies to regulate its production and exports. It also controls the iron and steel plant near Concepción. A saturation law gives an agency of the Chilean government power to exclude further investment in areas where production is deemed sufficient. In Colombia the iron and steel plant at Paz del Río has the unqualified support of the government. In Mexico the government through Pemex operates a petroleum monopoly, and the railroads are nationalized. The Nacional Financiera, most of the stock of which is owned by the Mexican government, is the channel for negotiating and administering loans obtained from abroad and is the principal source of funds for industrial development in the country. In Uruguay the government controls the production of electricity and operates the telephones, railroads, the port of Montevideo, insurance banks, and the meatpacking industry; it also imports and distributes coal, distills alcohol, and manufactures all cement used for public works. In Bolivia the central administration produces and markets most of the country's tin.

Economic diversification. By the 1960s industry and commerce accounted for 50 percent or more of the gross products of Argentina, Brazil, Chile, Colombia, Mexico, Uruguay, and Venezuela. At that time about 15 percent of the total labour force of Latin America was in manufacturing as opposed to about 45 percent in agriculture and 30 percent in service occupations. By the late 20th century industry and commerce accounted for more than 80 percent of the gross domestic product of Latin-American countries.

Industrial development has contributed appreciably to diversification within those sectors of the economies producing for internal consumption. The more highly industrialized countries are today producing a major share of their textiles, appliances, pharmaceuticals, and building materials, and are moving ahead in the manufacture of industrial chemicals. The republics have enjoyed only limited success in modifying the composition of their exports. Even as late as the postwar years they continued to export the same primary materials that they exported before World War II and a considerable portion of their foreign exchange earnings continued to come from a single commodity.

Foreign trade. To a greater extent than in the case of the countries that have reached a higher level of industrial development, the economies of the Latin-American republics have remained geared to overseas markets. Throughout the late 20th century it was common in many of the countries for the value of exports to contribute substantially to the gross national production. This situation, generally considered unfavourable by the leadership of the republics, has tended to increase Latin America's share of world trade.

The quite limited trade between the republics has given rise to a marked effort to expand their regional ties. The first major effort, the Central American Common Market (Cacom), encompassing El Salvador, Guatemala, Honduras, Nicaragua, and Costa Rica, created in 1957, was remarkably successful until civil wars and national rivalries in the 1980s halted progress.

LAFTA

The Latin American Free Trade Association (LAFTA), presently involving all the countries (except Guyana) in South America plus Mexico, went into effect in June 1961. LAFTA had many strong supporters and prospects for a successful future, as it succeeded in bringing about an increase of trade between the republics and in encouraging the private sectors. Many problems confront it, but none more serious than the need to increase its share of world trade, since foreign exchange is vital to the continued development of the republics.

Foreign investment. The disastrous experiences of the Depression and the conduct of nationalistic governments in the late 1930s made foreign investors most reluctant, after World War II, to send capital to Latin America in the form of portfolio investments. On the other hand, there has been since the war a relatively heavy flow of private capital into the area as a result of direct investments by foreign enterprises. Between 1945 and 1950 private capital went primarily into petroleum, but since 1950 investors have shown increased interest in manufactures and the extraction of minerals other than petroleum.

Latin-American governments in general have welcomed foreign investors since World War II but have insisted that since foreign investors will not put their capital into the development of those sectors basic to long-range growth, such as power and transportation, aid in these areas must come from international lending agencies.

Various agencies of the U.S. government have also been active in assisting the republics to develop their human and natural resources. The Export-Import Bank of Washington is the oldest of such agencies. After many years of agitation by the Latin-American republics, the U.S. government in 1959 approved an Inter-American Development Bank. This agency was originally capitalized at $1,000,000,000, of which $450,000,000 was contributed by the U.S. The bank approved its first loan in 1961. Loans and grants were made to the republics by the U.S. Agency for International Development (AID).

Urbanization. After World War II, three factors combined to produce a major urbanization movement: industrial and commercial expansion, a population explosion that was annually adding almost 3 percent to the total number of inhabitants, and the advantages afforded by the cities. In 1925 Latin America's population was 33 percent urban; in 1960 it was 45 percent urban and the percentage increased rapidly thereafter. Uruguay became the most urbanized of the republics, but nearly all the republics showed a pronounced trend toward urban concentration. During the decade after World War II Colombia's urban population increased 58 percent, Venezuela's 57 percent, Panama's 54 percent, and Mexico's 50 percent. Almost everywhere the tendency was for the largest cities to grow faster than the smaller ones.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Centralization

One of the profound effects of state intervention in the social and economic fields has been the general strengthening of the already strong central government at the expense of the states and municipalities. In several of the smaller nations the state as a separate entity does not exist at all and the municipality thus serves as the only buffer between the central government and the individual. In many cases, notably Colombia, the states and municipalities have been largely stripped of their power to tax and are consequently dependent upon subsidies from the central authority.

The executive. Historically, the tendency in Latin America has been toward the concentration of power in the executive branch of government, a tendency that was not ordinarily deterred by any theoretical balance of power or by federalism. The strengthening of the central governments at the expense of the states and municipalities in most instances has encouraged the trend toward a strong president. Since World War II the legislative branches have not in general exercised their constitutional prerogatives, and the legislatures have been little more than rubber stamps. The chief executive, also, generally participates in, if he does not actually make, judicial appointments, and thus is able to exercise considerable influence over the judiciary. Moreover, programs of social legislation, which have tended to place social interests above the rights of individuals as private citizens, have been more closely linked with administrative acts than with private civil law. This innovation gives the executive branch power to regulate the rights of individuals by administrative decrees and thus to reduce the power of the regular courts.

The electorate. In a drive for general enfranchisement, property requirements have been removed in most instances; age and literacy requirements have been reduced and in some cases discontinued; and women have been granted the right to vote. These measures have been highly successful in broadening popular participation in politics.

Political parties. Personalism and the political influence that the family traditionally exercised have given the erroneous impression that organized political parties are either nonexistent or are of recent origin in the Latin-American area. It is true that in several of the republics major and even dominant parties date from World War II. But it is also true that by the mid-19th century two relatively homogeneous party traditions had emerged, represented by the Liberals and the Conservatives. The Liberal and Conservative parties of Colombia and Chile all date from before 1850. In the late 19th century nationalist parties arose with programs for national development. Parties with middle-class leaders and depending upon the electoral support of labour groups were active in Argentina, Chile, and Uruguay before 1920. These older parties usually were organized from the top down and consequently lacked local spontaneity.

Single-party and multiparty governments

After World War II all types of party systems were used. In the late 20th century a single party held an effective monopoly on public power in Mexico, Paraguay, Nicaragua, and Cuba. Multiparty arrangements existed in Argentina, Bolivia, Brazil, Chile, Costa Rica, Guatemala, Panama, Peru, and Venezuela.

Quasi-Fascist parties—Peronista in Argentina, Integralista in Brazil, Nacista in Chile, and Sinarquista in Mexico—have had large followings at times. The agrarian-populistic Aprista parties have indigenous roots. The Apristas have sought far-reaching social and economic reform and the integration of the lower classes into the political process. Acción Democrática, after World War II the predominant civilian party in Venezuela, is often considered to be an Aprista party. After 1959 a number of parties based on ideologies similar to that of Castro made their appearance, notably in Brazil and Venezuela.

The Roman Catholic Church. The Roman Catholic Church exerts a profound influence over the people of Latin America, 90 percent of whom consider themselves Catholic. Nevertheless, as indicated in the section dealing with the 19th century, the church's influence did not prevent it from coming under persistent and often violent attacks from anticlericals. After World War II, however, there was a notable lessening of tension, and in the late 20th century there was no major anti-church movement anywhere in Latin America, except in Cuba.

Three developments after World War I appear to have contributed significantly to the change. (1) National governments became stronger, which allowed them to be somewhat more tolerant of the opposition. (2) The middle class became politically powerful at the expense of the old allies of the church. In the light of this development the church reexamined the ideology of constitutional democracy that such groups advocated, and by 1960 the local churches had, in most instances, revised their position to agree with the church's official social doctrine of concern for the plight of the working and underprivileged classes. In giving practical application to this new position, the church cooperated with liberal democratic groups in the overthrow of the Argentine dictator, Juan Perón, the Colombian military chief, Gustavo Rojas Pinilla, and the Venezuelan tyrant, Marcos Pérez Jiménez. In 1961 the church went into open opposition against Fidel Castro in Cuba. (3) The Communists established their ability to create ideological confusion in the minds of the members of both the lower classes and the intellectual middle class. This development was of great concern to the church, which emphasized social and economic reform, including, at least in Colombia and Venezuela, land reform.

Education. The schools have been the stepchildren of the social and economic planners. The public neglect of schools accounts in large part for the fact that developments which might be expected eventually to produce basic changes in educational thinking have been slow in appearing, and the direction that some of them will finally take is not yet clear.

Except at the university level, where public institutions generally are superior, private schools continue to be preferred by those families who can afford to send their children to them. As has been the case historically, educational opportunities at every level are far greater in the urban centres than in the rural areas. Meanwhile, the advantages afforded in the cities serve to attract the best young minds from the farms and villages, and once in the cities the youths show little inclination to return to the backward areas from which they came. What education is available in the rural districts is often determined in the capital cities and may have little relevance to the needs of the students or, for that matter, of the nation. The final and perhaps most obvious tendency is for educational planning, administration, and financing to come under the control of the central governments, as have many other social and economic activities.

Summary. Certain broad generalizations may be drawn regarding Latin America in the late 20th century. Despite the rash of military coups that unseated constitutionally elected governments in Argentina, Peru, Guatemala, Ecuador, the Dominican Republic, Honduras, Bolivia, and Brazil, it was evident that the long-range trend in Latin America was toward more government by law and by the people. Appreciable progress had been made toward institutionalizing the electoral process. Issue-oriented political parties had grown in strength. These trends were in part counterbalanced by more demagoguery; the concentration of power in the central government at the expense of the states and municipalities; and the general failure to make inroads on the extended powers of the executive branch of government, a situation conducive to the continuance of dictatorship in the republics.

In the social-religious area the problems of the underprivileged had been given recognition; people of colour had penetrated the privileged classes without producing significant social clashes; and the Roman Catholic Church had found it easier to live at peace with the political leadership. However, little had been done to create a responsible labour movement; education had failed to keep pace with progress in other areas; and a large part of the lower classes continued to live marginal existences.

Industrial expansion had often been at the expense of agriculture; factories had failed to absorb all those entering the labour market; industry had not measurably improved the living levels of the working groups; and increased industrialization had, by increasing the need for raw materials, fuels, and capital goods, tied the republics ever closer to the world trading community.

The most recent outgrowth of these phenomena has been the emergence of a dynamic nationalism in the most advanced republics. The new nationalism is the instrument of the growing managerial and intellectual classes. It reflects the deep conviction among public figures that their nations can be developed only within a global framework of which they are a part. The new nationalism accepts the necessity for foreign aid but insists that such assistance must not mean economic submission and cultural dependence as it often has in the past. (J.J.Jo./Ed.)

BIBLIOGRAPHY

General works: CHARLES GIBSON, *Spain in America* (1966), the best short survey, incorporates the latest scholarship and has a valuable bibliographical essay; JOHN H. PARRY covers the same ground somewhat more amply in his well-written *The Spanish Seaborne Empire* (1966). SILVIO A. ZAVALA, *El mundo americano en la época colonial,* 2 vol. (1967), is a work of synthesis by a leading Mexican historian. STANLEY J. and BARBARA H. STEIN, *The Colonial Heritage of Latin America: Essays* *on Economic Dependence in Perspective* (1970), brief but insightful, shows the continuity of Latin-American economic patterns from colonial times to the present. Among older works, CLARENCE H. HARING, *The Spanish Empire in America* (1947, reprinted 1963), remains indispensable for colonial institutions. BAILEY W. DIFFIE, *Latin American Civilization: Colonial Period* (1945), is a work of formidable scholarship and sharply defined viewpoints on various controversial topics. For a comprehensive collection of source materials on the colonial period, see BENJAMIN KEEN (comp.), *Readings in Latin America Civilization, 1492 to the Present,* 2nd ed. (1967).

Spanish conquest of America: CARL O. SAUER, *The Early Spanish Main* (1966), critically examines the impact of the discovery on the Indian population of the Caribbean. WILLIAM H. PRESCOTT, *History of the Conquest of Mexico,* 3 vol. (1843, reprinted 1966), and *History of the Conquest of Peru,* 2 vol. (1908, reprinted 1963), remain classics, unsurpassed for breadth of conception and literary charm, but their romantic attitudes clearly reveal the books' age. JOHN HEMMING, *The Conquest of the Incas* (1970), supersedes all previous accounts.

Spain's colonial empire: PHILIP W. POWELL, *Soldiers, Indians, and Silver: The Northward Advance of New Spain, 1550–1600* (1952); and EUGENE H. KORTH, *Spanish Policy in Colonial Chile* (1968), examine Spanish–Indian relations. LEWIS U. HANKE, *The First Social Experiments in America* (1935), and *The Spanish Struggle for Justice in the Conquest of America* (1949), discuss the intellectual history of the Conquest; while his *Aristotle and the American Indians* (1959) studies the controversy between Las Casas and Juan Ginés de Sepúlveda over the nature and capacity of the Indians. For Messianic and Utopian influences in the Conquest, see JOHN L. PHELAN, *The Millenial Kingdom of the Franciscans in the New World,* rev. ed. (1969). For a reappraisal of Spain's Indian policy that questions some of Lewis Hanke's premises, see BENJAMIN KEEN, "The Black Legend Revisited: Assumptions and Realities," *Hispanic American Historical Review,* 49:703–719 (1969). On race relations in the colonial period, see the thoughtful survey of MAGNUS MORNER, *Race Mixture in the History of Latin America* (1967). Controversy surrounds the question of the size of pre- and post-Conquest Indian populations. S.F. COOK and WOODROW BORAH made an illuminating series of studies that indicate a very large Indian population in 1519, followed by a demographic catastrophe in the 16th and 17th centuries, consolidating their findings in *The Indian Population of Central Mexico, 1531–1610* (1960). For comparison, see the sharply divergent findings of ANGEL ROSENBLAT, *La población de América en 1492: Viejos y nuevos cálculos* (1967). FRANCOIS CHEVALIER, *La Formation des grands domaines au Mexique* (1952; Eng. trans., *Land and Society in Colonial Mexico: The Great Hacienda,* 1963), a landmark in the study of colonial land systems, traces the interplay of demographic and economic trends in the development of the hacienda. JAMES LOCKHART deftly analyzes the structure of Spanish society in colonial Peru in *Spanish Peru, 1532–1560* (1968).

The *encomienda* has been the subject of careful study. Two pioneer works are L.B. SIMPSON, *The Encomienda in New Spain,* rev. ed. (1950); and SILVIO A. ZAVALA, *La encomienda indiana* (1935). CLARENCE H. HARING, *Trade and Navigation Between Spain and the Indies in the Time of the Hapsburgs* (1918), remains a standard work. On land systems, see J.M. OTS CAPDEQUI, *España en América: El régimen de tierras en la época colonial* (1959). On mining, see MODESTO BARGALLO, *La minería y la metalurgía en la América Española durante la época colonial* (1955), for a general survey.

For the relations between church and state, see JOHN L. MECHAM, *Church and State in Latin America,* rev. ed. (1966); and WILLIAM E. SHIELS, *King and Church: The Rise and Fall of the Patronato Real* (1961).

Colonial Brazil: CAIO PRADO, *Formacão do Brasil Contemporañeo* (1942; Eng. trans., *The Colonial Background of Modern Brazil,* 1967), is an excellent introductory survey. See also CHARLES R. BOXER, *The Dutch in Brazil, 1624–1654* (1957), and *The Golden Age of Brazil, 1695–1750* (1962). DAURIL ALDEN, *Royal Government in Colonial Brazil* (1968), is a model monograph. GILBERTO FREYRE, *Casa Grande e Senzala,* 2 vol. (1943; Eng. trans., *The Masters and the Slaves,* 1946), a work of great literary charm, portrays Portuguese colonizers as almost free from colour prejudice and Brazilian slavery as relatively mild. For opposing points of view, see CHARLES R. BOXER, *Race Relations in the Portuguese Colonial Empire, 1415–1825* (1963); and MARVIN HARRIS, *Patterns of Race in the Americas* (1964). On the economics of colonial Brazil, see ALEXANDER MARCHANT, *From Barter to Slavery* (1942), which stresses the mixed feudal-capitalist character of early Brazilian economy; and CAIO PRADO, *op. cit.*

The Bourbon reforms and Spanish America: JOHN LYNCH, *Spanish Colonial Administration, 1782–1810* (1958), discusses colonial political reform. For the transmittal of Enlightenment ideas to Latin America, see the essays in ARTHUR P. WHITAKER (ed.), *Latin America and the Enlightenment,* 2nd ed. (1961).

Wars of independence: ROBIN A. HUMPHREYS, "The Fall of the Spanish American Empire," *History,* 37:213–277 (1952), is a perceptive interpretive essay, while Humphreys and JOHN LYNCH (eds.) bring together a variety of viewpoints in their anthology, *The Origins of the Latin American Revolutions, 1808–1826* (1965).

(B.K.)

Latin-American Literature

Latin-American literature consists of the national literatures of the Spanish-speaking countries of the Western Hemisphere and Portuguese-speaking Brazil. It also includes the literary expression of the highly developed Indian civilizations conquered by the Spaniards. Over the years Latin-American literature has developed into one of the finest literatures of the Western world, displaying a richness and diversity of themes, forms, and styles. A concise survey of its development is provided here.

The article is divided into the following sections:

THE COLONIAL YEARS (1492–1826)

Literature of the conquest (1492–1600). With the discovery of new lands beyond the ocean, Spain and Portugal embarked on a Christian crusade that was to stamp the colonial seal on vast areas of the Americas. This adventure was chronicled from the day Columbus set sail; his letters to King Ferdinand and Queen Isabella of Spain marked the beginnings of a rich body of colonial writings. The discovery and conquest are told in countless letters, chronicles, histories, polemical tracts, dictionaries, grammars, religious pieces, and epic poems. The great cultures discovered and conquered by the Spaniards also possessed a rich heritage of poetry, theatre, and mythicohistorical writing, the most poignant of which are their chronicles of the conquest and their defeat and destruction.

The thrill of the first contact and the adventures and problems that followed are recorded in various writings, including the five dispatches, or *Cartas de relación* (1519–26; *Five Letters*), sent by Hernán Cortés to his emperor, Charles V; the indignant account of Bernal Díaz del Castillo, who spoke for the common soldier in his *Verdadera historia de la conquista de la Nueva España* (1632; *The True History of the Conquest of New Spain*); the impassioned pages of the *Brevísima relación de la destrucción de las Indias* (1552; *The Tears of the Indian*) by the Dominican friar Bartolomé de las Casas, which became the cornerstone of the "black legend" of colonial Spain; and the observations on the tragic consequences of the struggle between cultures embodied in the *Comentarios reales que tratan del origen de los Incas* (1609–17; *First Part of the Royal Commentaries of the Yncas*) by Garcilaso de la Vega, a mestizo born of an Incan princess and a Spanish conqueror.

Epic beginnings

The conquerors did not chronicle their deeds in prose alone. Fired by the heroic fervour of their age and by the example of Ariosto, Tasso, and the Latin masters, they sang in epic measures their adventures from distant Antarctica to the upper reaches of the Río Grande. *La Araucana* (1569–89; *The Araucaniad*), the first and the best of these epics, was begun on scraps of paper and bark by the young captain and courtier Alonso de Ercilla y Zúñiga while he was serving with the Spanish armies fighting against the Araucanians of Chile. Its great attraction still is the immediacy of the experience and the balanced attitude toward the Indians. Enormously popular, *La Araucana* inspired similar efforts elsewhere: Juan de Castellanos, *Elegías de varones ilustres de Indias* (1588; "Elegies of the Illustrious Gentlemen of the Indies"); Martín del Barco Centenera, *La Argentina;* and Gaspar de Villagrá, *Conquista de la Nueva México* (1610). These later poems pale in poetic import as they gain in documentary significance. Only the *Arauco domado* (1596; *Arauco Tamed*) of Pedro de Oña, Chile's first outstanding native poet, approaches Ercilla's epic in artistic achievement. *La cristiada* (1611; "The Saviour"), written in Lima by the Dominican Diego de Ojeda, is often called the best sacred epic in the Spanish language. Finally there is Bernardo de Balbuena, whose work best represents the rich and varied literary expression of the dawning Baroque age. His epic of Spanish history, *Bernardo, o la victoria de Roncesvalles* (1624; "Bernard, or the Victory of the Roncesvalles"), and his eclogues are overshadowed by his *La grandeza mexicana* (1604; "The Grandeur of Mexico"), a paean to the attractions of the new metropolis of Mexico City.

Origin of Brazilian literature

The discovery and conquest of the tropical heartland of America produced no rich body of Brazilian counterparts to the spirited epics in prose and verse of the Spanish conquerors. The Portuguese found no centuries-old native civilizations upon which to build a new society. Indians were few, and, rather than resist, they retreated into the interior, leaving the invaders to settle at widely scattered points along the extended narrow coastal plain. There was, then, no epic clash, but there were the beautiful and bountiful coast and forest lands, and these set the tone even today for *brasilidade,* the love for the land of Brazil. Brazil's "literary baptism" is to be found in the letter of discovery sent to Dom Manuel I of Portugal by the scribe Pero Vaz de Caminha, who accompanied the explorer Pedro Álvares Cabral in 1500. The missionary labours were largely in the hands of the Jesuits, and one of their number, José de Anchieta, is hailed the father of Brazilian literature. For the period, however, there is only one epic, and this, except for some descriptive passages, sings of Portuguese exploits against the Moors in feeble imitation of Luís Vaz de Camões. The epic *Prosopopéa* (1601), by Bento Teixeira Pinto, is commonly regarded as the first book by a Brazilian-born writer.

Cultural influence of the homeland

The flowering of colonial letters (1600–1808). By 1600 the day of the conqueror was past, and the New World entered a less spectacular phase as colonial society stabilized. The first printing press in the Americas was established in Mexico City about 1539, another in Lima in 1584, and both cities were granted university charters in 1551. In Brazil no university was founded or printing permitted until the close of the colonial period. Everywhere in the Americas a growing leisure class devoted more time to intellectual and artistic pursuits, and close ties with the homeland encouraged the development of parallel literary patterns. Life became a reflection of trends across the Atlantic, and there was little in these two centuries that had originality, although it is possible to find the imprint of Indian traditions among even cultivated writers.

The theatre, emerging from medieval forms at the time of the conquest, served during the 16th century as a missionary medium in the Indian tongues for the conversion of the natives. There were also a few secular representations in the Indian languages of pre-Conquest memories. (Some of these were written by Creole authors.) For the bulk of the non-Indian population consisting of the Creoles and urbanized mestizos, however, the dramatic repertoire was mainly Spanish and Portuguese. Plays written in America were patterned after the drama of the Spanish Golden Age, and the greatest of the native dramatists, the Mexican Juan Ruiz de Alarcón, even lived and wrote in Spain.

Fiction, banned by the Spanish crown, only emerged with independence, leaving the field in prose to historical and biographical works. One of the best was *El lazarillo de ciegos caminantes* (*c.* 1773; *A Guide for Inexperienced Travelers*), a satire in the form of a travelogue of a journey between Buenos Aires and Lima. Its author, Alonso Carrió de la Vandera, was a Spanish postal official; he used the pseudonym Concolorcorvo to protect himself from official reprisal.

The writings of Sor Juana Inés de la Cruz

Poetry was the most popular form of literature, and a contest held in Mexico in 1585 attracted more than 300 mostly indifferent disciples of the Spanish fashion. Their poems are in the curious anthology *Ramillete de varias flores poéticas* (1675; "A Nosegay of Poetic Flowers"), gathered by Jacinto de Evia of Ecuador. One later poet stood out above all in the colonial world: the Mexican nun Sor Juana Inés de la Cruz. Her metaphysical poem "Primero sueño" ("First Dream") and, above all, her beautiful profane lyrics are among the greatest literary achievements of Latin America.

During the colonial period, there appeared a satirical tradition protesting against social and individual suffering: the Peruvian Juan del Valle y Caviedes and Gregório de Matos Guerra, the epigrammatic "devil's mouthpiece" of Brazil, both of whom were influenced by the Spanish satirist Francisco Gómez de Quevedo y Villegas. Their verse was widely circulated and avidly read. Toward the end of the colonial period, it developed popular forms that stirred the revolutionary spirit.

Literature of rebellion (1808–26). Political unrest spread in Latin America as the 18th century advanced. French ideas broke through the relaxed controls held by Spain and Portugal over New World thought. Early stirrings became more purposeful under the impact of the French and American revolutions. Printing presses and periodicals sprang up, and ideals quickened in literary societies founded by Latin America's young liberals.

One of the earliest fruits of contacts with foreign thought was an abortive Brazilian *inconfidência mineria* of 1789, a "conspiracy of poets" headed by Joaquim José da Silva Xavier and supported by a number of exceptional writers who were members of the Minas school of epic and Neoclassical poets, which had no equal in the Spanish colonies of that day. José Basílio da Gama's *Uraguai* (1769) and José de Santa Rita Durão's *Caramuru* (1781) were two native epics of Brazil. Love of country and the appearance of the Indian as a literary character stamped both works as forerunners of Brazilian intellectual independence as well as of incipient Romanticism. When, after a bloodless victory in 1822, Brazil emerged as an empire, only one literary figure, José Bonifácio de Andrade e Silva, stood out as a patriarch of Brazil's struggle for independence. Author of vigorous, passionate verse (*Poesias,* 1825), he is considered Brazil's first Romantic poet.

Francisco de Miranda, a Venezuelan leader of Spanish-American independence, left a remarkable journal that revealed the influence of his contacts in the United States with the "great American experiment." His compatriot Simón Bolívar was later christened "the thinker of the Revolution" for his prophetic analyses of the sociopolitical scene. Bolívar's writings exhibit some of the best of French thought, particularly that of Montesquieu and Rousseau. But the Mexican José Joaquín Fernández de Lizardi, referred to as *el pensador mexicano* ("the Mexican thinker"), is the only liberator remembered primarily as a man of letters. His fame rests largely on a picaresque tale, *El periquillo sarniento* (1816; *The Itching Parrot*), conceded to be the first Latin-American novel.

The revolution found popular expression in balladry and heroic verse. More enduring patriotic poetry came with

Patriotic poetry

Precursors of the Romantic movement

victory in the work of three outstanding poets. Of the three, only José Joaquín Olmedo, an Ecuadorean, limited his work almost exclusively to the themes and spirit of the revolutionary years. His best known poem, “La victoria de Junín, canto a Bolívar” (1825), is a fine example of heroic poetry in the classical style. In his Virgilian *Silva a la agricultura de la zona tórrida* (1826; *A Georgic of the Tropics*), Andrés Bello of Venezuela exhorted his fellow Americans to turn their swords into plowshares and to cultivate the natural riches of America. In later years Bello distinguished himself as a grammarian, critic, lawgiver, and educator and as the intellectual father of Chile's Romantic generation. José María Heredia of Cuba lived most of his life in political exile (Venezuela, New York City, Mexico), where nature's grander moods and the Indian past inspired his widely acclaimed poems *En el teocalli de Cholula* (1820) and *Oda al Níagara* (1824). He anticipated the new literary movement of Romanticism with which the young Latin-American nations were born to political independence.

THE FORMATIVE YEARS (1826–1910)

Emergence of gaucho literature

Romanticism. Political independence from Spain and Portugal did not bring freedom from political despotism and anarchy; economic and political stability for most new nations came late and with difficulty. American themes had fired the imagination of the liberators, but Neoclassic forms were still dominant. European Romanticism pointed the way to cultural independence also, even though that way lay largely along a route marked out by French, English, and Spanish writers. The controlled form and restraint of Neoclassicism yielded to the freedom, individualism, and emotional intensity of Romanticism, and the European cult of the medieval became in many cases a passion for the Indian—his present and past. The most illustrious early Romanticists were Argentine political refugees who fled from the dictator Juan Manuel de Rosas. Their leader was Esteban Echeverría, who, after a stay in France (1826–30) where Romanticism was at its height, showed the way in “La cautiva” (1837; “The Captive”), one of the earliest fusions of native themes and scenes with newer free verse forms. Domingo Faustino Sarmiento provided the first serious study of the great plain (or Pampas) and of gaucho (cowboy) lore in his *Facundo* (1845; *Life in the Argentine Republic in the Days of the Tyrants, or Civilization and Barbarism*), a novel written in passionate denunciation of Rosas. This emphasis on the national scene gave birth to an indigenous literary genre without European prototype, the gaucho literature of Argentina and Uruguay. The gaucho had long been the subject of folktale and ballad and soon figured in some of the best verse, as in Rafael Obligado's poem (1887) on the legendary minstrel Santos Vega, and the humorous *Fausto* (1866) by Estanislao del Campo, until finally he received epic treatment in *Martín Fierro* (1872–79; *The Gaucho, Martin Fierro*), by José Hernández.

This Romantic evocation of national themes and types reached its poetic climax while also pointing the way to the next period in the elegiac *Tabaré* (1886) by the Uruguayan Juan Zorrilla de San Martín. Zorrilla's epic poem related the fate of the aboriginal Charrúas, vanquished by Spanish invaders. In prose, the Colombian Jorge Isaacs wrote the popular lachrymose idyll *María* (1867), while Juan León Mera of Ecuador contributed *Cumandá* (1871) and Manuel de Jesús Galván of Santo Domingo added *Enriquillo* (1879–82) to a growing number of fictionalized portrayals of idealized Indian life. José Mármol gave a dramatic depiction of life in Argentina under the tyranny of Rosas in *Amalia* (1851–55). The *tradición*, patterned after the historical anecdote, was perfected by the Peruvian master of humorous prose, Ricardo Palma, whose *Tradiciones peruanas* (*The Knights of the Cape*) appeared between 1872 and 1910.

The Brazilian Romantics had no such “glorious” past to idealize and preferred to extol the natural beauties of their homeland and the simple Indian life. Domingo José Gonçalves de Magalhães, though still eclectic, launched the Romantic movement with *Suspiros Poéticos e Saudades* (1836; “Poetic Sighs and Longings”), but the best and most representative of Brazil's Romantic poets is Antônio Gonçalvez Dias. Later phases were exemplified in a poetry of doubt and despair by Manuel Antônio Alvares de Azevedo, author of *A Noite na Taverna* (*c.* 1851; “The Night in the Tavern”) and in the sociopolitical verse of Antônio de Castro Alves, author of *Os Escravos* (1876; “The Slaves”).

The theatre was not one of Latin America's strongest genres during the Romantic period. It consisted primarily of melodrama, regional comedy, and political pamphleteering. One exception was Cuba's Gertrudis Gómez de Avellaneda, who demonstrated that she was capable of serious psychological analysis and careful dramatic structure. She produced her plays in Spain, however, where she spent most of her life.

Cuadro de costumbres

Realism and Naturalism. The Romanticist's interest in the picturesque and unusual helped him discover evidence of a budding national way of life, and shortly after mid-century the *cuadro de costumbres* (sketch of contemporary customs) developed into a realistic novel of manners, often with an urban setting. From that time, the novel assumed a more commanding role in Latin-American letters, but it appeared almost concurrently in the several types representative of successive literary trends in Europe whose masters provided the molds into which Latin-American themes were poured. Alberto Blest Gana began producing a series of *costumbrista* novels on Chilean life, of which *Martín Rivas* (1862) was considered the best. Naturalism in the manner of Émile Zola made its appearance with the Argentine Eugenio Cambaceres. After mid-century several late-Romantic political writers, both Brazilian and Spanish-American, distinguished themselves in essay form. Among them were Juan Montalvo, Eugenio María de Hostos, Joaquim Nabuco, and Rui Barbosa. Manuel González Prada, an ironic experimental poet and essayist, was the chief figure, and his verse paved the way for the new poetry of a coming generation of rebels.

The true novel appeared first in Brazil in 1844 with *A Moreninha* (“The Little Brunette”) by Joaquim Manuel de Macedo. Still one of the most widely read of his country's novelists, José Martiniano de Alencar initiated a vogue of the Brazilian *Indianista* novel with *O Guarani* (1857; “The Guarani Indian”) and *Iracema* (1865). These romantic tales of love between Indian and white, however, represented only one aspect of Alencar's varied literary activity. He also turned to the life and customs of Brazil's backlands, and *O Gaúcho* (1870) and *O Sertanejo* (1876; “The Man of the Backlands”), though still markedly Romantic in spirit, were among the forerunners of a flourishing regional genre. Two other contributors to this transitional genre were Alfredo d'Escragnolle Taunay, whose *Inocência* (1872) became a universal favourite, and Bernardo Guimarães, whose abolitionist *Escrava Isaura* (1875) was a decisive step forward in the direction of the novel of social protest.

True Realism

True Realism, with a definite leaning toward Naturalism, was initiated by *Memórias de um Sargento de Milícias* (1854) by Manuel Antônio de Almeida, but it was not until the mid-1870s that the novel began to expose cankers of social and psychological maladjustments to a rapidly changing economic scene. Aluízio Azevedo, an early example of social protest in the manner of 20th-century novelists, wrote such favourites as *O Mulato* (1881) and *O Cortiço* (1890; *A Brazilian Tenement*). Less occupied with external aspects of Brazilian life, Joaquim Maria Machado de Assís pried into the psychological complex of the Brazilian and distinguished himself as his country's most original and gifted writer. His trilogy, *Bras Cubas* (1881; *Epitaph of a Small Winner*), *Quincas Borba* (1891; *Philosopher or Dog?*), and *Dom Casmurro* (1899), was a landmark in Latin-American letters.

Rubén Darío and the Modernists. In Spanish America, a measure of political and economic stability aided the emergence of a cosmopolitan awareness of life and letters that resulted in the revolt against the sentimental romantic writers who filled the literary pages of the newspapers and magazines. Young writers across the Americas immersed themselves in the mainstream of world thought and writing. Somewhat disparagingly labelled *modernistas* by the older generation, they wrote on exotic themes, often shutting themselves off from their immediate environment in artificial worlds of their own making—the ancient past, the distant Orient, and the lands of childhood fancy and sheer creation. Beauty was their goddess and "art for art's sake" their creed. Influenced by French movements, they followed no regular path; Symbolism, Parnassianism, Decadentism, and all the rest coexisted in any given individual or followed each other in any order.

Foremost among the early Modernists were the Mexican Manuel Gutiérrez Nájera, whose elegiac verse and restrained rhythmical prose sketches and tales best represented the transition from the excesses of Romanticism to the more filigreed Modernism; the Colombian José Asunción Silva, who wrote a small but influential body of savagely ironic and elegiac poems; the Cuban Julián del Casal, cultivator of the exquisite Parnassian sonnet; and his compatriot José Martí, martyr and symbol of Cuba's struggle for freedom from Spain, whose inspired prose style and deceptively simple, sincere verse set his work above and apart from all schools and movements.

The full flowering of Modernism, however, came under the leadership of one of the greatest poets in Spanish, Rubén Darío of Nicaragua. His collection of verse and prose, *Azul,* published in 1888, pointed the initial way, but *Prosas profanas* (1896; "Lay Hymns") represented the high point of the escapist, cosmopolitan phase of the movement. Darío blended the best of Modernist formal experimentation with an expression of inner despair or an almost metaphysical joy in *Cantos de vida y esperanza* (1905; "Songs of Life and Hope"). When Spain's empire crumbled in 1898 and mutual sympathy allayed the old distrust between Spain and its former colonies, he turned to Hispanic traditions as he had always turned to Hispanic forms; and in the face of U.S. imperialism, he spoke for Hispanic solidarity. Darío's imitators, particularly of his early experimental, escapist phase, were often slavish copiers who deserved the scornful name of *rubendaristas,* but Darío and his fellow Modernists brought about the greatest revitalization of language and poetic technique in Spanish since the 17th century. Many of his contemporaries were writers of considerable merit: Mexico's Amado Nervo, whose Orientally influenced mysticism was reflected in *Serenidad* (1914) and *Elevación* (1917); Peru's José Santos Chocano, whose exalted Americanism gave birth to *Alma América* (1906; "American Soul"); Bolivia's Ricardo Jaimes Freyre, who drew upon Scandinavian mythology for *Castalia bárbara* (1899; "The Barbarous Castalia"); Colombia's Guillermo Valencia, whose classic bent was manifest in *Ritos* (1898; "Rituals"); and Uruguay's philosopher and essayist José Enrique Rodó, whose *Ariel* (1900) distinguished him as the leading theoretician and exponent of Modernist ideals.

Modernist revitalization of the Spanish language and poetic technique

Even after Darío's death, the majority of Spanish-American Modernists continued to be spellbound by the verbal magic and brilliance of his *Prosas profanas.* There were, however, other gigantic figures: Leopoldo Lugones of Argentina underwent drastic shifts ranging from Laforguian irony through Baroque intensity to an intense nationalism expressing itself in poems based on popular folk songs; Enrique González Martínez of Mexico wrestled with social and ethical problems in Symbolist sonnets; and Julio Herrera y Reissig of Uruguay was perhaps the outstanding Symbolist of Modernism, finding new visions in a landscape represented in startling images.

In Portuguese-speaking Brazil, reaction against Romantic verse never produced the rich mosaic of Spanish-American Modernism. The Brazilian Parnassians with their formalistic, detached poetry were challenged by poets attracted to French Symbolism, but a Symbolist movement as such never materialized in Brazil. Parnassianism, as epitomized in the poetry of Raimundo Correia, Alberto de Oliveira, and Olavo Bilac, was unchanged until the developments of the 1920s altered Brazilian letters drastically.

Parnassianism in Brazil

LITERARY DEVELOPMENTS OF THE 20TH CENTURY

The horror and bloodshed of the Mexican Revolution (1910–17) shocked much of the complacent intellectual minority into a realization of the plight of their country's submerged masses: the Indian and mestizo (or person of European and Indian parentage) found able champions of their cause. The revolution had a similar effect almost everywhere throughout Latin America. Other events also played a significant role in altering the perspective and general orientation of Latin-American men of letters. World Wars I and II, along with the intervening worldwide economic depression of the 1930s and the Spanish Civil War, thrust their nations into the international scene. Given this situation, regionalistic preoccupations gave way to more universal concerns, a shift that was accompanied not only by new themes but also by new literary modes and stylistic techniques as well. The works that emerged during the second half of the century give testimony to the full maturing of Latin-American literature and its entry into the mainstream of Western letters.

Vanguard literature. Shortly after 1900 the Modernist leaders Darío and Lugones initiated a return from their formal and exotic innovations to more traditional forms and especially to themes dealing with the troubled external world. Their followers, however, never wholly rejected the very complex versification and language and countless new thematic sources with which Modernism had enriched Hispanic poetry. Prominent among them was an extraordinary group of women whose lyrics were more widely enjoyed than most of the work of male poets in the first half of the 20th century. Love in its impassioned and transcendental manifestations, maternal longing, and social protest were the themes of the Uruguayans Delmira Agustini and Juana de Ibarbourou and of the Chilean Gabriela Mistral, winner of the 1945 Nobel Prize for Literature. The Argentine Alfonsina Storni dealt with personal anguish and the difficulties of being a woman in a stern male world in a harsh, ironic idiom. Stemming from various aspects of the Modernist tradition were the allusive, interiorized lyrics of the Peruvian José María Eguren or the less complex subtleties of the Argentine Enrique Banchs; the idiosyncratic blend of novel imagery, Baudelairean struggle, and provincial themes of the Mexican Ramón López Velarde; or the brilliant sonnet sequence of the tropical scene in *Tierra de promisión* (1921; "Land of Promise") by the Colombian José Eustasio Rivera.

Women lyricists

Poetry in the modern period is of an extraordinary richness and complexity. Some groups, following the nihilistic waves of post-World War I "isms," experimented with free verse, often daring to use obscure imagery that gave a mistaken impression of a coldly intellectual mood. Many of these vanguard experimenters wrote out of a sociopolitical commitment as well. In the experiments of the Puerto Rican Luis Palés Matos, the Cuban mulatto poet Nicolás Guillén, and others, the voice and song of the African tradition were carried to a high artistic level.

Vicente Huidobro of Chile, who initiated the *creacionista* movement, is important for his insistence on the poet's total creation of an autonomous world, an approach that has been extremely influential on younger figures. Argentina's Jorge Luis Borges launched *ultraísmo* in Buenos Aires in 1921 but evolved from these avant-garde beginnings to a poetry reflecting a strong love for his city, a familiarity with numerous foreign literatures, and a preoccupation with metaphysical themes. These latter elements are also discernible in Borges' short stories, which have brought him international fame. César Vallejo of Peru fused social concerns with Surrealism and the heritage of Modernism to create an intensely subjective and often obscure but vital poetry. The Chilean Pablo Neruda, winner of the 1971 Nobel Prize for Literature, also blended Marxism with Surrealism in his earlier work; he attempted a poetic synthesis of the suffering of the Americas in the *Canto general* (1950; *General Song of Chile*). In his enormous production, Neruda made poetry of even the smallest aspects of the world he saw about him. It would be impossible to mention all the distinguished poets of the Post-Modernist period who were closely associated with such journals as *Martín Fierro* in Argentina, *Contemporáneos* in Mexico, and *Colónida* in Peru. Suffice it to say that the period from the beginnings of Modernism to the end of Post-Modernism rivals the Spanish Golden Age of the 16th and 17th centuries in wealth of poetry.

Borges and *ultraísmo*

The vanguard revolt in Brazil, usually referred to as Modernism (not at all the same as Spanish-American Modernism), broke away noisily from academicism and colonial cultural bondage at the noted Modern Art Week program in São Paulo in 1922. The Brazilian Modernists' primary aim was to modernize national thought and life, casting aside the persistent vestiges of the 19th century. This high-pitched, often theatrical, self-searching period of aesthetic reevaluation and analysis of the immediate Brazilian present served as a sorely needed purge and produced such important figures in Brazilian literature as the movement's high priest, Mário de Andrade, a gifted poet and musicologist; his lieutenant, Oswald de Andrade; Ronald de Carvalho, a critic and bard; and Manuel Bandeira, who has been acclaimed the country's greatest modern lyric poet. Preoccupation with social and metaphysical problems and an imperative urge toward untrammelled self-expression characterized the poetry of Modernist contemporaries or followers—namely, Jorge de Lima, Cecília Meireles, and Augusto Federico Schmidt.

Literature of social protest. Martí and González Prada were hailed as the intellectual progenitors of men aware of their responsibilities in guiding the Americas in a rapidly changing world. Rufino Blanco Fombona of Venezuela assailed his country's tyrants in *El hombre de hierro* (1907; "The Man of Iron") and *El hombre de oro* (1916; *The Man of Gold*). He was also among the first to attack "Yankee imperialism," abetted by a militant Argentine, Manuel Ugarte. The latter's *Porvenir de América latina* (1911; "The Future of Latin America") and *Destino de un continente* (1923; *The Destiny of a Continent*), together with the writings of the Peruvian Francisco García Calderón, envisioned Latin America as the future guardian of the Latin tradition. Ricardo Rojas, an Argentine literary historian and critic, and José Vasconcelos, a controversial Mexican philosopher and educator, were more concerned with racial and cultural aspects within the American family of nations. Alfonso Reyes, a Mexican poet, scholar, and critic, made the essay a vital, intimate force and raised it to a new level of artistic excellence.

It was in prose fiction that the grandeur of the American scene in all its drama was best described. The disheartening years following the abolition of slavery in 1888 and the establishment of a republic in 1889 made serious-minded Brazilians analyze their troubled homeland as an extraordinary amalgam of man, land, and climate. Euclides da Cunha revealed the bedrock of Brazilian life in the epic *Os Sertões* (1902; *Rebellion in the Backlands*); this work was the first written protest on behalf of Brazil's forgotten man, the emerging Brazilian of the backlands. In *Canaã* (1902; *The Canaan*), a novel of ideas, José Pereira da Graça Aranha focussed on the effects of recent European immigration upon this evolving Brazilian type. José Bento Monteiro Lobato's collection of short stories *Urupês* (1918; "Shelf Fungi") showed that intellectuals were still probing for native traits in a search that gave direction to the Modernist outburst of the 1920s.

Regionalist Brazilian literature

A new cultural regionalism of Brazil's "Northeastern school" that flowered after 1930 produced gifted prose writers, including a sociologist, Gilberto Freyre, whose *Casa Grande e Senzala* (1933; *The Masters and the Slaves*) was fundamental to an understanding of the region. José Lins do Rego, in a highly personal, evocative style, depicted the clash of the old and new way of life in his classic "Sugarcane" cycle and in *Pedra Bonita* (1938); Jorge Amado gave Brazil some of America's best proletarian literature in *Cacau* (1933), *Jubiabá* (1935), and *Terras do Sem Fim* (1942; *The Violent Land*). *Angústia* (1936), by Graciliano Ramos, attested to the fact that the individual inner struggle had also been adroitly plumbed by these new regionalists. Érico Veríssimo was one of Latin America's most distinguished cosmopolitan writers, as demonstrated in *Olhai os Lírios do Campo* (1938; *Consider the Lilies of the Field*) and *O Tempo e o Vento* (1950; *Time and the Wind*).

New trends in the Spanish-American novel

Contrary to its exceptional development in Brazil, the Spanish-American novel had been left largely to a few Modernist novelists, such as Manuel Díaz Rodríguez of Venezuela and Enrique Larreta of Argentina, and to the followers of the French Naturalist Émile Zola. Baldomero Lillo of Chile gave artistic dimension to the sufferings of oppressed miners in *Sub sole* (1906), and Federico Gamboa of Mexico, in *Santa* (1903), dealt with a prostitute reminiscent of Zola's protagonist in *Nana* from a Catholic point of view—the only version of Naturalism possible in the Hispanic world. Later novelists, essayists, and short-story writers developed new and more effective techniques under the influence of various foreign innovators, including Marcel Proust, James Joyce, Franz Kafka, and William Faulkner. The most widely acclaimed work that resulted from the Mexican Revolution was the novel *Los de abajo* (1915; *The Underdogs*) by Mariano Azuela, an army doctor of one of the bands of the revolutionary general Pancho Villa. Azuela chronicled the revolution from the point of view of the humble peasants who were its soldiers and its victims. From the late 1920s on, Martín Luis Guzmán, Gregorio López y Fuentes, and José Rubén Romero, among many others, gave further breadth and scope to this turbulent period without ever equalling Azuela's treatment.

The 20th-century Indian was not the uncompromising hero of the epic *La Araucana,* nor the symbol of colonial revolt against tyrannical Spain, and much less the "noble savage" of the untamed romantic wilderness. He was rather the victim of political and economic forces that kept the masses in abject bondage to colonial institutions. This new *indianista* literature had its roots in such novels as *Aves sin nido* (1889; *Birds Without a Nest*) by Clorinda Matto de Turner of Peru. The Indian's cause, moreover, was advanced by González Prada, a precursor of the militant pro-Indian, social reform party APRA (Alianza Popular Revolucionaria Americana) in 1923. The most extreme example of this committed literature was the brutally realistic school of Ecuadorean writers. Its most influential member, Jorge Icaza, produced mass-directed, vernacular novels such as *Huasipungo* (1934; *The Villagers*) and *En las calles* (1935; "In the Streets").

The effects of the tensions between the pampa and the city

The clash between the forces of nature and powerful economic pressures, between the land and the city, had by no means died with the passing of 19th-century local chiefs. The land and its resources were more zealously sought than ever before, and the struggle became more violent as man fought against man for the possession of these riches. In the growing urban centres, modern industrial economy exerted an even more insidious control over the destiny of the masses. The clash between the old and the new on the pampas of Argentina and Uruguay created the sombre descriptive pages of Uruguay's short-story writer Javier de Viana; the psychological portrayal of rural types in *El terruño* (1916; "The Native Country") by Carlos Reyles, also of Uruguay; the brilliant imagery of the mythic re-creation of the gaucho, *Don Segundo Sombra* (1926) by Ricardo Güiraldes of Argentina; and the tragic depths of the irreconcilable difference between the city and pampa in *El inglés de los güesos* (1924; "The English Archaeologist") by another Argentine, Benito Lynch. In *Doña Bárbara* (1929) Rómulo Gallegos gave a dramatic depiction of similar forces at work on the Venezuelan *llanos,* or "plains."

From the tropics appeared two artists who spoke for a growing number of younger writers who had discovered the *selva,* or "jungle": Horacio Quiroga of Uruguay, a consummate short-story artist, who excelled both in fantasy and in dramatic descriptions of the struggle for life in the jungle of Misiones, Argentina; and José Eustasio Rivera, a Colombian poet, whose sole prose work, *La vorágine* (1924; *The Vortex*), was a powerful denunciation of exploitation in the upper Amazon during the rubber boom of the early 1900s.

In *La maestra normal* (1914; "The Grade-School Teacher") Manuel Gálvez, an Argentine novelist, captured the pettiness and monotony of smaller provincial centres before modern mechanized manners shattered old colonial ways. Chile's underprivileged had two champions in the novelist Joaquín Edwards Bello and the short-story writer Manuel Rojas. Man's struggle with the deeper forces within and beyond himself was re-created in the psychological novels of the Chilean Eduardo Barrios and in the tales of the Cuban Alfonso Hernández Catá.

Major characteristics of contemporary Latin-American literature

Recent trends. The literature of the last half of the 20th century has been characterized by an increased preoccupation with man as the victim of alienating forces, solitude, identity, anguish, and evil and by a marked determination to create new forms and techniques. Above all, it has displayed a new language more responsive to the demands imposed by increasingly complex spiritual, social, and ideological concerns. These concerns found a direct expression in the essay, a form that has been cultivated with distinction by writers equally competent in other genres. Individual and collective preoccupation with analysis ran the gamut from the ruthless dissection of his country's ills by the Argentine Ezequiel Martínez Estrada in *Radiografía de la pampa* (1933; *X-Ray of the Pampa*) to the more social and psychological probings of the Mexican Octavio Paz (*El laberinto de la soledad* [1950; *The Labyrinth of Solitude*]). Paz is also notable as a poet; he has wedded the heritage of Surrealism with Indian and popular traditions and Oriental mysticism. Many poets veered from experimental and hermetic emphasis to seek answers to the same concerns in a more personalized style, often striving for a direct, conversational, and sometimes heavily ironic tone.

Experimentation in the theatre

The drama had enjoyed considerable popularity in the larger cities since the late 1800s, although relatively few plays were of real note. By the end of the century, Cuba and Argentina both had a flourishing satirical popular theatre based on regional types and language. In Argentina and Uruguay this led to an important regional realistic theatre anchored in social problems, particularly the complex impact of intense immigration. Its outstanding practitioner was the Uruguayan Florencio Sânchez. The late 1920s saw an outburst of experimental activity, reflecting influences ranging from Expressionism to Eugene O'Neill, Luigi Pirandello, and Jean Cocteau. Since that time, the Latin-American theatre has absorbed forces as varied as Arthur Miller, Tennessee Williams, Samuel Beckett, and Eugène Ionesco. Among the best of those leading the movement of renovation were the Brazilian Joracy Camargo (*Deus lhe Pague* [1932; "God Pays Them"]) and the Mexican Rodolfo Usigli (*El gesticulador* [1937] and *Corona de sombra* [1943; *Crown of Shadows*]), a caustic commentator of his compatriots' foibles and a seeker of a national self-comprehension. A younger generation, including Sebastián Salazar Bondy of Peru, Emilio Carballido of Mexico, Osvaldo Dragún of Argentina, Jorge Díaz of Chile, José Triana of Cuba, and Ariano Suassuna of Brazil won international recognition for its technical expertise and its humane responses to social problems. After the unrest of the late 1960s, the Latin-American theatre became increasingly politicized, and a series of groups under the inspiration of Enrique Buenaventura of Colombia and Augusto Boal of Brazil developed a theory and practice of theatre as a collective creation.

Predominance of prose fiction

Prose fiction has occupied the centre of the contemporary Latin-American literary scene. This new fiction has given full voice to social concerns through techniques that launched an unremitting attack on both the form of the novel and the structure of the language. The new narrative prose was already in evidence in the work of a small band of earlier writers, exemplified especially in the extraordinary tales of the Argentine Jorge Luis Borges (*El Aleph* [1949; *The Aleph and Other Stories, 1933–1969,* or *Labyrinths*]); in the terrifying portrait of tyranny, *El señor presidente* (1948; *The President*), of the Guatemalan Miguel Angel Asturias, winner of the 1967 Nobel Prize for Literature; in the Joycean examination of the social and psychological roots of the revolution of 1910 by the Mexican Agustín Yáñez (*Al filo del agua* [1947; *The Edge of the Storm*]); in the ideological and philosophical novels of the Argentine Eduardo Mallea (*La bahía de silencio* [1940; *The Bay of Silence*] and *Todo verdor perecerá* [1941; *All Green Shall Perish*]); in the vision of America as a new and different reality expressed in Baroque language by the Cuban Alejo Carpentier (*Los pasos perdidos* [1953; *The Lost Steps*]); in the earthy, enigmatic prose of the Brazilian João Guimarães Rosa (*Grande Sertão Veredas* [1956; *The Devil to Pay in the Backlands*]); and in the Rabelaisian romances of another Brazilian, Jorge Amado (*Gabriela, Cravo e Canela* [1958; *Gabriela, Clove and Cinnamon*]).

Principal among the creators of a new Latin-American literature was the Argentine Julio Cortázar, who explored interlocking realities in novels and short stories. His *Rayuela* (1963; *Hopscotch*) was a challenge to the whole concept of the structure of fiction. It was on the bedrock of such pioneers that the younger generation planted the invigorating and challenging contemporary novel. The leaders of this generation were the Colombian Gabriel García Márquez, winner of the 1982 Nobel Prize for Literature for his mythic re-creations of Spanish-American history and society, *Cien años de soledad* (1967; *One Hundred Years of Solitude*) and *Crónica de una muerte anunciada* (1981; *Chronicle of a Death Foretold*); the Peruvian José María Arguedas, preoccupied with the problem of the rural Indian and the linguistic interactions of Spanish and Indian tongues; the Mexican Carlos Fuentes, engaged in the effort to examine the interaction between the present and the still-living past in the Spanish-American psyche; the Peruvian Mario Vargas Llosa, author of massive re-creations of his homeland in structures learned from the romance of chivalry; and the Cubans Guillermo Cabrera Infante, Reinaldo Arenas, and José Lezama Lima and the Puerto Rican Luis Rafael Sânchez, each in his own way revolutionizing the structures of the language to re-create a highly personal vision of his nation and his people. Other leaders were Chile's José Donoso, who examines the decline of feudal society in novels that range from realistic through the bizarre; Manuel Puig of Argentina, who re-creates the impact of popular culture; or Juan Rulfo of Mexico, who transforms the arid highlands into complex metaphors of salvation and damnation.

BIBLIOGRAPHY. Several good anthologies of Latin-American literature in English translation are available: JOHN M. COHEN (ed.), *Latin American Writing Today* (1967), translations of prose and poetry from Argentina, Brazil, Chile, Colombia, Cuba, Mexico, Peru, and Uruguay; WILLIAM I. OLIVER (trans. and ed.), *Voices of Change in the Spanish American Theatre* (1971), a collection of six plays; GEORGE WOODYARD (ed.), *The Modern Stage in Latin America* (1971), a collection. Useful literary histories and critical studies include ISAAC GOLDBERG, *Brazilian Literature* (1922, reprinted 1978); LUIS HARSS and BARBARA DOHMANN, *Into the Mainstream: Conversations with Latin-American Writers* (1967, reprinted 1969); JEAN FRANCO, *An Introduction to Spanish American Literature* (1969); JOHN A. NIST, *The Modernist Movement in Brazil: A Literary Study* (1967); WILSON MARTINS, *The Modernist Idea: A Critical Survey of Brazilian Writing in the Twentieth Century* (1970, reprinted 1979; originally published in Spanish, 1969); DAVID P. GALLAGHER, *Modern Latin American Literature* (1973); JOHN S. BRUSHWOOD, *The Spanish American Novel: A Twentieth-Century Survey* (1975); DAVID WILLIAM FOSTER and VIRGINIA RAMOS FOSTER (eds.), *Modern Latin American Literature,* 2 vol. (1975), a critical commentary on more than 100 Latin-American writers; MARGARET SAYERS PEDEN (ed.), *The Latin American Short Story: A Critical History* (1983). Special topics are discussed in LUCÍA FOX-LOCKERT, *Women Novelists in Spain and Spanish America* (1979); RICHARD L. JACKSON, *Black Writers in Latin America* (1979); DAVID T. HABERLY, *Three Sad Races: Racial Identity and National Consciousness in Brazilian Literature* (1983); BRAULIO MUÑOZ, *Sons of the Wind: The Search for Identity in Spanish American Indian Literature* (1982). *The Handbook of Latin American Studies,* prepared by the Latin American, Portuguese, and Spanish Division of the Library of Congress, is an annual critical bibliography.

(J.E.E./F.N.D.)

Latin Literature

Latin literature was the product of the Roman Republic and the Roman Empire. When Rome fell, Latin remained the literary language of the Western medieval world until it was superseded by the Romance languages it had generated and by other modern languages. After the Renaissance the writing of Latin was increasingly confined to the narrow limits of certain ecclesiastical and academic publications. This article focuses primarily on ancient Latin literature. It does, however, provide a broad overview of the literary works produced in Latin by European writers during the Middle Ages and Renaissance.

The article is divided into the following sections:

Ancient Latin literature

Literature in Latin began as translation from the Greek, a fact that conditioned its development. Latin authors used earlier writers as sources of stock themes and motifs, at their best using their relationship to tradition to produce a new species of originality. They were more distinguished as verbal artists than as thinkers; the finest of them have a superb command of concrete detail and vivid illustration. Their noblest ideal was *humanitas,* a blend of culture and kindliness, approximating the quality of being "civilized" as the word is used today of individuals.

Little need be said of the preliterary period. Hellenistic influence came from the south, Etrusco-Hellenic from the north. Improvised farce, with stock characters in masks, may have been a native invention from the Campania region (the countryside of modern Naples). The historian Livy traced quasi-dramatic *satura* (medley) to the Etruscans. The statesman-writer Cato and the scholar Varro said that in former times the praises of heroes were sung after feasts, sometimes to the accompaniment of the flute, which was perhaps an Etruscan custom. If they existed, these *carmina convivalia,* or festal songs, would be behind some of the legends that came down to Livy. There were also the rude verses improvised at harvest festivals and weddings and liturgical formulas, whose scanty remains show alliteration and assonance. The nearest approach to literature must have been in public and private records and in recorded speeches.

STYLISTIC PERIODS

Ancient Latin literature may be divided into four periods: early writers, to 70 BC; Golden Age, 70 BC–AD 18; Silver Age, AD 18–133; and later writers.

Greek influence of Livius Andronicus

Early writers. The ground for Roman literature was prepared by an influx from the early 3rd century BC onward of Greek slaves, some of whom were put to tutoring young Roman nobles. Among them was Livius Andronicus, who was later freed and who is considered to be the first Latin writer. In 240 BC, to celebrate Rome's victory over Carthage, he composed a genuine drama adapted from the Greek. His success established a tradition of performing such plays alongside the cruder native entertainments. He also made a translation of the *Odyssey.* For his plays Livius adapted the Greek metres to suit the Latin tongue; but for his *Odyssey* he retained a traditional Italian measure, as did Gnaeus Naevius for his epic on the First Punic War against Carthage. Scholars are uncertain as to how much this metre depended on quantity or stress. A half-Greek Calabrian called Ennius adopted and Latinized the Greek hexameter for his epic *Annales,* thus further acquainting Rome with the Hellenistic world. Unfortunately his work survives only in fragments.

The Greek character thus imposed on literature made it more a preserve of the educated elite. In Rome coteries emerged such as that formed around the Roman consul and general Scipio Aemilianus. This circle included the statesman-orator Gaius Laelius; the Greek Stoic philosopher Panaetius; the Greek historian Polybius; the satirist Lucilius; and an African-born slave of genius, the comic playwright Terence. Soon after Rome absorbed Greece as a Roman province, Greek became a second language to educated Romans. Early in the 1st century BC, however, Latin declamation established itself, and, borrowing from Greek, it attained polish and artistry.

Plautus, the leading poet of comedy, is one of the chief sources for colloquial Latin. But Ennius sought to heighten epic and tragic diction, and from his time onward, with a few exceptions, literary language became ever more divorced from that of the people, until reaction came in the 2nd century AD.

The Ciceronian period

Golden Age, 70 BC–AD 18. The Golden Age of Latin literature spanned the last years of the republic and the virtual establishment of the Roman Empire under the reign of Augustus (27 BC–AD 14). The first part of this period, from 70 to 42 BC, is justly called the Ciceronian. It produced writers of distinction, most of them also men of action, among whom Julius Caesar stands out. The most prolific was Varro, "most learned of the Romans," but it was Cicero, a statesman, orator, poet, critic, and philosopher, who developed the Latin language to express abstract and complicated thought with clarity. Subsequently, prose style was either a reaction against, or a return to, Cicero's. As a poet, although uninspired, he was technically skillful. He edited the *De rerum natura* of the philosophic poet Lucretius. Like Lucretius, he admired Ennius and the old Roman poetry and, though apparently interested in Hellenistic work, spoke ironically of its extreme champions, the *neōteroi* ("newer poets").

After the destruction of Carthage and Corinth in 146 BC, prosperity and external security had allowed the cultivation of a literature of self-expression and entertainment. In this climate flourished the *neōteroi,* largely non-Roman Italians from the north, who introduced the mentality of "art for art's sake." None is known at first hand except Catullus, who was from Verona. These poets reacted against the grandiose, the Ennian tradition of "gravity," and their complicated allusive poetry consciously emulated the Callimacheans of 3rd-century Alexandria. The Neoteric influence persisted into the next generation through Cornelius Gallus to Virgil.

Virgil, born near Mantua and schooled at Cremona and Milan, chose Theocritus as his first model. The self-consciously beautiful cadences of the *Eclogues* depict shepherds living in a landscape half real, half fantastic; these allusive poems hover between the actual and the artificial. They are shot through with topical allusions, and in the fourth he already appears as a national prophet. Virgil was drawn into the circle being formed by Maecenas, Augustus' chief minister. In 38 BC he and Varius introduced the young poet Horace to Maecenas; and by the final victory of Augustus in 30 BC, the circle was consolidated.

The Augustan Age

With the reign of Augustus began the second phase of the Golden Age, known as the Augustan Age. It gave encouragement to the classical notion that a writer should not so much try to say new things as to say old things better. The rhetorical figures of thought and speech were mastered until they became instinctive. Alliteration and onomatopoeia (accommodation of sound and rhythm to sense), previously overdone by the Ennians and therefore eschewed by the *neōteroi,* were now used effectively with due discretion. Perfection of form characterizes the odes of Horace; elegy too became more polished.

The decade of the first impetus of Augustanism, 29–19 BC, saw the publication of Virgil's *Georgics* and the composition of the whole *Aeneid* by his death in 19 BC; Horace's *Odes,* books I–III and *Epistles,* book I; in elegy, books I–III of Propertius (also of Maecenas' circle) and books I–II of Tibullus, with others from the circle of Marcus Valerius Messalla Corvinus, and doubtless the first recitations by a still younger member of his circle, Ovid. About 28 or 27 BC Livy began his monumental history.

Maecenas' circle was not a propaganda bureau; his talent for tactful pressure guided his poets toward praise of Augustus and the regime without excessively cramping their freedom. Propertius, when admitted to the circle, was simply a youth with an anti-Caesarian background who had gained favour with passionate love elegies. He and Horace quarreled, and after Virgil's death the group broke up. Would-be poets now abounded, such as Horace's protégés, who occur in the *Epistles;* Ovid's friends, whom he remembers wistfully in exile; and Manilius, whom no one mentions at all. Poems were recited in literary circles and in public, hence the importance attached to euphony, smoothness, and artistic structure. They thus became known piecemeal and might be improved by friendly suggestions. When finally they were assembled in books, great care was taken over arrangement, which was artistic or significant (but not chronological).

Meanwhile, in prose the Ciceronian climax had been followed by a reaction led by Sallust. In 43 BC he began to publish a series of historical works in a terse, epigrammatic style studded with archaisms and avoiding the copiousness of Cicero. Later, eloquence, deprived of political influence, migrated from the forum to the schools, where cleverness and point counted rather than rolling periods. Thus developed the epigrammatic style of the younger Seneca and, ultimately, of Tacitus. Spreading to verse, it conditioned the witty couplets of Ovid, the tragedies of Seneca, and the satire of Juvenal. Though Livy stood out, Ciceronianism only found a real champion again in the rhetorician Quintilian.

Silver Age, AD 18–133. After the first flush of enthusiasm for Augustan ideals of national regeneration, literature paid the price of political patronage. It became subtly sterilized; and Ovid was but the first of many writers actually suppressed or inhibited by fear. Only Tacitus and Juvenal, writing under comparatively tolerant emperors, turned emotions pent up under Domitian's reign of terror into the driving force of great literature. Late Augustans such as Livy already sensed that Rome had passed its summit. Yet the title of Silver Age is not undeserved by a period that produced, in addition to Tacitus and Juvenal, the two Senecas, Lucan, Persius, the two Plinys, Quintilian, Petronius, Statius, Martial, and, of lesser stature, Manilius, Valerius Flaccus, Silius Italicus, and Suetonius.

Later writers. The decentralization of the empire under Hadrian and the Antonines weakened the Roman pride and passion for liberty. Romans began again to write in Greek as well as Latin. The "new sophistic" movement in Greece affected the "novel poets" such as Florus. An effete culture devoted itself to philology, archaism, and preciosity. After Juvenal, 250 years elapsed before Ausonius of Bordeaux (4th century AD) and the last of the true classics, Claudian (flourished about 400), appeared. The anonymous *Pervigilium Veneris* ("Vigil of Venus"), of uncertain date, presages the Middle Ages in its vitality and touch of stressed metre. Ausonius, though in the pagan literary tradition, was a Christian and contemporary with a truly original Christian poet, the Spaniard Prudentius. Henceforward, Christian literature overlaps pagan and generally surpasses it.

The *Meditations* of Marcus Aurelius

In prose these centuries have somewhat more to boast, though the greatest work by a Roman was written in Greek, the *Meditations* of the emperor Marcus Aurelius. *Elocutio novella,* a blend of archaisms and colloquial speech, is seen to best advantage in Apuleius (born about 125). Other writers of note were Aulus Gellius and Macrobius. The 4th century AD was the age of the grammarians and commentators, but in prose some of the most interesting work is again Christian.

THE GENRES

Comedy. Roman comedy was based on the New Comedy fashionable in Greece, whose classic representative was Menander. But whereas this was imitation of life to the Greeks, to the Romans it was escape to fantasy and literary convention. Livius' successor, Naevius, who developed this "drama in Greek cloak" (*fabula palliata*), may have been the first to introduce recitative and song, thereby increasing its unreality. But he slipped in details of Roman life and outspoken criticisms of powerful men. His imprisonment warned comedy off topical references, but the Roman audience became alert in applying ancient lines to modern situations and in demonstrating their feelings by appropriate clamour.

Plautus

Unlike his predecessors, Plautus specialized, writing only comedy involving high spirits, oaths, linguistic play, slapstick humour, music, and skillful adaptation of rhythm to subject matter. Some of his plays can be thought of almost as comic opera. Part of the fun consisted in the sudden intrusion of Roman things into this conventional Greek world. "The Plautine in Plautus" consists in pervasive qualities rather than supposed innovations of plot or technique.

Terence

As Greek influence on Roman culture increased, Roman drama became more dependent on Greek models. Terence's comedy was very different from Plautus'. Singing almost disappeared from his plays, and recitative was less prominent. From Menander he learned to exhibit refinements of psychology and to construct ingenious plots; but he lacked comic force, being uninterested in raising easy laughs. His pride was refined language—the avoidance of vulgarity, obscurity, or slang. His characters were less differentiated in speech than those of Plautus, but they talk with an elegant charm. The society Terence portrayed was more sensitive than that of Plautine comedy; lovers tended to be loyal and sons obedient. His historical significance has been enhanced by the loss of nearly all of Menander's work.

Though often revived, plays modeled on Greek drama were rarely written after Terence. The Ciceronian was the great age of acting, and in 55 BC Pompey gave Rome a permanent theatre. Plays having an Italian setting came into vogue, their framework being Greek New Comedy but their subject Roman society. A native form of farce was also revived. Under Julius Caesar, this yielded in popularity to verse mime of Greek origin that was realistic, often obscene, and full of quotable apothegms. Finally, when mime gave rise to the dumb show of the *pantomimus* with choral accompaniment and when exotic spectacles had become the rage, Roman comedy faded out.

Tragedy. Livius introduced both Greek tragedy (*fabula crepidata,* "buskined") and comedy to Latin. He was followed by Naevius and Ennius, who loved Euripides. Pacuvius, probably a greater tragedian, liked Sophocles, and heightened tragic diction even more than Ennius. His successor, Accius, was more rhetorical and impetuous. The fragments of these poets betoken grandeur in "the high Roman fashion," but they also have a certain ruggedness. They did not always deal in Greek mythology: occasionally they exploited Roman legend or even recent history. The Roman chorus, unlike the Greek, performed on stage and was inextricably involved in the action.

Classical tragedy was seldom composed after Accius, though its plays were constantly revived. Writing plays, once a function of slaves and freedmen, became a pastime of aristocratic dilettantes. Such writers had commonly no thought of production: post-Augustan drama was for reading. The nine extant tragedies of the younger Seneca probably were not written for public performance. They are melodramas of horror and violence, marked by sensational pseudo-realism and rhetorical cleverness. Characterization is crude, and philosophical moralizing obtrusive. Yet Seneca was a model for 16th- and early 17th-century tragedy, especially in France, and influenced English revenge tragedy.

Epic and epyllion. Livius' pioneering *Odyssey* was, to judge from the fragments, primitive, as was the *Bellum Punicum* of Naevius, important for Virgil because it began with the legendary origins of Carthage in Phoenicia and Rome in Troy. But Ennius' *Annales* soon followed. This compound of legendary origins and history was in Latin, in a transplanted metre, and by a poet who had imagination and a realization of the emergent greatness of Rome. In form his work must have been ill-balanced; he almost ignored the First Punic War in consideration of Naevius and became more detailed as he added books about his own times. But his great merit shines out from the fragments—nobility of ethos matched with nobility of language. On receptive spirits, such as Cicero, Lucretius, and Virgil, his influence was profound.

Virgil's *Aeneid*

Little is known of the "strong epic" for which Virgil's friend Varius is renowned, but Virgil's *Aeneid* was certainly something new. Recent history would have been too particularized a theme. Instead, Virgil developed Naevius' version of Aeneas' pilgrimage from Troy to found Rome. The poem is in part an Odyssey of travel (with an interlude of love) followed by an Iliad of conquest, and in part a symbolic epic of contemporary Roman relevance. Aeneas has Homeric traits but also qualities that look forward to the character of the Roman hero of the future. His fault was to have lingered at Carthage. The command to leave the Carthaginian queen Dido shakes him ruthlessly out of the last great temptation to seek individual happiness. But it is only the vision of Rome's future greatness, seen when he visits Elysium, that kindles obedient acceptance into imaginative enthusiasm. It was just such a sacrifice of the individual that the Augustan ideal demanded. The second half of the poem represents the fusing in the crucible of war of the civilized graces of Troy with the manly virtues of Italy. The tempering of Roman culture by Italian hardiness was another part of the Augustan ideal. So was a revival of interest in ancient customs and religious observances, which Virgil could appropriately indulge. The verse throughout is superbly varied, musical, and rhetorical in the best sense.

With his *Hecale,* Callimachus had inaugurated the short, carefully composed hexameter narrative (called epyllion by modern scholars) to replace grand epic. The *Hecale* had started a convention of insetting an independent story. Catullus inset the story of Ariadne on Naxos into that of the marriage of Peleus and Thetis, and the poem has a mannered, lyrical beauty. But the story of Aristaeus at the end of Virgil's *Georgics,* with that of Orpheus and Eurydice inset, shows what heights epyllion could attain.

Ovid's *Metamorphoses*

Ovid's *Metamorphoses* is a nexus of some 50 epyllia with shorter episodes. He created a convincing imaginative world with a magical logic of its own. His continuous poem, meandering from the creation of the world to the apotheosis of Julius Caesar, is a great Baroque conception, executed in swift, clear hexameters. Its frequent irony and humour are striking. Thereafter epics proliferated. Statius' *Thebaid* and inchoate *Achilleid* and Valerius' *Argonautica* are justly less read now than they were. Lucan's unfinished *Pharsalia* has a more interesting subject, namely the struggle between Caesar and Pompey, whom he favours. He left out the gods. His brilliant rhetoric comes close to making the poem a success, but it is too strained and monochromatic.

Didactic poetry. Ennius essayed didactic poetry in his *Epicharmus,* a work on the nature of the physical universe. Lucretius' *De rerum natura* is an account of Epicurus' atomic theory of matter, its aim being to free men from superstition and the fear of death. Its combination of moral urgency, intellectual force, and precise observation of the physical world makes it one of the summits of classical literature.

This poem profoundly affected Virgil, but his poetic reaction was delayed for some 17 years; and the *Georgics,* though deeply influenced by Lucretius, were not truly didactic. Country bred though he was, Virgil wrote for literary readers like himself, selecting whatever would contribute picturesque detail to his impressionistic picture of rural life. The *Georgics* portrayed the recently united land of Italy and taught that the idle Golden Age of the fourth *Eclogue* was a mirage: relentless work, introduced by a paternal Jupiter to sharpen men's wits, creates "the glory of the divine countryside." The compensation is the infinite variety of civilized life. Insofar as it had a political intention, it encouraged revival of an agriculture devastated in wars, of the old Italian virtues, and of the idea of Rome's extending its works over Italy and civilizing the world.

Ovid's *Ars amatoria* was comedy or satire in the burlesque guise of didactic, an amusing commentary on the psychology of love. The *Fasti* was didactic in popularizing the new calendar; but its object was clearly to entertain.

Satire. *Satura* meant a medley. The word was applied to variety performances introduced, according to Livy, by the Etruscans. Literary satire begins with Ennius, but it was Lucilius who established the genre. After experimenting, he settled on hexameters, thus making them its recognized vehicle. A tendency to break into dialogue may be a vestige of a dramatic element in nonliterary *satura.* Lucilius used this medium for self-expression, fearlessly criticizing public as well as private conduct. He owed much to the Cynic-Stoic "diatribes" (racy sermons in prose or verse) of Greeks such as Bion; but in extant Hellenistic literature he is most clearly presaged by the fragments of Callimachus' iambs. "Menippean" satire, which descended from the Greek prototype of Menippus of Gadara and mingled prose and verse, was introduced to Rome by Varro.

The satire of Horace and Juvenal

Horace saw that satire was still awaiting improvement: Lucilius had been an uncouth versifier. *Satires* I, 1–3 are essays in the Lucilian manner. But Horace's nature was to laugh, not to flay, and his incidental butts were either insignificant or dead. He came to appreciate that the real point about Lucilius was not his denunciations but his self-revelation. This encouraged him to talk about himself. In *Satires* II he developed in parts the satire of moral diatribe presaging Juvenal. His successor Persius blended Lucilius, Horace, diatribe, and mime into pungent sermons in verse. The great declaimer was Juvenal, who fixed the idea of satire for posterity. Gone was the personal approach of Lucilius and Horace. His anger may at times have been cultivated for effect, but his epigrammatic power and brilliant eye for detail make him a great poet.

The younger Seneca's *Apocolocyntosis* was a medley of prose and verse, but its pitiless skit on the deification of the emperor Claudius was Lucilian satire. The *Satyricon* of Petronius is also Menippean inasmuch as it contains varied digressions and occasional verse; essentially, however, it comes under fiction.

With Lucilian satire may be classed the fables of Augustus' freedman Phaedrus, the Roman Aesop, whose beast fables include contemporary allusions.

Iambic, lyric, and epigram. The short poems of Catullus were called by himself *nugae* ("trifles"). They vary remarkably in mood and intention, and he uses iambic metre normally associated with invective not only for his abuse of Caesar and Pompey but also for his tender homecoming to Sirmio. Catullus alone used the hendecasyllable, the metre of skits and lampoons, as a medium for love poetry.

Horace was a pioneer. In his *Epodes* he used iambic verse to express devotion to Maecenas and for brutal invective in the manner of the Greek poet Archilochus. But his primary aim was to create literature, whereas his models had been venting their feelings. In the *Odes* he adapted other Greek metres and claimed immortality for introducing early Greek lyric to Latin. The *Odes* rarely show the passion now associated with lyric but are marked by elegance, dignity, and the studied perfectionism of their craftsmanship.

Martial went back to Catullus for his metres and his often obscene wit. He fixed the notion of epigram for posterity by making it characteristically pointed. Occasionally his poems are touching.

Catullus' love poems

Elegy. The elegiac couplet of hexameter and pentameter (verse line of five feet) was taken over by Catullus, who broke with tradition by filling elegy with personal emotion. One of his most intense poems in this metre, about Lesbia, extends to 26 lines; another is a long poem of involved design in which the fabled love of Laodameia for Protesilaus is incidentally used as a paradigm. These two poems make him the inventor of the "subjective" love elegy dealing with the poet's own passion. Gallus, whose work is lost, established the genre; Tibullus and Propertius smoothed out the metre.

Propertius' first book is still Catullan in that it seems genuinely inspired by his passion for Cynthia: the involvement of Tibullus is less certain. Later Propertius grew more interested in manipulating literary conventions. Tibullus' elegy is constructed of sections of placid couplets with subtle transitions. These two poets established the convention of the "soft poet," valiant only in the campaigns of love, immortalized through them and the Muses. Propertius was at first impervious to Augustan ideals, glorying in his abject slavery to love and his naughtiness (*nequitia*), though later he became acclimatized to Maecenas' circle.

Tibullus, a lover of peace, country life, and old religious customs, had grace and quiet humour. Propertius, too, could be charming, but he was far more. He often wrote impetuously, straining language and associative sequence with passion or irony or sombre imagination.

Ovid's aim was not to unburden his soul but to entertain. In the *Amores* he is outrageous and amusing in the role adopted from Propertius, his Corinna being probably a fiction. Elegy became his characteristic medium. He carried the couplet of his predecessors to its logical extreme, characterized by parallelism, regular flow and ebb, and a neat wit.

OTHER LANGUAGE AND LITERARY ART FORMS

Cicero's influence on oratory

Rhetoric and oratory. Speaking in the forum and law courts was the essence of a public career at Rome and hence of educational practice. After the 2nd century BC, Greek art affected Latin oratory. The dominant style in Cicero's time was the "Asiatic"—emotional, rhythmical, and ornate. Cicero, Asiatic at first, early learned to tone down his style. Criticized later by the revivers of plain style, he insisted that style should vary with subject. But in public speaking he held that crowds were swayed less by argument than emotion. He was the acknowledged master speaker from 70 BC until his death (43 BC). He expounded the history of Roman oratory in the *Brutus* and his own methods in the *De oratore.*

The establishment of monarchy robbed eloquence of its public importance, but rhetoric remained the crown of education. Insofar as this taught boys to marshal material clearly and to express themselves cogently, it performed the function of the modern essay; but insofar as the temptations of applause made it strained and affected, it did harm.

In the *De oratore,* Cicero had pleaded that an orator's training should be in all liberal arts. Education without rhetoric was inconceivable; but what Cicero was proposing was to graft onto it a complete system of higher education. Quintilian, in his *Institutio oratoria,* went back to Cicero for inspiration as well as style. Much of that work is conventional, but the first and last books in particular show admirable common sense and humanity; and his work greatly influenced Renaissance education.

History. Quintus Fabius Pictor wrote his pioneering history of Rome during the Second Punic War, using public and private records and writing in Greek. His immediate successors followed suit. Latin historical writing began with Cato's *Origines.* After him there were as many historiasters, or worthless historians, as the poetasters disdained by Cicero. The first great exception is Caesar's *Commentaries,* a political apologia in the guise of unvarnished narrative. The style is dignified, terse, clear, and unrhetorical.

Sallust took Thucydides as his model. He interpreted, using speeches, and ascribed motives. In his extant monographs *Bellum Catilinae* and *Bellum Jugurthinum,* he displays a sardonic moralism, using history to emphasize the decadence of the dominant caste. The revolution in style he inaugurated gives him importance in an account of Latin literature.

Livy began his 40 years' task as Augustus came to power. His work consummated the annalistic tradition. If in historical method he fell short of modern standards, he had the literary virtues of a historian. He could vividly describe past events and interpret the participants' views in eloquent speeches. He inherited from Cicero his literary conception of history, his copiousness, and his principle of accommodating style to subject. Indeed, he was perhaps the greatest of Latin stylists. His earlier books, where his imagination has freer play, are the most readable. In the later books, the more historical the times become, the more disturbing are his uncritical methods and his patriotic bias. Livy's work now is judged mainly as literature.

The histories of Tacitus

Tacitus, on the other hand, stands higher now than in antiquity. Though his anti-imperial bias in attributing motives is plain, his facts can rarely be impugned; and his evocation of the terrors of tyranny is unforgettable. He is read for his penetrating characterizations, his drama, his ironical epigrams, and his unpredictability. His is an extreme development of the Sallustian style, coloured with archaic and poetic words, with a careful avoidance of the commonplace.

Suetonian biography apart, historiography thereafter degenerated into handbooks and epitomes until Ammianus Marcellinus appeared. He was refreshingly detached, rather ornate in style, but capable of vivid narrative and description. He continued Tacitus' account from Domitian's death to AD 378, more than half his work dealing with his own times.

Biography and letters. The idea of comparing Romans with foreigners was taken up by Cornelius Nepos, a friend of Cicero and Catullus. Of his *De viris illustribus* all that survive are 24 hack pieces about worthies long dead and one of real merit about his friend Atticus. The very fact that Atticus and Tiro decided to publish nearly 1,000 of Cicero's letters is evidence of public interest in people. Admiration of these fascinating letters gave rise to letter writing as a literary genre. The younger Pliny's letters, anticipating publication, convey a possibly rose-tinted picture of civilized life. They are nothing to his spontaneous correspondence with Trajan, where one learns of routine problems, for instance with Christians confronting a provincial governor in Bithynia. The letter as a verse form, beginning with striking examples by Catullus, was established by Horace, whose *Epistles* carry still further the humane refinement of his gentler satires.

Suetonius' biographies

Suetonius' lives of the Caesars and of poets contain much valuable information, especially since he had access to the imperial archives. His method was to cite in categories whatever he found, favourable or hostile, and to leave this raw material to the judgment of the reader. The *Historia Augusta,* covering the emperors from 117 to 284, is a collection of lives in the Suetonian tradition. Tacitus' *Agricola* was an admiring, but not necessarily overcoloured, biographical study.

Some of the most valuable autobiography was incidental, such as Cicero's account of his oratorical career in the *Brutus.* Horace's largely autobiographical *Epistles* I was sealed with a miniature self-portrait. Ovid, in exile and afraid of fading from Rome's memory, gave an invaluable account of his life in *Tristia* IV.

Philosophical and learned writings. The practical Roman mind produced no original philosopher. Apart from Lucretius the only name that demands consideration is Cicero's. He was trained at Athens in the eclectic New Academy, and eclectic he apparently remained, seeking a philosophy to fit his own constitution rather than a logical system valid for all. He used the dialogue form, avowedly in order to make people think for themselves instead of following authority. Essentially, he was a philosophic journalist, composing works that became one of the means by which Greek thought was absorbed into early Christian thinking. The *De officiis* is a treatise on ethics. The dialogues do not follow the Platonic, or dialectic, pattern but the Aristotelian, in which speakers expounded already formed opinions at greater length.

Nor were the Romans any more original in science. Instead, they produced encyclopaedists such as Varro and Celsus. Pliny's *Natural History* is a fascinating ragbag, especially valuable for art history, though it shows to what extent Hellenistic achievement in science had become confused or lost.

Literary criticism. Cicero's *Brutus* and the 10th book of Quintilian's *Institutio oratoria* provide examples of general criticism. Cicero stressed the importance of a well-stocked mind and native wit against mere handbook technique. By Horace's day, however, it had become more timely to insist on the equal importance of art. Some of Horace's best criticism is in the *Satires* (I, 4 and 10; II, 1), in the epistle to Florus (II, 2), and in the epistle to Augustus (II, 1), a vindication of the Augustans against archaists. But it was his epistle to a Piso and his sons (later called *Ars poetica*) that was so influential throughout Europe in the 18th century. It supported, among acceptable if trite theses, the dubious one that poetry is necessarily best when it mingles the useful (particularly moral) with the pleasing. Much of the work concerned itself with drama. The Romans were better at discussing literary trends than fundamental principles—there is much good sense about this in Quintilian, and Tacitus' *Dialogus* is an acute discussion of the decline of oratory.

Horace's *Ars poetica*

Fiction. Republican and early imperial Rome knew no Latin fiction beyond such things as Sisenna's translation of Aristides' *Milesian Tales*. But two considerable works have survived from imperial times. Of Petronius' *Satyricon*, a rambling picaresque novel, one long extract and some fragments remain. The disreputable characters have varied adventures and talk lively colloquial Latin. The description of the vulgar parvenu Trimalchio's banquet is justly famous. Apuleius' *Metamorphoses* (*The Golden Ass*) has a hero who has accidentally been changed into an ass. After strange adventures he is restored to human shape by the goddess Isis. Many passages, notably the story of Cupid and Psyche, have a beauty that culminates in the apparition of Isis and the initiation of the hero into her mysteries. (L.P.Wi./R.H.A.J.)

Medieval Latin literature

From about 500 to 1500 Latin was the principal language of the church, as well as of administration, theology, philosophy, science, history, biography, and belles lettres, and medieval Latin literature is therefore remarkably rich. Two themes dominate the linguistic and literary development of medieval Latin: its close and creative adaptation of the classical heritage from which it emerged and its changing relationship with the medieval vernacular languages. Within these two broad themes a number of subsidiary yet significant strains can be distinguished: the emergence of national characteristics in the Latin literature produced in different parts of Europe; the refinement of the polarity between popular and learned Latin by the clergy's use of a colloquialism intelligible to its audience as a lingua franca; and the effect of certain periods of special vigour and artistic self-awareness, such as the Carolingian revival of the 8th and 9th centuries and the new impulse given to learned and vernacular literature in the 12th.

THE 3RD TO THE 5TH CENTURY: THE RISE OF CHRISTIAN LATIN LITERATURE

The early history of medieval Latin literature is in part the story of the reception of the classical past by the Christians, to whom it represented secular culture. Old forms and genres were continuously renewed over the millennium following the entrance of Christians to the circle of literary production, dated for convenience to the conversion of Constantine to Christianity (in about AD 313). For example, the Latin epic persisted in recognizable form throughout the period, and its authors remained in continuous contact with the great classical exponents Lucan, Statius, and, above all, Virgil. From the 4th century, the degree of scholarly interpretation applied to these epic poets, especially Virgil, was intensified. Virgilian technique was imitated by many poets, among them the 4th-century Spaniard Juvencus, who versified a portion of the Bible, and the author of the epic poem *Waltharius* (probably 9th century), written in hexameters.

Persistence of the Latin epic

Even before the conversion of Constantine, Christians were developing new forms of literature, which persisted throughout the ensuing centuries. The production of hagiographical texts (lives of the saints) was widespread in the Middle Ages. The first Acts of the Martyrs in Latin were written during the 3rd century, and the flowering of the form after the end of the period of persecution of Christians shows the powerful appeal that it exercised at all levels of society. The *Passio Sanctarum Perpetuae et Felicitatis* (*The Passion of S. Perpetua*), written in a style that owes little to classical precedent, is a distinctive early example of the genre.

The 3rd and 4th centuries were above all an age of translation. Among the Greek patristic writings diffused to a wider audience in the West in Latin versions, the lives of the Desert Fathers occupied an important place. The Latin translation by Evagrius, bishop of Antioch, of Athanasius' *Life of Saint Antony* enjoyed the widest transmission, and its influence is as marked by contrast in the early Latin Lives of the Saints as it is by imitation. Sulpicius Severus' biography of St. Martin, an original Latin work, greatly influenced hagiography over many centuries. (A further, equally influential example of the genre was the *Dialogues* of Pope Gregory the Great, written in about 593.)

The most important work of translation appeared at the end of the 4th century: the Vulgate, completed by the monastic leader Jerome, replaced sporadic earlier attempts to render the Bible into Latin. The idiom and style of the Bible's original languages were apparent through the veil of Jerome's Latin, however, and provided a counterweight to the classical styles that continued to be taught and practiced through the schools in the West. Exegesis of the text occupied many of the greatest minds of the Middle Ages for the largest part of their careers, and the literary work of many major authors, from Augustine and Gregory to Bede, reflects their individual understanding of Scripture.

The early Christian liturgy also gave birth to new forms of literature. From the ancient practice of psalmody in the churches derives the hymn. Ambrose, bishop of Milan in the second half of the 4th century, wrote the earliest prosaic hymns, which incorporated nonliturgical texts into the mass to be sung by the congregation. These were rapidly imitated, notably by the Spanish poet Prudentius at the end of the century, and remained in continuous use in churches and monasteries for more than a millennium.

A major problem of Christian thinkers in these centuries was the integration of the history of the pagan empire with the history of salvation. Synthesis and epitome of biblical and classical history appeared in the *Historiarum adversus paganos libri VII* (*7 Books of Histories Against the Pagans*) of Orosius and the briefer *Chronica* (*c.* 402–404) of Sulpicius Severus. On a larger scale, Augustine's *De civitate Dei* (*The City of God*) offered a comprehensive view of past history, the present, and the world to come in the light of scriptural revelation. His spiritual autobiography, the *Confessiones* (*Confessions*), was an exploration of the philosophical and emotional development of an individual soul. The distinctive originality of this work owed little to classical autobiography and was unmatched by later imitations.

Writings of Augustine

The Gallic schools of the 5th century gave rise to a literary culture unique in this period. Versification of the Bible developed a new degree of exegetical and stylistic refinement, while the letters of Paulinus of Nola and Sidonius Apollinaris, bishop of Auvergne, display a picture of cultivated aristocratic and ecclesiastical society. Both men were also admired as poets, Sidonius in particular as an encomiast. On the secular side, at the beginning of the century in Rome the Egyptian poet Claudian produced the most elaborate examples of imperial verse panegyric to a succession of dignitaries. His *Raptus Proserpinae* (*c.* 400; *The Rape of Proserpine*) is one of the last examples of an extended narrative in verse that dwells wholly in the world of pagan mythology.

THE 6TH TO THE 8TH CENTURY

Gaul's literary history is interrupted by the Frankish invasions, though there are signs that abbots and bishops began to perceive the benefit of using literature to promote the cults of local saints. Two figures of note are Gregory of Tours and Venantius Fortunatus, bishop of Poitiers. In addition to a vast corpus of hagiography, Gregory produced the monumental *Historia Francorum* (605–664; *History of the Franks*), the most extensive history of a barbarian people that had yet been written. He set the arrival of the Franks in Gaul, and their recent past, in the perspective of universal history. An element of local patriotism is also discernible. Gregory was one of the many patrons who inspired the poet Fortunatus, whose astute and pliable talent achieved distinction in both secular panegyric and hymnody. His hagiography, in verse and in prose, also occupies a prominent place in his output. His style exercised a powerful appeal upon the poets of the Carolingian renaissance.

Three figures of encyclopaedic learning dominate the literature of the 6th and 7th centuries. In the course of his long retirement from a career in public service under the Ostrogothic kings in Italy, Cassiodorus combined zealous preservation of the literature of the classical past with an enormously influential educational plan. His late 6th-century compendium of sacred and secular learning, *Institutiones divinarum et humanarum lectionum* (*An Introduction to Divine and Human Readings*), was among the shaping influences upon monastic culture. The Roman Boethius, a Neoplatonist philosopher, wrote on arithmetic and music, but his most popular and influential work was *De consolatione philosophiae* (1882–91; *The Consolation of Philosophy*), written in about 524, when Boethius was imprisoned under sentence of execution. The Spaniard Isidore produced a series of encyclopaedic compilations that were used as repositories of diverse learning by later centuries. It was midway through the 6th century that the last major Latin work was produced in the Eastern Empire: the epic *Iohannis* of the African poet Corippus.

Writings of the Venerable Bede

The conversion of the Saxons began to bear literary fruit during the 7th and early 8th centuries. In an elaborate and allusive style, Aldhelm, bishop of Sherborne, wrote, first in prose and later in verse, a treatise on sainthood called *De Virginitate.* In the kingdom of Northumbria, particularly open to influence of Irish monastic learning, St. Bede the Venerable devoted his life to scholarship. The culmination of his work is the *Historia ecclesiastica gentis Anglorum* (*The Venerable Bede's Ecclesiastical History of England*), completed in 731. Synthesized from a variety of sources, literary and nonliterary, the work charts the involvement of God with the English people and the relation of the English church to the Christian world centred on Rome.

THE CAROLINGIAN RENAISSANCE

The revival of letters, accompanied by wide-scale copying of classical texts, to which the reign of Charlemagne (768–814) gave fresh impetus, produced some of the most brilliant literary achievements of the Latin Middle Ages. An international elite of scholars, among whom the most distinguished were the Anglo-Saxon Alcuin, the Visigoth Theodulf of Orléans, and the Italians Paulinus of Aquileia and Paul the Deacon, produced a body of lyric, epic, and didactic poetry (both sacred and secular, both religious and political) unmatched in the earlier period. The revival of epic, and the secularization of the sacred hero, occurred in the extant third book of a lost and larger Virgilian epic, anonymously transmitted but known by the title *Karolus Magnus et Leo Papa* ("Charlemagne and Pope Leo"). Its example was followed in the next generation by Ermoldus Nigellus, writing about the deeds of Louis the Pious, and the tradition of earlier Carolingian authors is extended by two major political poets, Walafrid Strabo and Sedulius Scottus (also the author of an uproarious mock epyllion). In prose the major achievements lie in the fields of biography, with Einhard's *Vita Karoli Magni* (*c.* 830; *Life of Charlemagne*); of religious controversy, with Theodulf's *Libri Carolini* (defenses written at Charlemagne's request); and of theology, with John Duns Scotus' metaphysical masterpiece, the *Periphyseon.*

THE 9TH TO THE 11TH CENTURY

The sequence and the liturgical drama

From the later 9th century on, the liturgy gave rise to two new literary forms: the sequence and the liturgical drama. Notker Balbulus, monk of St. Gall, was not the first to compose sequences, but his *Liber hymnorum* ("Book of Hymns"), begun about 860, is an integrated collection of texts that spans the whole of the church year in an ordered cycle. Performed between the biblical readings in the mass, each sequence is a free meditation upon scriptural themes, often drawing upon and synthesizing disparate texts. Among later exponents of the genre, Adam of St. Victor was the most distinguished, though the mystical sequences of Hildegard of Bingen exercise a potent appeal. During the same period the enormous expansion of the cult of the Virgin left a notable mark upon hymnody, the early 11th century seeing the composition of Marian hymns, including such ubiquitous texts as "Salve Regina" ("Hail, Queen") and "Alma Redemptoris Mater" ("Sweet Mother of the Redeemer").

Notker's sequences are alive with dramatic possibility, and at St. Gall the practice of troping, or embellishing, liturgical texts also took dramatic form. The *Quem quaeritis* trope from St. Martial, an abbey at Limoges, was one of the earliest such pieces to demand dramatic performance. From this beginning developed the long tradition of liturgical drama, which, like the sequence, is centred upon the major feasts of the church year.

Two narrative works stand out in this period. The *Waltharius* epic is set in the years of the invasions of Attila the Hun. The polished sophistication of its narrative technique contrasts delightfully with its Germanic subject matter. The *Ruodlieb*, a romance written perhaps in about 1050 in a language heavily influenced by vernacular usage, reveals a comparable narrative subtlety. Even in its fragmentary state, the variety and vigour of its episodes are apparent.

The ease with which religious forms such as the sequence are adapted for secular use is nowhere seen better than in the 11th-century compilation known as the *Cambridge Songs.* The blend of humorous contes, hymnody, and lyric testifies to a diverse taste in the unknown anthologist. Other lyric collections from the next century, such as the Ripoll and Arundel lyrics, may draw upon work of earlier provenance. To the chance survival of individual compilations such as these derives the bulk of knowledge of the secular lyric, which is one of the chief distinctions of the 12th and 13th centuries.

THE 12TH TO THE 14TH CENTURY

The *Carmina Burana*

The *Carmina Burana* ("Songs from Bavaria"), the largest and greatest collection of secular lyrics, comes from the Benediktbeuern, a Benedictine monastery in Bavaria. It was put together in the 13th century, though most of the songs are much older, and contains work by many of the finest poets of the age. The contents are divided by subject into moral and satirical verse, love poetry, drinking songs, and liturgical dramas. Walter of Châtillon and Philip the Chancellor are conspicuous among the authors of the satires, the force of their works deriving from learned and allusive use of Scripture. Peter of Blois is found in the section of satirical verse and the section of love poetry. His verse forms achieve a new degree of delicacy and sophistication, and his erotic poetry owes much to a close study of classical poets, particularly Ovid. Yet many of the forms in evidence, the pastourelle (a love debate between a knight and a shepherdess) for example, have no classical antecedent. In the complexity of its argument, and profusion of imagery, a poem such as "Dum Diane vitrea" ("While Shining Diane") far exceeds the imagination of any classical author. Among the drinking songs in the third section are works of the anonymous German "Archpoet" and of Hugh Primas of Orléans, a slightly earlier figure. Under the cover of a pointedly low-life persona, these poets, both prominent men in court society, practiced a robust form of satire in which much of the humour is deflected upon themselves. Grander forms of poetry are not neglected: Walter of Châtillon's foray into epic, the *Alexandreis* (written *c.* 1180), is one of the most distinguished products of the medieval fascination with the legends of Alexander the Great, and it exercised an immense influence on subsequent vernacular literature.

Works of Peter Abelard

The 12th century was an age of philosophical development, above all in the cathedral schools (as at Chartres) and new universities (as at Paris). Scholars such as Alain of Lille (Alanus de Insulis) and John of Salisbury returned to philosophical problems that had been posed in the days of Boethius. With Roger Bacon, Duns Scotus, and Robert Grosseteste, the first chancellor of Oxford University, a significant English contribution is discernible. Peter Abelard trained at Paris, where he taught John of Salisbury. Of Abelard's philosophical works, *Sic et non* (completed *c.* 1136; "Yes and No") is the most notable, probing critically the vast bulk of received authority. In three of his most original literary works, the relationship with Héloïse is a prominent feature. The *Hymnarius Paraclitensis* is a collection of hymns for Héloïse's convent, where the reading of Scripture is complex and shows the imprint of novel theological thought. The six *planctus* ("laments") are meditations on guilt and suffering, set in the mouths of biblical personages, while the correspondence between Abelard and Héloïse reflects themes found in both verse collections. Abelard's autobiographical work, the *Historia calamitatum* (written *c.* 1136; *The Story of Abelard's Adversities*), recounts the story of his tragic love affair and its theological consequences.

Liturgical and cultic innovation left its mark upon Latin literature during the 13th and 14th centuries. John of Garland's compilation of hymns to the Virgin is a late testimony to the force of Marian inspiration. From the early 13th century derive two of the latest sequences to feature in the liturgy in all countries, the "Dies irae" ("The Day of Wrath") and the "Stabat Mater" ("The Mother Stands"). The cults of the Holy Cross and of the Passion are the impetus to the poetry of two Franciscans, the Italian St. Bonaventura and John Pecham in England. Pecham's *Philomena praevia* is an extended lyrical meditation that blends the story of the Redemption with the liturgical course of a single day.

The theology of the 13th century is dominated in bulk and stature by the writings of St. Thomas Aquinas. The culmination of a career centred upon Paris and Rome is the *Summa theologiae* (written between 1265 and 1272), a systematic exposition of the essentials of faith, grounded in Aristotelian principles. The translation of Aristotle into Latin continued throughout the century. Aquinas' liturgical works also remained prevalent. (P.Go.)

Renaissance Latin literature

The term Renaissance Latin is associated, for 14th-century Italy, mainly with Dante, Petrarch, and Boccaccio, though mention should also be made of the Florentine historian Leonardo Bruni and the humanist scholars Albertino Mussato, Coluccio Salutati, and Aeneas Silvius Piccolomini (Pope Pius II). In verse there was a general return to classical models and elegance, while in prose Latin was still a necessary medium for the abundant humanistic, scientific, philosophical, and religious literature that was a mark of the new age.

In Italy there were three main centres of learning and literature in the 15th and 16th centuries: Florence, Rome, and Naples. Each of these centres had its own circle of writers and scholars. The Florentine group was chiefly noted for the Platonist philosophers Poggio Bracciolini, Marsilio Ficino, Giovanni Pico della Mirandola, and a poet and scholar, Angelo Poliziano. Rome was the centre for a grammarian, Pietro Bembo, and for Marco Vida, author of a Latin epic on the redemption, while Naples was the home of many poets and scholars, notably Giovanni Pontano, Jacopo Sannazzaro, Lorenzo Valla, and Girolamo Fracastoro.

Influence of the Brethren of the Common Life

Germany and the Low Countries also made a large contribution in prose and verse to Latin literature in the 15th and 16th centuries. Many humanists owed their early education to the Brethren of the Common Life, a Dutch Christian community that laid great emphasis on the classics. Among these was Desiderius Erasmus, the greatest figure of the northern Renaissance. Bred in the rhetorical tradition of literary humanism, he had little interest in the scientific premonitions of the age. As an editor and expositor of classical texts and the writings of the Church Fathers, as a commentator on the ecclesiastical conflicts of his time, and as a scholar, wit, and satirist, he was unsurpassed by any humanist in northern Europe. A German abbot, Johannes Trithemius, was a historian and scholar with an immense range of interests and knowledge; Conradus Celtis was conspicuous as a humanist and poet; while Petrus Lotichius wrote elegant verse.

Spanish humanism was best seen in the scholar and friend of Erasmus, Juan Vives, while in England the statesman and scholar Sir Thomas More was the outstanding figure. Polydore Virgil, an Italian, brought the new methods of historical writing into England, though a poet and historian, Tito Livio Frulovisi, had written a life of Henry V that influenced later English writers. Among many Latin poets should be mentioned George Buchanan and John Barclay, both Scots. The strong English tradition of classical verse composition in the schools was shown in the Latin poems of such 17th-century poets as John Milton, Henry Vaughan, Richard Crashaw, and Abraham Cowley.

In France, where, as in England, the Renaissance came late, some members of the group of writers known as La Pléiade wrote Latin verse. Despite the eventual triumph of the French vernacular, Latin poems continued to be written, and several hymns composed in classical forms were included in church services in the 17th and 18th centuries. (F.J.E.R.)

Until the early 18th century, Latin was recognized as the best medium for historical and scientific work if it were intended to reach a European audience. For this reason Marsilio Ficino and Pico della Mirandola, Erasmus and More, and later Francis Bacon, Hugo Grotius, René Descartes, Benedict Spinoza, and Sir Isaac Newton used what was still an international language.

BIBLIOGRAPHY

Ancient: Detailed and documented accounts include W.S. TEUFFEL, *Teuffel's History of Roman Literature,* new ed., 2 vol. (1891–92, reprinted 1967; originally published in German, 5th rev. ed. 1890); MARTIN SCHANZ, *Geschichte der römischen Literatur bis zum Gesetzgebungswerk des Kaisers Justinian,* rev. by CARL HOSIUS and G. KRUGER, 4 vol. in 5 (1914–35, reissued 1966–71); AUGUSTO ROSTAGNI, *Storia della letterature latina,* 3rd ed., 3 vol. (1964), sumptuously illustrated; EDUARD NORDEN, *Die römische Literature,* 6th ed. (1961); H.J. ROSE, *A Handbook of Latin Literature from the Earliest Times to the Death of St. Augustine,* 3rd ed. (1954, reprinted with supp. bibliog., 1966); and E.J. KENNEY and W.V. CLAUSEN (eds.), *The Cambridge History of Classical Literature,* vol. 2, *Latin Literature* (1982). J. WIGHT DUFF, *A Literary History of Rome: From the Origins to the Close of the Golden Age,* 3rd ed., ed. by A.M. DUFF (1953, reprinted 1967), and *A Literary History of Rome in the Silver Age: From Tiberius to Hadrian,* 3rd ed., ed. by A.M. DUFF (1964), is a standard introduction containing comprehensive and scholarly surveys, with supplementary bibliographies. Shorter accounts are J.W. MACKAIL, *Latin Literature,* 2nd rev. ed. (1896, reissued 1966); MICHAEL GRANT, *Roman Literature,* new ed. (1958); and KARL BUCHNER, *Römische Literaturgeschichte,* 5th ed. (1980). On poets, see W.Y. SELLAR, *The Roman Poets of the Republic,* 3rd ed. (1889, reprinted 1965), *The Roman Poets of the Augustan Age: Virgil,* 3rd ed. (1897, reprinted 1965), and *The Roman Poets of the Augustan Age: Horace and the Elegiac Poets,* 2nd ed. (1899, reprinted 1965); H.E. BUTLER, *Post-Augustan Poetry from Seneca to Juvenal* (1909, reprinted 1977); and J. WIGHT

DUFF, *Roman Satire* (1936, reissued 1964). General works include H. BARDON, *La Littérature latine inconnue* (1952); F. KLINGNER, *Römische Geisteswelt,* 5th ed. (1965, reprinted 1979); M.L. CLARKE, *Rhetoric at Rome* (1953, reprinted 1968), and *The Roman Mind* (1956, reissued 1968); S.F. BONNER, *Roman Declamation in the Late Republic and Early Empire* (1949, reprinted 1969); W. BEARE, *The Roman Stage,* 3rd rev. ed. (1964, reprinted 1977), and *Latin Verse and European Song* (1957); and GORDON WILLIAMS, *Tradition and Originality in Roman Poetry* (1968). For an account of Latin studies see M. PLATNAUER (ed.), *Fifty Years (and Twelve) of Classical Scholarship,* 2nd rev. ed. (1968); and the annual bibliography in *L'Année philologique,* published by the International Society of Classical Bibliography in Paris. Posthumous influence of various authors is traced in G. HIGHET, *The Classical Tradition* (1949, reprinted 1957); and R.R. BOLGAR, *The Classical Heritage and Its Beneficiaries* (1954, reprinted 1977).

Middle Ages: ADOLF EBERT, *Allgemeine Geschichte der Literatur des Mittelalters in Abendland,* 2nd ed., 3 vol. (1880–89, reprinted 1971); GUSTAVO VINAY, *Alto midioevo latino: conversazioni e no* (1978); BERNHARD BISCHOFF, *Mittelalterliche Studien,* 2 vol. (1966–67); CHRISTINE MOHRMANN, *Études sur le latin des chrétiens,* 3 vol. (1958–61); K. STRECKER, *Introduction to Medieval Latin* (1957, reissued 1968; originally published in German, 2nd ed., 1929); GIOVANNI CREMASCHI, *Guida allo studio del latino medievale* (1959); F.A. WRIGHT and T.A. SINCLAIR, *A History of Later Latin Literature from the Middle of the Fourth to the End of the Seventeenth Century* (1931, reprinted 1969); M. HÉLIN, *A History of Medieval Latin Literature,* rev. ed. (1949, originally published in French, 1943); PIERRE DE LABRIOLLE, *Histoire de la littérature latine chrétienne,* 2 vol., 3rd ed. (1947); M.L.W. LAISTNER, *Thought and Letters in Western Europe, A.D. 500 to 900,* new ed. (1957); MAX MANITIUS, *Geschichte der lateinischen Literatur des Mittelalters,* 3 vol. (1923–31, reprinted 1965–74); J. DE GHELLINCK, *Littérature latine au moyen age,* 2 vol. (1939), *L'Essor de la littérature latine au XII^e,* 2 vol. (1946); C.H. HASKINS, *The Renaissance of the Twelfth Century* (1927, reissued 1971); and ERNST ROBERT CURTIUS, *European Literature and the Latin Middle Ages* (1953, reprinted 1983; originally published in German, 1948). On the poetry of the period, see FREDERIC J.E. RABY, *A History of Christian-Latin Poetry from the Beginnings to the Close of the Middle Ages,* 2nd ed. (1953, reprinted 1966), and *A History of Secular Latin Poetry in the Middle Ages,* 2 vol., 2nd ed. (1957, reprinted 1967); MAX MANITIUS, *Geschichte der christlich-lateinischen Poesie* (1891); DAG L. NORBERG, *La Poésie latine rhythmique du Haut Moyen Age* (1954); JACQUES FONTAINE, *Naissance de la poésie dans l'occident Chrétien* (1981); and PETER GODMAN (ed.), *Poetry of the Carolingian Renaissance* (1985).

Renaissance: P. VAN TIEGHEM, *La Littérature latine de la Renaissance* (1944, reprinted 1966); WILFRED P. MUSTARD (ed.), *Studies in the Renaissance Pastoral,* 6 vol. (1911–31); GEORG ELLINGER, *Geschichte der neulateinischen Literatur Deutschlands im sechzehnten Jahrhundert,* 3 vol. (1929–33, reprinted 1969); WOLFGANG MANN, *Lateinische Dichtung in England vom Ausgang der Frühhumanismus bis zum Regierungsantritt Elisabeths* (1939); JOHN SPARROW, "Latin Verse of the High Renaissance," in E.F. JACOB (ed.), *Italian Renaissance Studies* (1960); ALESSANDRO PEROSA and JOHN SPARROW (eds. and comps.), *Renaissance Latin Verse: An Anthology* (1979); ROBERTO WEISS, *The Dawn of Humanism in Italy* (1947, reprinted 1970), *The Spread of Italian Humanism* (1964), and *Humanism in England During the Fifteenth Century,* 3rd ed. (1967); and HANS BARON, *The Crisis of the Early Italian Renaissance* (1966).

(Ed.)

Lavoisier

A French chemist and the father of modern chemistry, Antoine-Laurent Lavoisier was a brilliant experimenter and many-sided genius who was active in public affairs as well as in science. He developed a new theory of combustion that led to the overthrow of the phlogistic doctrine, which had dominated the course of chemistry for more than a century. His fundamental studies on oxidation demonstrated the role of oxygen in chemical processes and showed quantitatively the similarity between oxidation and respiration. He formulated the principle of the conservation of matter in chemical reactions. He clarified the distinction between elements and compounds and was instrumental in devising the modern system of chemical nomenclature. Lavoisier was one of the first scientific workers to introduce quantitative procedures into chemical investigations. His experimental ingenuity, exact methods, and cogent reasoning, no less than his discoveries, revolutionized chemistry. His name is indissolubly linked to the establishment of the foundations upon which modern science rests.

Education and early work

Lavoisier was born in Paris on August 26, 1743. His father, an *avocat au parlement* (parliamentary counsel), gave him an excellent education at the Collège Mazarin, where, along with a solid classical grounding in language, literature, and philosophy, he received the best available training in the sciences, including mathematics, astronomy, chemistry, and botany. Following his family's tradition, he pursued the study of law, and he received his license to practice in 1764. His inquiring mind, however, continually drew him to science. In 1766 he received a gold medal from the Academy of Sciences for an essay on the best means of lighting a large town. Among his early work were papers on the Aurora Borealis, on thunder, and on the composition of gypsum. Pursuing an early interest in rocks and minerals, he accompanied the geologist J.-E. Guettard on a long geological trip and assisted him in preparing his mineralogical atlas of France. In 1768, after presenting a paper on the analysis of water samples, Lavoisier was admitted to the academy as *adjoint-chimiste* (associate chemist). He passed through all the grades in the academic structure and was made director in 1785 and treasurer in 1791.

Lavoisier, with his wife, oil painting by Jacques-Louis David, 1788. In the collection of Rockefeller University, New York City.

By courtesy of The Rockefeller University, New York City

Through his family, Lavoisier became independently wealthy in his early 20s. In 1771 he married Marie Paulze, who would later assist him in his work by illustrating his experiments, recording results, and translating scientific articles from English. In accordance with a common practice among the wealthy bourgeoisie at the time, his father bought him a title of nobility in 1772, and a few years later Lavoisier purchased the country estate of Fréchines, near Blois.

Scientific achievements. Lavoisier's name gained wide recognition when, in 1770, he refuted the then prevalent belief that water is converted into earth by repeated distillation. By carefully weighing both the earthy residue and the distilling apparatus, he demonstrated that the solid matter came from the glass vessels and not from the water.

Study of combustion

Speculating on the nature of the traditional four elements—earth, water, air, and fire—Lavoisier began to investigate the role of air in combustion. On November 1, 1772, he deposited with the Academy of Sciences a note stating that sulfur and phosphorus when burned increased in weight because they absorbed "air," while the metallic lead formed when litharge was heated with charcoal weighed less than the original litharge because it had lost "air." The exact nature of the airs concerned in the processes he could not yet explain, and he proceeded to study the question extensively. In 1774 he published his first book, *Opuscules physiques et chimiques,* in which he presented the results of both his reading and his experimentation. That year Joseph Priestley prepared "dephlogisticated air" (oxygen) by heating "red precipitate of mercury." Lavoisier confirmed and extended Priestley's work. Perceiving that in combustion and the calcination of metals only a portion of a given volume of common air was used up, he concluded that the active agent was Priestley's new "air," which was absorbed by burning, and that "nonvital air," or azote (nitrogen), remained behind. He observed that birds lived longer in the new "eminently respirable air," as he described it, and he showed that this air combined with carbon to produce the "fixed air" (carbon dioxide) obtained by Joseph Black in 1754.

Recognition that the atmosphere is composed of different gases that take part in chemical reactions made it possible to identify the composition of many substances, particularly the acids. In a memoir presented to the academy in 1777, read in 1779 but not published until 1781, Lavoisier assigned to dephlogisticated air the name oxygen, or "acid producer," on the erroneous supposition that all acids were formed by its union with a simple, usually nonmetallic body. He explained combustion not as the result of the liberation of a hypothetical fire principle, phlogiston, but as the result of the combination of the burning substance with oxygen. On June 25, 1783, he announced to the academy that water was the product formed by the combination of hydrogen and oxygen; in this, however, he had been anticipated by the English chemist Henry Cavendish. As a member of a committee for finding ways to improve lighter-than-air flight with the newly invented balloons, he produced quantities of hydrogen, called "inflammable air," by decomposing water into its constituent gases. From his knowledge of the composition of water, Lavoisier was led to the beginnings of quantitative organic analysis. He burned alcohol and other combustible organic compounds in oxygen, and from the weight of water and carbon dioxide produced he calculated their composition.

Lavoisier published a brilliant attack on the phlogistic theory in 1786. Despite the opposition of Priestley and others, a growing number of scientists began to adopt his views. In 1787 a group of French chemists published the *Méthode de nomenclature chimique,* which classified and renamed the known elements and compounds. Reflecting Lavoisier's new discoveries and theories, the *Nomenclature* exerted a wide influence. Also influential was the revision in 1788 of Antoine-François de Fourcroy's popular *Élémens d'histoire naturelle et de chimie,* which was completely recast in terms of Lavoisier's views and according to the new chemical nomenclature. The following year Lavoisier and others established the *Annales de chimie,* a journal devoted to the new chemistry. Gradually the older approach based on the phlogistic theory lost adherents, and eventually Lavoisier's ideas were adopted universally.

The spread of Lavoisier's doctrines was greatly facilitated by the defined and logical form in which he presented them in his *Traité élémentaire de chimie* (1789). This classic book provided a concise exposition of his work and that of his followers and offered an introduction to the new approach to chemistry. In the prefatory "Discours préliminaire" Lavoisier set forth his views on the proper methods of scientific inquiry and scientific teaching, and he defended the new nomenclature. Those substances that could not be decomposed he termed *substances simples,* the elements out of which other matter was made. To a large extent the modern concept of an element, as against the ancient Greek idea, stems from Lavoisier. In the *Traité* he furnished a clear statement of his principle of the conservation of matter in chemical reactions. Nothing, he said, is created or destroyed; there are only alterations and modifications, and there is an equal quantity—an equation—of matter before and after the operation.

Other scientific work

In addition to his purely chemical work, Lavoisier, mostly in conjunction with the mathematician and astronomer Pierre-Simon Laplace, devoted considerable attention to physical problems, especially those connected with heat. The two carried out some of the earliest thermochemical investigations, devised an apparatus for measuring linear and cubical expansions, and employed a modification of Black's ice calorimeter in a series of determinations of specific heats. Regarding heat (*matière du feu*) as a peculiar kind of imponderable matter, Lavoisier held that the three states of aggregation—solid, liquid, and gas—were modes of matter, each depending on the amount of *matière du feu* with which the substances concerned were associated. He also worked at fermentation, respiration, and animal heat, looking upon the processes concerned as essentially chemical in nature. From measurements made in his pioneering biochemical experiments on animal heat and on the gases exchanged during respiration, he concluded that respiration was a type of oxidation reaction similar to the burning of carbon. A paper discovered many years after his death showed that he had anticipated later thinkers in explaining the cyclical process of animal and vegetable life.

Public service. Throughout Lavoisier's extraordinary career as a scientist, he carried on a simultaneous career as a public servant of remarkable versatility, contributing his talents in the areas of finance, economics, agriculture, education, and social welfare, among others. In 1768 he became an assistant in one of the revenue-collecting departments of the government, subsequently becoming a full titular member of the Ferme Générale, the main tax-collecting agency. The financial and organizational abilities he displayed as a farmer-general, along with his undoubted scientific and technical capacity, led in 1775 to his appointment as *régisseur des poudres* (a director of the gunpowder administration). With his customary energy he set about making improvements in the chaotic powder industry. He abolished the vexatious search for saltpetre in the cellars of private houses, increased the production of the salt, and improved the manufacture of gunpowder. The post enabled Lavoisier to move to the Arsenal of Paris, where he took up residence and equipped a superb laboratory. This establishment soon became a gathering place for the scientists and advanced thinkers of the day, and the dinners presided over by his wife became famous. After dinner the guests often would be escorted to the laboratory to witness or take part in a demonstration of some new experiment. Although an increasing number of public duties claimed Lavoisier's time, he regularly set aside one day a week for scientific investigations.

As his influence in the Academy of Sciences grew, so did his responsibilities. He was a member of numerous official committees to look into matters concerning the public. In 1781 the notorious Franz Anton Mesmer arrived in Paris, and Lavoisier (along with Benjamin Franklin) served on a committee to investigate his cures by "animal magnetism," pronouncing them a hoax. With another committee, he explored the hospitals and prisons of Paris and recommended remedies for their deplorable state. At Fréchines he started a model farm, where he demonstrated the advantages of scientific agriculture. In 1785 he was named to the government's committee on agriculture and as its secretary drew up reports and instructions on the cultivation of crops, promulgating various agricultural schemes. As a landowner in the province of Orléans, Lavoisier was chosen a member of the provincial assembly in 1787. There he devised measures for improving social and economic conditions in the area by such means as savings banks, insurance societies, canals, workhouses, and tax reforms. He advanced money without interest to the towns of Blois and Romorantin for the purchase of barley during the famine of 1788. He was associated with committees on hygiene, coinage, the casting of cannon, and public education. He was secretary and treasurer of the commission appointed in 1790 to secure uniformity of weights and measures throughout France, work that led to the establishment of the metric system.

Lavoisier's fate in the Revolution

A reformer and political liberal opposed to many aspects of the ancien régime, Lavoisier took an active role in the French Revolution. When the States General was reconvened in 1789, he became an alternate deputy and drew up the code of instructions for guidance of the deputies. He was elected to the commune of Paris and joined the moderate Society of 1789, a planning group. He became an administrator of the national treasury and published detailed analyses of the state of the nation's finances and its agriculture. But his membership in the unpopular Ferme Générale was alone sufficient to make him an object of suspicion to the authorities, and, despite his many services to the nation and his wide renown as a scientist, he came under increasingly severe attack from the more radical pamphleteers. In 1787, at Lavoisier's suggestion, a wall had been erected around Paris to halt the flow of contraband into the city. The extremist revolutionary Jean-Paul Marat accused him of putting Paris in prison and of stopping the circulation of air. In 1791 the Ferme Générale was abolished, and Lavoisier was subsequently removed from his position in the gunpowder administration and forced to leave his home and laboratory in the Arsenal. In 1793 the Reign of Terror commenced, and, in spite of strenuous efforts by Lavoisier, the Academy of Sciences, along with the other learned societies, was suppressed. At the end of the year the Revolutionary Convention ordered the arrest of the former members of the Ferme Générale, and in May 1794 they were tried by the revolutionary tribunal. The trial lasted less than a day. Lavoisier and 27 others were condemned to death. That same afternoon, May 8, he and his companions, including his father-in-law, were guillotined at the Place de la Révolution (now Concorde). His body was thrown into a common grave. (D.I.D.)

BIBLIOGRAPHY. ANTOINE-LAURENT LAVOISIER, *Oeuvres de Lavoisier,* 6 vol. (1862–93, reprinted 1965), is a collected edition that includes all of his principal works and memoirs from the academy volumes, as well as numerous letters, notes, and reports. DENIS I. DUVEEN and HERBERT S. KLICKSTEIN, *A Bibliography of the Works of Antoine Laurent Lavoisier, 1743–1794* (1955), with the *Supplement* (1965), is a definitive work that contains a description of Lavoisier's writings, together with annotations. DOUGLAS MCKIE, *Antoine Lavoisier: Scientist, Economist, Social Reformer* (1952, reissued 1962), is an authoritative biography. A good account of his life and career can be found in SARAH R. RIEDMAN, *Antoine Lavoisier: Scientist and Citizen* (1957, reissued 1967). A biography based on 20th-century research is LÉON VELLUZ, *Vie de Lavoisier* (1966). Lavoisier's place in the history of science is explored in KENNETH S. DAVIS, *The Cautionary Scientists: Priestley, Lavoisier, and the Founding of Modern Chemistry* (1966). HENRY GUERLAC, *Lavoisier: The Crucial Year* (1961), examines his first experiments on combustion in 1772, and his *Antoine-Laurent Lavoisier: Chemist and Revolutionary* (1975) is an illustrated reprint of the article from the *Dictionary of Scientific Biography.* ROGER HAHN, *The Anatomy of a Scientific Institution: The Paris Academy of Sciences, 1666–1803* (1971), includes information on Lavoisier's activities.

The Profession and Practice of Law

The primary function of the profession and practice of law is to apply the law in specific cases—to individualize it. This function is manifest in the work of the advocate and the judge in the process of trying and deciding cases. In Anglo-American systems a lawyer investigates the facts and the evidence by conferring with his client, interviewing witnesses, and reviewing documents. He may seek a summary dismissal because the opponent evidently has no case, or, through discovery proceedings he may force the other side to reveal more fully the issues and facts on which it relies. At the trial he introduces evidence, objects to improper evidence from the other side, and advances partisan positions on questions of law and of fact. In continental European countries the judge has greater responsibility for investigation of the facts. At trial he plays an active role in taking evidence, questioning witnesses, and framing the issues. Continental lawyers suggest lines of factual inquiry to the judge and, like their Anglo-American counterparts, advance legal theories and argue the law in accord with the interests of their clients. In either system, if a lawyer loses his client's case, he may seek a new trial or relief in an appellate court.

Even controversies that are not resolved in court require the aid of lawyers. Negotiation, reconciliation, compromise—in all of which lawyers have a large part—bring about the settlement of most cases without trial.

The profession also applies and utilizes the law in the less dramatic setting of the office. The lawyer as counselor and negotiator may aid in shaping a transaction so as to avoid disputes or legal difficulties in the future, or so as to achieve advantages for his client, such as the minimization of taxes. The law gives to private persons extensive but not unlimited power to arrange and determine their legal rights in many matters and in various ways, such as through wills, contracts, leases, or corporate bylaws. In structuring these arrangements the lawyer is helping to particularize the legal rights of the parties.

Another field of legal work, which has developed rapidly in the 20th century, is the representation of clients before administrative commissions, commissions of inquiry, and, in some countries, legislative committees. This development has been a result of the increase of government regulation of economic life.

A lawyer has several loyalties in his work, including loyalty to his client, to the administration of justice, to the community, to his associates in practice, and to himself—whether to his economic interests or to his ethical standards. These diverse and at times competing loyalties must be reconciled with wisdom. It is the purpose of the standards of the profession to effect the reconciliation. (M.A.Gl.)

This article is divided into the following sections:

Legal profession

One definition of the legal profession is "the vocation based on expertness in the law and its application." This simple definition may be best, despite the fact that in some countries there are several professions and even some occupations (*e.g.,* police service) that require this expertness but may not be considered to be within the "legal profession" at all.

HISTORY

Distinct legal systems emerged relatively early in history, but legal professions of size and importance are relatively modern. There is not the slightest trace in ancient times of a distinct legal profession in the modern sense. The earliest known legal specialist was the judge, and he was only a part-time specialist. The chief, prince, or king of small societies discharged the judicial function as part of the general role of political leader. As his power spread, he delegated the function, though not to legal specialists; in the secular stages of the early systems, legal duties were taken over by royal officials who were "generalists." In the wake of powerful religious or quasi-religious movements priests or wise men often judged or advised the judges, a situation that persisted in Muslim countries and in China until the 20th century AD. It may be suspected that in some of these cases specialized legal aid to the ordinary citizen did exist, but at levels of social status below the notice of chroniclers or tomb inscriptions and perhaps without benefit of official approval.

Classical beginnings of a legal profession. A distinct class of legal specialists other than judges first emerged in the Greco-Roman civilization, and as with the law itself, the main contribution was from Rome in the period from 200 BC to AD 600. In the early stages of both Greece and Rome, as later among the German tribes who overran the Roman Empire, there was a prejudice against the idea of specialists in law being generally available for fee. The assumption was that the citizen knew the customary law and would apply it in transactions or in litigation personally with advice from kinsmen. As the law became more complex, men prominent in public life—usually patricians—found it necessary to acquire legal knowledge, and some acquired a reputation as experts. Often they also spent periods serving as magistrates and in Rome as priests of the official religion, having special powers in matters of family law. Among the German tribes noble experts were allowed to assist in litigation, not in a partisan fashion but as interpreters (*Vorsprecher*) for those unready of speech who wished to present a case. The peculiar system of development of the early Roman law, by annual edict and by the extension of trial formulas, gave the Roman patrician legal expert an influential position; he became the jurisconsult, the first nonofficial lawyer to be regarded with social approbation, but he owed this partly to the fact that he did not attempt to act as an advocate at trial—a function left to the separate class of orators—and was prohibited from receiving fees.

The jurisconsult and his successors

The modern legal professional, earning his living by fee-paid legal services, first became clearly visible in the later Roman Empire when the fiction that a jurisconsult received only gifts was abandoned and when at the same time the permissible fees were regulated. Changes in the methods of trial and other legal developments caused the jurisconsult to disappear in time. The orator, who now was required to obtain legal training, became the advocate. A subordinate legal agent of the classical system, the procurator, who attended to the formal aspects of litigation, took on added importance because later imperial legal procedure depended largely on written documents drawn by procurators. The jurisconsults had been important as teachers and writers on law; with their decline this function passed to government-conducted law schools at Rome, Constantinople, and Berytus and to their salaried professors. There was also a humbler class of paid legal documentary experts, the *tabelliones,* useful in nonlitigious transactions.

Medieval Europe. This late Roman pattern of legal organization profoundly influenced the Europe that began to arise after the barbarian invasions from AD 1000 on; and even during the invasions the methods of Roman imperial administration never ceased to exist in some parts of southern France and in central Italy. The Christian Church, which became the official Roman imperial church after AD 381, developed its own canon law, courts, and practitioners and followed the general outline of later Roman legal organization. Because of its success among the invaders the church was in a position to establish its jurisdiction in many matters of family law and inheritance. Hence both the idea of a legal profession and the method of its operation retained sufficient force to offset Germanic and feudal objections to legal representation. After the revival of learning in the 12th century, in particular the revived study of Roman law at Bologna, the influence of the late Roman professional system was greatly strengthened.

Classes of legal practitioners

From then on every country in continental Europe acquired, by various stages and with numerous local variations, a legal profession in which four main constituents could be observed. Procurators attended to the formal and especially the documentary steps in litigation. Advocates, who usually were university graduates in Romanist learning, gave direct advice to clients and to procurators and presented oral arguments in court. Among a miscellany of legal scribes the notaries acquired importance because, in addition to being drafting experts, they also provided officially recognized document authentication and archives. University teachers of law took over the main task of explaining and of adapting the mixture of Roman law and Germanic custom that produced the modern laws of the major European countries and continued to dominate in the scholarly interpretation of the law even after the 19th-century codifications. The relative importance of these classes varied enormously from place to place and from century to century. At times the teaching doctors almost supplanted the advocates; in some courts the procurators swallowed up the advocates and in others the converse occurred; only the notaries managed to survive with little change.

England after the Conquest. England after the Norman Conquest also was influenced by Roman example, and the clerics who staffed the Norman and Plantagenet monarchies and who provided the earliest of their judges enabled the notion of a legal profession and especially of litigious representation to be accepted. Only in the ecclesiastical and admiralty courts, however, did procurators (proctors) and doctors of the civil and canon laws become established as practitioners. The native "common law" was developed by a specialized legal society, the Inns of Court in London; there, through lectures and apprentice training, men acquired admission to practice before the royal courts. More particularly, they could become serjeants—the most dignified of the advocates, from whom alone after about 1300 the royal judges were appointed. Various agents for litigation resembling procurators also became known. The "attorneys," authorized by legislation, at first shared the life of the Inns with the "apprentices" in advocacy, who themselves in time acquired the title of barristers. Indeed there were cases of men working as both barristers and attorneys. When in the 16th century the Court of Chancery was established as the dispenser of "equity," the appropriate agent for litigation was called a solicitor, but the common-law serjeants and barristers secured the right of advocacy in that court. It was not until the 17th century that the attorneys and solicitors were expelled from the Inns and the division between advocate and attorney became rigid, and not until the 18th century that the barristers accepted a rule that they would function only on the engagement of an attorney—not directly for the client. Other types of legal agents also developed in England, but in the 19th century all of the nonbarristers were brought under the one name, solicitor. The order of serjeants was wound up, leaving only barristers, of whom the most senior could be made Queen's (or King's) Counsel.

The English division

In its final development the English legal profession thus bore a resemblance to the European—particularly to that of northern France, where the *parlements* (courts) had a corporate life and apprentice training not unlike that of the Inns. But there were four significant differences between England and the Continent. No distinct class of university teachers and commentators on the national law developed in England. Development of the law took place chiefly through precedent based on the reported judgments of the courts, rather than through legislation. The continental monarchies also developed a system of career judicial office, in which the young university licentiate went straight into government service, whereas in England appointment of judges from the senior practicing profession remained the settled practice. In addition, the division between barristers and solicitors ultimately became much more rigid in England than did the division between the advocate and procurator in Europe, and Europe never adopted an equivalent of the English practice requiring a barrister to be employed by a solicitor; both the procurator and the advocate were separately and directly employed by the client. England never developed the profession of notary, so that the whole burden of transactional work fell on those who are now the solicitors, with legal advice from the bar.

Worldwide legal profession. The main patterns both of law and of legal practice were exported by the continental European powers and England to their overseas colonies and possessions, and most of the noncolonial countries of the rest of the world imitated one or the other system. Thus the Romano-Germanic practices (frequently called civil law) became the norm for Scandinavia, Scotland, Latin America, and most of the Muslim countries of the Middle East, for French-speaking areas and Portuguese and Spanish Africa, and for Japan, Thailand, and the former French parts of Southeast Asia. They have also influenced practice in what are now the socialist countries of eastern Europe. The English system provided the model for English-speaking North America, for most former English colonies in Africa, including South Africa, for most of the Indian subcontinent, and for Malaysia, Australia, and New Zealand. The original model has undergone considerable modification by both the countries of export and the countries of reception. In particular, the specialization of procurator–advocate and solicitor–barrister has tended to be replaced by a "fused" profession of legal practitioners qualified to perform both functions and usually doing so. Such a fusion occurred gradually in Germany between the 16th and 18th centuries. It has taken place more recently in France except before the courts of appeal and, while the division still formally exists in Italy, it is no longer of practical importance. In Latin America the fused profession is general. Notaries as a separate specialized branch of the profession exist, however, in most civil-law countries.

CHARACTERISTICS OF THE PROFESSION

Social role. The legal profession has always had an ambiguous social position. Leading lawyers have usually been socially prominent and respected—the sections of the profession so favoured varying with the general structure of the law in the particular community. The family status of early Roman jurisconsults may have been more important than their legal expertise in securing such a position, but by the time of the principate it was their legal eminence that made them respected. The English serjeants lived magnificently, especially in Elizabethan times, and the French Ordre des Avocats was established (14th century AD) by feudal aristocrats in circumstances reminiscent of early Rome—including an insistence on receiving gifts rather than fees. The early Italian doctors of the civil and canon law (12th–15th centuries) were revered throughout Europe. In England and the countries influenced by its system the highest prestige gradually came to be concentrated on the judges rather than on the order of serjeants, of which they were members, and the judges of high-level courts remain the only legal class in the liberal capitalist common-law countries of today to command great respect. In the Romano-Germanic systems it is the notaries and the advocates who have come to be most trusted or admired, the judiciary being more closely identified with the civil service.

The legal profession and social conflict

Yet along with this high repute, sustained over two millennia, lawyers have also been among the most hated and distrusted elements in society. In a few cases this has been the consequence of a general hostility to the whole idea of law, China being the most important example. Confucian teaching (6th century BC) opposed the use of civil law as a major means of social control, and this influence remained powerful there and in Japan until the 20th century. In the Soviet Union the early leaders (1917–22) imagined that law and lawyers were the instruments of the ruling classes and that law would soon wither away in classless Communism. This belief was revived in the early days of Chinese Communism (1947–55). Further experience persuaded these governments that there was room for "socialist legality" and for lawyers to serve it, but a degree of mistrust remains and the repute of the legal expert is lower than that of the political and technological expert.

Most lawyers are conservative because the law itself is predominantly intended to satisfy expectations arising from an inherited pattern of behaviour; in a particular social setting this tends to identify the lawyer with the established and successful classes and to make him seem an enemy to oppressed classes or "new men." Individual lawyers have, nevertheless, often been on the side of rebels: Robespierre and Lenin were both lawyers. But the dominant attitude of the legal profession is one of moderation. Thus many lawyers took the British side in the American Revolution, and even among the lawyers who took the other side the predominant influence was against any attempt to turn the political revolution into a social revolution.

Along with these ideological and political reasons for popular distrust, and even more deep-seated, are the inherent difficulties associated with law and with some of the legal functions. Most people would like law to be so certain that its application is of equal certainty in all cases and so simple that any person of sense can see how it applies. In a discipline sharing the imperfection and complexity of society itself, no such situation is attainable, and the lawyers are blamed for the basic difficulty of their craft—which in some instances they intensify needlessly themselves by multiplying obscurities, contradictions, and complexities. The legal function likely to be most distrusted by the average person, though it also produces some of the law's heroes, is litigious advocacy, particularly in the criminal law. Plato and Aristotle condemned the advocate as one who was paid to make the worse cause appear the better or endeavoured by sophisticated tricks of argument to establish as true what any person of common sense could see was false. The feeling against advocacy in the criminal law was so strong that, at least in the case of the more serious kinds of crime, a right to representation by a trained advocate was nowhere generally recognized until the 18th century AD.

Governments and the members of organized legal professions have from the infancy of the craft endeavoured to meet the basic problem of representation by a basic rule of professional ethics—that the dominant duty of the advocate is not to his client but to truth and the law. Since the later Roman Empire, advocates have been required to take oaths to this effect, and lawyers are often technically classed as "officers of court." The duty of the advocate is to fight for the rights of his client, but only up to the point where an honourable person could fairly put the case on his own behalf. He must not identify with his client's possible willingness to tell untruths or to misrepresent the relevant law. (See below *Legal ethics.*)

Private practice. Client-directed lawyers often are called counselors, but in the original sense of that word—giving advice as to how the law stands—this is rarely an independent function; it is an inseparable part of the other functions. In his client-directed activities the lawyer is concerned with how the law affects specific circumstances, which can for convenience be divided into two main types: transactional and litigious.

In the transactional type the lawyer is concerned with the validity or legal efficacy of a transaction independent of any immediate concern with the outcome of litigation. Such activities comprise the largest area of professional activity whether considered from the point of view of the number of lawyers involved, or of the time they have to spend on the task, or of the number of clients affected. If the events constituting the transaction in question happen before the lawyer is consulted, he can only advise on their legal significance and perhaps suggest methods of overcoming legal deficiencies in what has been done. If future conduct is involved, he is better placed to plan his client's course of conduct so as to secure the required end in the most economical fashion that the law permits and to minimize the chances of future litigation. Transactions may concern words and acts, but characteristically they require the drafting of documents. In the Romano-Germanic systems these often require notarization. Typical activities falling in this category today include the following: transferring interests in land; transmitting property on death; settling property within a family; making an agreement (especially if a commercial agreement of some complexity and duration is involved); incorporating or winding up a corporate entity; varying the terms on which a corporate entity is conducted (classes of shares, managerial rights, distribution of profits, etc.); and adjusting the ownership and control of property and income so as to comply with the requirements of taxation laws and minimize their impact on the property and income in question or so as to ensure the proper management of the assets and distribution of the proceeds among beneficiaries (estate planning) or both. In the Romano-Germanic systems many of these functions are discharged by notaries and in the English and similar divided systems by solicitors, though in difficult situations the opinions of advocates or barristers may be obtained. In the fused professions of North America some firms of attorneys, or departments within firms, specialize in business of this type and avoid, so far as they can, the litigious function.

Typical legal transactions

The litigious function is subdivided into three main stages. First is the preparation of the case—interviewing the client and investigating the circumstances in the light of the leads provided by the client, and attending to the formal requirements of the procedure in question, which may involve writs, summonses, filing of statements of claim or defense, and preparing for trial. Second is the trial proper, in which the facts and law are established and argued before the judge and a decision is made. Third is the execution of the judgment—obtaining payment of damages, delivery of property, or performance of obligation in civil cases; payment of fine or imprisonment, etc., in criminal matters. Similar stages arise on appeal. In the divided professions the sharing of these functions is intricate and varies between one system and another. The advocate or barrister is especially responsible for the second stage, but he may advise upon or draft many of the documents used in other stages. If incidental disputes concerning procedure have to be litigated, he is likely to conduct the proceedings; and if the procedure includes a pretrial conference, he is likely to represent the client. Otherwise the first and third stages are mainly the province of procurator or solicitor.

Public-directed practice. Many law graduates choose to enter public service rather than private practice. Of the public roles played by members of the legal profession that of judge is most visible, but the status of judge and the mode of entry into this branch of the profession vary considerably from country to country.

The traditional independence, power, creativity, and prestige of the Anglo-American judge contrasts with the rather ordinary civil servant status of most continental judges. In the countries of Anglo-American influence, at least until recently, appointment (or, in some U.S. states, election) to a judgeship has been viewed as the crown of a long and distinguished legal career. In the continental countries, by contrast, a law graduate who wishes to be a judge has merely to complete a training period and pass an examination to get a job deciding cases. The beginning civil-law judge can expect to start at the lowest level and, like any other civil servant, to rise in the hierarchy through a series of promotions. In continental Europe ordinarily only positions on the highest courts are open to distinguished practitioners or professors as well as to career civil servants. Lateral entry into the judiciary at any level is uncommon. It has frequently been observed that, because of their standardized training, continental judges tend to share a common outlook and that their concerns about advancement promote a civil service mentality which discourages initiative and independence. Any tendency toward individualism is apt to be further inhibited by the fact that continental judges, even at the lowest levels, usually sit in panels and that their decisions are presented in unsigned opinions. Except in a few courts, such as the West German Federal Constitutional Court, disagreement among the judges is not revealed, either in the form of a dissenting opinion or in a record of the judges' votes.

In the late 20th century, however, the contrast between continental and Anglo-American judicial roles has diminished. In the United States the prestige of judgeships, except at the higher levels, has declined somewhat. It is not unusual for judges to resign and return to private practice or for eminent lawyers to decline to be considered for judicial positions. Relatively low judicial salaries are often mentioned as a reason why lower court judgeships in the United States have ceased to attract the best candidates. Meanwhile, in some continental countries, such as West Germany, judges are being recruited from among the best law graduates. Because of the special training for their positions, continental judges are almost uniformly professional and competent.

Governments have always required a staff of legal specialists, and the scope for such employment today is enormous. There is usually a senior political officer—minister of justice, attorney general, solicitor general—who by convention needs to be a lawyer, and a government department concerned mainly with the legal problems of the government as client (in the English-derived systems usually the office of the attorney general). Increasingly, however, the great departments of state need their own legal subbranch. In some countries, West Germany for example, lawyers dominate the higher offices in the civil service, while in others, such as France, the various official bureaus are more likely to be staffed by nonlawyers who have been trained in a special school of administration. In the socialist countries of eastern Europe most lawyers tend to work for government or for collectivized industrial and farm organizations.

One of the oldest and still most difficult of the governmental legal functions is that of prosecutor in criminal cases. Prosecution is sometimes in part carried on by private persons acting through private lawyers, but the trend is to concentrate the function in government legal officers. In most Commonwealth countries the crown, or public, prosecutor is a specialized officer under the general control of the attorney general. England has an independent "director of public prosecutions" concerned only with the most serious types of crime, but most prosecutions have been conducted by private barristers briefed by him or by the police. A 1985 law, however, provides for the establishment of a body of official prosecutors, following the Scottish system, which relies on public prosecutors (procurators fiscal). In the United States this function has come to be mainly local, and prosecutors, whose most common title is district attorney, are elected for short terms.

In the Romano-Germanic systems prosecuting is a career service. In Italy and France the prosecutor is a member of the judiciary. Both prosecutors and judges receive the same training and both may move from one role to the other in the course of their advancement in the civil service. In West Germany, although the prosecutor is not technically a member of the judiciary, he is not strictly separate from it, and individuals move easily from one position to the other.

The prosecuting function is particularly delicate because criminal prosecution can be used as an instrument of oppression and political persecution, even where conviction is not obtained, and because in most systems prosecutors are expected to act with a degree of fairness and restraint not necessarily expected of the parties to civil litigation. Many Romano-Germanic systems employ officers who keep a general supervision over the working of the courts and especially their criminal jurisdiction. This is the office of the "prosecutor general," or "officer of justice"; and a similar service exists in most of the socialist countries of eastern Europe.

Another branch of government, the legislature, usually requires legal assistance. Legislation needs to be expressed in language readily comprehensible by judges and lawyers and to be framed in harmony with the existing body of law. This requires the service of parliamentary draftsmen who are expert lawyers. A further specialized branch of advisory activity associated with legislation has become prominent—the law reform commission or committee.

Teaching and scholarship. Since Roman times teaching and scholarship in the law have provided prominent roles in the legal profession. Until the 18th century, teaching of the English common law was vested exclusively in the Inns of Court, and a good deal of continental European teaching for professional practice—particularly in the case of notaries and procurators—was also professionally organized. Even university law teaching in Europe was rarely aloof from legal practice; there was usually a fruitful interchange between practitioner and teacher, exemplified in such great figures as the French 18th-century teacher, advocate, and judge Robert Joseph Pothier, whose commentaries provided the foundation for the Napoleonic Code of civil law. Much law teaching in the new university law schools that sprang up in the United States, the United Kingdom, and the Commonwealth in the 19th and 20th centuries was carried on part-time by attorneys, barristers, and judges, and some still is. Sir William Blackstone, the first Vinerian professor of English law at Oxford, came from the bar and became a judge. Law teaching in the late 20th century has tended to become a distinct, full-time profession, usually carried on at a university.

Teachers and practitioners in all countries contribute to an enormous professional literature, with students' texts, practical manuals, theoretical monographs, and a periodical literature whose bulk is coming to be almost as big a problem as the enormous bulk of reported judicial decisions that are consulted for guidance and precedent. Civil-law judges pay close attention to the views of legal scholars as expressed in general and specialized treatises, commentaries on the codes, monographs, law review articles and case notes, and expert opinions rendered in connection with litigation. Persistent scholarly criticism often prompts reexamination of a legal doctrine and sometimes even leads to the abandonment of an established judicial position. In the Anglo-American systems certain kinds of legal writing, such as leading treatises, have become highly influential, as evidenced by the measurable increase in citations to secondary sources in contemporary judicial opinions. But the degree of deference to academic opinion is in general much less than in the continental countries.

Trends in the profession. Beginning in the 1960s nearly everywhere the legal profession experienced unprecedented expansion, which began to slow down in the 1980s. Major elements in this expansion were the postwar baby boom, the opening of new university law faculties, and the entry of women into the profession in significant numbers for the first time. The growth of the profession has increased competitive pressures for jobs and business.

Trends toward specialization and bureaucratization are also clearly noticeable in the modern legal profession. Although formal divisions among lawyers based on functions have declined, there tends to be in most places a de facto division of labour between those lawyers who advise clients and those who appear and argue before tribunals as well as increasing specialization in various legal fields, such as tax law, estate planning, and labour law. Bureaucratization has affected both the practice of law and the judiciary. In all countries the ranks of those lawyers who work as salaried employees of government, business, or law firms are growing more rapidly than those of traditional independent professionals. Changes in the administration of justice in the United States, brought about by increased litigation and crowded dockets, have bureaucratized the roles of judges. Courts in the United States increasingly depend on large and layered staffs to perform many functions that were previously considered to be within the province of the independent judge. While judges in the United States are far from being civil servants in the continental European sense, their work is becoming more administrative in nature. (G.S./M.A.Gl.)

AUTONOMY AND CONTROL

Issues of judicial independence. At least since classical Greece a recurring political theme has been the need for a government of laws rather than of men. Actually, however, as the legal philosopher Julius Stone has said, society of necessity has a government both of laws and of men, and the demand for legal autonomy is often seen in practice as a demand for freedom of the lawyers from political dictation or influence. The main issue has been the independence of the judiciary, and democracies have been particularly assiduous in cultivating both a spirit and traditions that respect judicial independence. The details of their governmental structure or constitutional guarantees tend in that direction, offering obstacles to the ready dismissal of judges, charging their salaries on consolidated revenue, and prohibiting the vesting of judicial functions other than in duly constituted courts of law.

Political context of the judiciary

The special position of the judiciary in constitutional states is usually considered to be an aspect of the division of powers, but it should also be considered in its relation to the structure of the legal profession. Since the late Roman Empire admission to the practice of law and the regulation of the practicing profession have been habitually vested in the judiciary. Furthermore, the duty to speak fearlessly for his client has often required courage of the advocate in the face of political threats directed against him, and, when these threats were directed also against the court before which he appeared, judicial courage was also required. The legal profession as a whole is then seen defending "the rule of law" against the political regime.

The issue of judicial independence may sometimes, however, be seen in the context of tension between the judges and the advocates. In the Romano-Germanic systems the judges often are subject to a strong corporate discipline within their own craft, and differences can occur between them and the body of advocates and also between them and the university teacher-commentators. These differences may relate to questions of legal ethics, especially the limits of advocate identification with client, or to questions of legal doctrine; the judges are then apt to be considered as representing "the state" and the advocates and teachers the autonomy of the law. In the English-derived systems judges are much less subject to corporate discipline, and disputes with the bar are more likely to arise with individual judges and to be highly personal. Even in stable countries, where the rule of law and the independence of judiciary and profession are respected, there is a less dramatic tension between the standards and tone of the lawyers on the one hand and the political administration on the other. For the lawyers, policy is largely concealed in the propositions that constitute the normative system, and legal reasoning usually involves definitions and processes of inference from the body of such propositions themselves rather than directly from the policies that the norms subserve. There have often been revolts against such "logic" within the legal profession itself, especially in the 20th century, but it still remains the most common method of thinking among lawyers and it is doubtful whether one can speak of a "rule of law" at all unless a good deal of legal reasoning is conceptual in style. Politicians and administrators, on the other hand, are more likely to reason directly from policies and purposes and from the considerations relevant to their attainment. This divergence of approach is often illustrated by referring to the tension beween the police officer, confident that he has the guilty man and intent only on putting him in jail, and the lawyers and judge, who insist on the need for "conviction according to law," which may involve applying rules of evidence that seem artificial and even absurd to the police officer. In rigid constitutional systems, where there is judicial review of legislation, politicians may be affronted at the way in which political issues are transformed by the lawyers into legal issues. In many modern countries there has been a tendency to remove certain kinds of disputes both from the courts and from the lawyers and to vest their determination in administrative bodies before which lawyers are denied standing, so as to escape what has been regarded as the blight of legal reasoning; as often there have been reactions in favour of restoring the "rule of law" and the lawyers. In such disputes it is often difficult to distinguish between lawyer attitudes that reflect the necessary features of a rule of law from those that merely reflect the temporary self-interest of particular lawyers or their clients.

Professional associations: protection and control

Regulation by statutes and bar associations. Since about 1800 most countries have brought their legal professions under systems of statutory control with three main principles: admission to practice automatically and compulsorily makes the lawyer a member of an appropriate professional association; those associations are given substantial powers in relation to legal education, admission to practice, and the disciplining of the profession, but they are subject to overriding powers vested in the courts and/or (especially in the Romano-Germanic systems) in government legal departments; the practice of law for reward is prohibited—generally or as to particular functions—to persons not admitted under the system. In the United States about half of the states have such a system, which is known as the "integrated bar"; in the other states bar associations are voluntary and have few controlling powers. England has retained the traditional Inns of Court (in whose management the judges play a leading role) for barristers, but solicitors are subject to a statutory system as above. In some countries (*e.g.,* France) professional organization is regionalized to correspond with judicial organization, and in some federal countries (*e.g.,* the United States, Canada, and Australia) professional control is vested in the states; such situations create the problem of a national organization that is generally a voluntary federation of regional bodies and therefore lacking in compulsive authority. The American Bar Association, established in 1878, is a leading example. In other federal countries (*e.g.,* West Germany and India) the central government has created national law associations responding to the need for a system of control. The law associations, apart from the functions already mentioned, help their members to understand and apply professional ethics, and they develop canons of ethics to cover new problems. They are often active in the prohibition of legal practice by unqualified persons, which tends to bring them into dispute with other professions—*e.g.,* tax accountants and land salesmen—whose members wish to perform legal functions in relation to their tasks and often have considerable knowledge of the relevant law.

Where the profession is divided, it is usually possible to transfer from one branch to another, though sometimes after delay or additional training. In many of the Romano-Germanic systems, however, professional mobility is severely restricted by another factor—numerical limits on the numbers admitted to a branch of the profession.

There are usually limits to the numbers of procurators and notaries, and in some cases, notably the highest French courts, advocate and procurator functions have been combined in relation to a particular jurisdiction and a limit has been placed on numbers; otherwise the number of advocates is generally not restricted. In the restricted cases a person admitted to practice can actually work in the profession only as an employee of an existing practitioner or after buying out such a practitioner.

The actual ability to enter or pursue the profession can also be much influenced by the varying national rules as to legal partnerships. They are prohibited for English barristers and for most divided bars derived from that system and among some of the Romano-Germanic specialized advocates and notaries. In France law partnerships are permitted, and the proportion of lawyers practicing in this manner is constantly increasing. Incorporation of legal practitioners is almost universally prohibited. These restrictions result from the emphasis on personal responsibility of the individual lawyer to his client, to the court, and to the ethical system. In countries with fused professions, however, partnership is usually permitted. West German law firms tend to be quite small, and even in the United States, despite the fact that some very large firms have developed, particularly among the "corporation lawyers" of New York, single-person practices and small partnerships are still common. (G.S./M.A.Gl.)

Legal ethics

Sources of legal ethics

Legal ethics are the principles of conduct that members of the profession are expected to observe in the practice of law. They are an outgrowth of the development of the legal profession itself.

Practitioners of law emerged when legal systems became too complex for those affected by them to understand and apply the law. Certain individuals with the required ability mastered the law and offered their skills for hire. No prescribed qualifications existed, and these specialists were not subject to legal controls. The incompetent, unscrupulous, and dishonest charged exorbitant fees, failed to perform as promised, and engaged in delaying and obstructive tactics in the tribunals before which they appeared. Action to prevent such abuses was taken by legislation and by judicial and other governmental measures. The right to practice law came to be limited to those who met prescribed qualifications. Expulsion from practice and criminal penalties were introduced for various types of misconduct.

These measures did more than correct the abuses. They also gave recognition to the social importance of the functions performed by lawyers and identified those who were qualified to perform them. A consciousness developed within the profession of the need for standards of conduct. This became the core of legal, or professional, ethics.

Prior statutes, court rules, and other government directives remained in force along with the profession's self-imposed ethical standards. Taken together, they constituted the sum total of the restraints placed upon lawyers in regard to their professional conduct. This pattern has continued to the present time.

In many countries professional associations of lawyers have sought to commit the principles of ethical conduct to written form, but a written code is not essential. Ethical principles may exist by common understanding as well as in the literature and writings of the profession. This is the case in England. A code, however, makes ethically obligatory principles readily available to the practitioner and thus helps to assure greater observance of them. When such a code does exist, it usually contains both statements of general ethical principles and particular rules governing specific problems of professional ethics. But no code can foresee every ethical problem that may arise in the practice of law. Hence, codes are supplemented by opinions rendered and published by committees of bar associations.

DIFFERENCES AMONG COUNTRIES

Principles of legal ethics, whether written or unwritten, not only seek to control the conduct of legal practice but also reflect the basic assumptions, premises, and methods of the legal system within which the lawyer operates. They reflect as well the profession's conception of its own role in the administration of justice. In western European legal systems the role of the practitioner in both civil and criminal litigation differs from what it is in Anglo-American legal systems, and this is reflected in the ethics of their legal professions. Practitioners in countries having a Communist form of government are ordinarily salaried employees of the government, and their ethical obligations consequently have a focus different from that in countries where lawyers engage in independent practice or are employed by private firms. In England and to a certain extent in France, where the profession is divided into separate branches, the principles of professional ethics reflect the relationships incident to that division.

In western European and Anglo-American countries and others with similar systems of justice, such as Japan and India, in which the lawyer is not an employee of the state but engages in private practice to serve clients who employ him, professional ethics are addressed to two basic aspects of the lawyer's status. On the one hand, he is employed by clients to serve and represent their interests; on the other, he is participating in an important social function—the application of rules of law through advice, trial of cases, preparation of legal documents, and negotiation with others for his clients. Hence, the principles of legal ethics stress that the lawyer's chief interest lies in serving his client and in securing justice—not in increasing his own income. He is an agent of his client but deemed to retain a large measure of independent judgment as to the proper course to pursue. He represents his client's interests but may not engage in tactics that defeat the fair administration of justice. The lawyer is engaged, it is said, in a profession and not in a business.

Conflicting interests of client and society

Naturally the interests of client and society sometimes conflict, and the principles of legal ethics do not always indicate how these conflicts should be resolved. Should a lawyer cross-examine an adverse witness in such a way as to undermine or destroy his testimony when the lawyer believes the witness is actually telling the truth? May he invoke rules of evidence to exclude points that would weigh against his case but that he considers to be probably true? May he take advantage of the errors of an unskilled opponent? Should he demand a jury trial for purposes of delay when a jury trial has no advantage for his client? These questions may be answered differently in legal systems that operate on different premises. A system in which a lawyer presents his client's case in the most favourable light permitted by law and in which the court must decide the merits of the case may well produce different answers than those produced in a system that assigns a higher priority to the lawyer's duty to the state to assure proper administration of justice.

AREAS OF APPLICATION

Conflicting interests. A lawyer is at times faced with the question of whether or not he may represent two or more clients whose interests conflict. Quite aside from his ethical obligation, the legal systems of the world generally forbid a lawyer from representing a client whose interests conflict with those of another, unless both consent.

In the Anglo-American legal systems the prohibition has three aspects. First, the attorney is not permitted to concurrently represent two or more clients if, in order to further the interests of one, he must forego advancing the conflicting interests of another. In short, he cannot be both for and against a client. Second, he cannot subsequently accept employment from another for the purpose of undoing what he had earlier been retained to accomplish. Third, he may not accept subsequent employment from another if it involves the use, the appearance of use, or possible use of confidential information received from his former client. Such actions are forbidden by law and by legal ethics.

Obligations to the client

To illustrate, an attorney may not ordinarily prepare an instrument for both buyer and seller in which their respective rights are defined. He may not prepare an instrument or negotiate a settlement for a client and later accept employment from another to defeat that instrument or settlement. He may not represent both a driver and his passenger in recovering damages from another party charged with negligent driving in a collision since the passenger may have a claim against his own driver as well. He may not represent two or more defendants in a criminal prosecution if their respective defenses are inconsistent or, possibly, even when the case against one is stronger than the case against the other. The same principles apply with respect to interests of the attorney that may detract from the full and faithful representation of his clients. For example, he may not purchase property that he has been retained to acquire for his client, nor may he draw a will in which he is a beneficiary.

These conflict-of-interest prohibitions are not absolute. The client may consent to the representation after full disclosure of the actual or possible conflict. But the client's consent may not suffice if public interest is deemed to be adversely affected.

The practicing lawyer who is also a member of a legislature is confronted with a conflict of interest whenever his clients enlist his support to promote or oppose legislation or to secure favourable decisions from administrative agencies that are dependent on legislative financial support. The problem is an important one in the United States, where members of legislatures frequently maintain private law practices, but it has received insufficient consideration by the U.S. legal profession. (M.E.P.)

Confidential communications. In the Anglo-American countries judicial decisions, legislation, and professional ethics forbid a lawyer to testify about confidential communications between himself and his client unless the client consents. Similar provisions are found in such diverse legal systems as those of Japan, West Germany, and the Soviet Union. In countries in which the attorney's obligation to protect state interests is given relatively greater emphasis, there may be a duty to disclose information when it is deemed to be to the state's advantage. In Anglo-American law the obligation does not apply when the client seems about to commit a crime. An attorney also may disclose his client's communication when the client sues him; *e.g.,* for malpractice.

Dignity of the profession

Advertising and solicitation. Traditionally, advertising by lawyers was forbidden almost everywhere. It has been a long-standing principle of professional ethics in Anglo-American countries that an attorney must not seek professional employment through advertising or solicitation, direct or indirect. The reasons commonly given have been that seeking employment through these means lowers the tone of the profession, that it leads to extravagant claims by attorneys and to unrealistic expectations on the part of clients, and that it is inconsistent with the personal relationship that should exist between attorney and client. A more basic reason appears to have been the social necessity of restraining the motive of personal gain and of stressing the objective of service. Until 1977 the legal profession in all Anglo-American countries took the position that, with some exceptions, the prohibition must be complete. The situation changed in the United States in 1977 when the U.S. Supreme Court ruled that lawyers could not be barred from advertising fees. The American Bar Association subsequently revised its code of ethics to include provisions and guidelines for advertising and suggested that lawyers limit their advertising to basic information about services and fees. Within narrow limits the same trend has made itself felt in England.

Fees. Attorneys are ethically enjoined to keep fees reasonable, neither too high nor too low. Attempts to control fees range from mandatory fees fixed by statute in West Germany, minutely regulating compensation for legal services of all sorts, to mandatory fees set by courts for solicitors in contentious matters in England and Wales, to advisory fee schedules set by the profession in Canada, France, Spain, and Japan. In the United States local bar associations sometimes enforced minimum fee schedules through disciplinary proceedings, but the U.S. Supreme Court held in 1975 that such practices violated the antitrust laws.

The profession in the United States has assumed, in principle, the obligation to serve poor clients without compensation. The task, however, has become so enormous, especially in view of the expansion of the constitutional right to counsel in criminal cases, that ways of providing paid legal services for the poor have emerged, such as through legal aid societies and public defenders. The growth of legal aid has been a significant 20th-century development in many other countries. In West Germany legal insurance plans are widespread as well.

Fees that are contingent on the successful outcome of litigation or settlement are widely used in the United States, particularly in automobile-accident and other negligence cases, and they are accepted as ethical by the U.S. legal profession. The fee is usually an agreed percentage (typically 20 to 40 percent) of the recovery. The justification given is that this arrangement makes the courts accessible to persons who would otherwise be unable for financial reasons to press their claims. But contingent fees give the attorney a financial stake in the outcome of litigation—which is ordinarily frowned upon. The converse consideration may be that in this type of case, where the outcome is difficult to predict, the lawyer also assumes the risk of losing his fee. Furthermore, although free legal aid has removed the need for a poor person to enter into such a transaction, legal aid is not available to persons who are not poor but are not wealthy enough to engage in extended litigation. In countries other than the United States contingent fees are, nevertheless, generally prohibited. Nor are they permitted in the United States in criminal and divorce cases, in cases to secure a pardon, or in the enactment of legislation.

Honesty in court

Criminal cases. Both the prosecution and the defense of criminal cases raise special ethical issues. The prosecutor represents the state, and the state's concern is not only in convicting the guilty but also in acquitting the innocent. The prosecutor also has an ethical and, in considerable measure, a legal duty to disclose to the defense any information known to him and unknown to the defense that might exonerate the defendant or mitigate the punishment. He must not employ trial tactics that may lead to unfair convictions, nor should he prosecute merely to enhance his political prospects.

The defense counsel has different concerns. Under Anglo-American law an accused may compel the state to prove that he is guilty beyond a reasonable doubt. The defense counsel, therefore, becomes ethically obligated to require the state to produce such proof, whether or not the attorney believes his client to be guilty. His client's guilt is for the tribunal to determine. The attorney may not, however, deliberately resort to perjured or other false testimony. Similar principles hold in civil-law countries. When the client, against the attorney's advice, insists on testifying falsely, the ethical course to be pursued has not been fully settled. Some maintain that the attorney should withdraw, if possible, or else merely permit the client to testify without aiding him or asserting the truth of the testimony given. (M.E.P./M.A.Gl.)

Legal education

The rise of legal education

Schools of law are of comparatively recent origin. The ancient Romans had schools of rhetoric that provided training useful to someone planning a career as an advocate, but there was no systematic study of the law. During the 3rd century BC, Tiberius Coruncanius, the first plebeian *pontifex maximus* (chief of the priestly officials) gave public legal instruction, and a class of nonpriests (*jurisprudentes*) who acted as legal consultants emerged. A student, in addition to reading the few lawbooks that were available, might attach himself to a particular *jurisprudens* and learn the law by attending consultations and by discussing points with his master. Over the ensuing centuries a body of legal literature developed, and some *jurisprudentes* set themselves up as regular law teachers.

In the medieval universities of Europe, including England, it was possible to study canon law and Roman law but not the local or customary legal system. The study of national laws at universities is in most European countries a development that began in the 18th century; the study of Swedish law at Uppsala dates from the early 17th century.

On the continent of Europe the transition to the study of national law was facilitated by the fact that modern legal systems grew mostly from Roman law. In England, on the other hand, the national law, known as the common law, was indigenous. In medieval times education in the common law was provided for legal practitioners by the Inns of Court through reading and practical exercises. These methods fell into a decline in the late 16th century, mainly because students came to rely on printed books, and after the middle of the 17th century there was virtually no organized education in English law until the introduction of apprenticeship for solicitors in 1729. The famous jurist Sir William Blackstone lectured on English law at Oxford in the 1750s, but university teaching of the common law did not develop significantly until the 19th century. The Council of Legal Education for barristers was established in 1852. In the United States systematic legal education began with the founding of the Harvard Law School in 1817. (L.A.S.)

THE AIMS OF LEGAL EDUCATION

Legal education generally has a number of theoretical and practical aims, not all of which are pursued simultaneously. The emphasis placed on various objectives differs from period to period, place to place, and even from one teacher to another. One aim is to make the student familiar with legal concepts and institutions and with characteristic modes of reasoning. Like most intellectual disciplines the law has its technical concepts, frequently expressed in technical terms. All lawyers must become acquainted too with the processes of making law, settling disputes, and regulating the legal profession. They must study the structure of government and the organization of courts of law, including the system of appeals and other adjudicating bodies.

Another aim of legal education is the teaching of law in its social, economic, political, and scientific contexts. While law schools have never ignored the social context of their subject, Anglo-American legal education has always been less interdisciplinary than that of continental Europe. With the development of a more or less scientific approach to social studies in the 20th century, however, this is changing. Some law schools appoint economists, psychologists, or sociologists to their staffs, while others require or permit their students to take courses outside the law school as part of their work toward a degree. This awareness of the other social studies is thought to be more advanced in the United States than in Great Britain. Continental legal education (in both eastern and western Europe) tends to be highly interdisciplinary, with nonlegal subjects compulsory for students taking their first degree in law.

Traditionally, legal education included the teaching of legal history, which was once regarded as an essential part of any educated lawyer's equipment. While legal history has lost prestige in the sense that separate courses in the subject are offered in few law schools and, when optional, are not very popular among students, much legal history is, nonetheless, taught in the context of other courses. Since the corpus of the law is a constantly evolving collection of rules and principles, many teachers consider it necessary to trace the development of the branch of law they are discussing. In countries where most parts of the law are codified (as, for example, in continental Europe, Central and South America, the countries in the Mediterranean basin and in Africa that were formerly under French influence, Thailand, and Japan) it is not generally thought necessary to go back beyond the codes. On the other hand, in countries with a common-law system (England, most members of the British Commonwealth, and most parts of the United States), in which few branches of law are codified, knowledge of the law has traditionally depended to a great extent on the study of court decisions and statutes out of which common law evolved. This made the study of legal history of more immediate significance in such countries. But as the former case-law areas have increasingly come under statutory and administrative regulation, the practical importance of legal history has receded.

The graduating law student is not expected to have studied the whole body of substantive law. He is, however, expected to be familiar with the general principles of the main branches of law. To this end certain subjects are regarded as basic: constitutional law, governing the major organs of state; the law of contract, governing obligations entered into by agreement; the law of tort (or delict), governing compensation for personal injury and damage to property, income, or reputation; the law of real (or immovable) property, governing transactions with land; and criminal (or penal) law, governing punishment, deterrence, rehabilitation, and prevention of offenses against the public order. The chief materials are the same everywhere: codes (where these exist), reports of court decisions, legislation, government and other public reports, institutional books (in civil-law countries), textbooks, and articles in learned periodicals. The aim is not so much that the student should remember "the law" as that he should understand basic concepts and methods and become sufficiently familiar with a law library to carry out the necessary research on any legal problem that may come his way.

STUDY AND PRACTICE

To some extent all law courses are out of harmony with legal practice, for in real life a case is not presented as neatly by a client to his lawyer as it is in a textbook. The case usually begins as a statement, often jumbled, of facts and problems that cut across pedagogical categories. A story of a road accident, for example, may involve the lawyer in considering questions of the civil responsibility for the cause of the accident; of contract (in relation to insurance); of criminal law (in relation to a traffic offense); and of other branches of law as well. It is therefore important, while making divisions of law for convenience of study and examination, to guard students against the danger of thinking in compartments.

Lawyers must also contend in practice with branches of law in which they have received no formal education. More importantly, new social problems requiring legal attention and new legal structures come into existence during every lawyer's lifetime. His task may be eased if he has learned to look to the experience of other nations. A good law school produces a graduate who is not constricted by pedagogy but is trained to adapt himself to—and perhaps lead in bringing about—legal changes related to social, economic, and political developments.

The curriculum of the law school must also allow for the great diversity of careers followed by those who have been trained in the law. In most countries large numbers of persons with a legal training seek a career outside the practicing legal profession, commonly in the civil service, in municipal government service, in legal education, and in commerce and industry. Students' requirements and tastes differ; and most law schools, therefore, offer a choice. It is common to prescribe a certain number of compulsory subjects, which are regarded as essential to any law student's education, and leave a freedom of selection as to other subjects, stipulating only the number of courses to be studied. Except in the Soviet Union there is little uniformity from law school to law school within the same country as to which subjects are compulsory, and lists of optional subjects vary markedly.

The extent to which legal education aims to teach practice and procedure varies from place to place. Attention is always given to the methods of ascertaining the law from the books but not always to the ways of using this knowledge of the law in various roles, such as legal adviser or judge. Discussion of these matters tends to be more widespread in universities in the United States and in countries where the main qualification to practice the law is a university degree than it is in England and continental Europe, where professional training is provided outside the university and after graduation. In recent years clinical programs, in which students can have real or simulated experience in law practice, have become a staple part of the American law school curriculum. In the Soviet Union students are required to take one semester of clinical work. On the Continent such training is typically part of a postgraduate apprenticeship program (see below).

Courses on the rules and principles of court procedure are typically compulsory in university law schools. In England, however, few universities teach these, leaving them to the bar and solicitor's examinations, though the law of evidence (governing what facts may be proved in court, and how) is usually an optional subject; some knowledge of civil and criminal procedure may be picked up incidentally during the study of substantive law.

TEACHING METHODS

Methods of legal education are constantly changing, but the requirement of a university degree has become more or less uniform, coupled in many countries with the need to pass a qualifying examination organized by the profession. Apprenticeship, once a usual way of entering the profession in the common-law countries, has everywhere been increasingly displaced by university education, to which it has now become a supplement.

University law schools tend to differ along national lines in their methods of teaching. In the United States, following pioneer work by Christopher Columbus Langdell at Harvard in the latter half of the 19th century, the case method came to prevail, in which the student reads reported cases and other materials collected in a casebook and the class answers questions about them instead of listening to a lecture by the teacher. The casebook method has been adopted at some institutions in England and other common-law countries but has found scarcely any adherents elsewhere. Even in the United States most law schools now use seminars and lectures as well. The case method has the advantage of emphasizing the characteristic feature of the common law—the evolution of principles from decisions in actual cases—and thus of focusing the student's attention on the processes of analogy and distinction. It has the disadvantages, first, of being relatively time-consuming in relation to the amount of knowledge of legal principle that can be imparted and, second, of concentrating on a source of law that has become just one of many in modern statutory and regulatory legal systems. The traditional teaching techniques in English universities have been lectures and tutorials (or seminars).

In continental European countries the backbone of legal education is the formal lecture. Class sizes are typically very large compared with those in the United States and Britain, and lectures tend to be magisterial performances. Attendance is frequently voluntary, and those who stay away are usually able to secure the text of what they have missed. Seminars are given too, particularly for specialized subjects. Similar methods are used in other countries with large numbers of law students. In the Soviet Union, as in western Europe, the lecture method supplemented by smaller discussion groups is typical. But in the Soviet Union the lectures are well attended and participation in seminars is mandatory.

Teaching methods are not unrelated to the nature of the legal system. The methodology of continental legal education has grown out of and perpetuates a legal tradition heavily influenced by scholars, while the methods in England and the United States have emerged from and contribute to the maintenance of the tradition of judge-made law. Methods were influenced also by the fact that in England legal education was from early times in the hands of the bar, while on the Continent from the 12th century on it was the province of the universities. The fact that in common-law systems principles of law are largely derived by a process of inductive reasoning from many decisions of higher courts lay behind the development of the case method. In continental Europe the fact that law is found mainly in systematic legislation is one of the chief reasons for the lecture method, in which the subject can be approached through its philosophical background. A desire to systematically expound a body of principles rather than approaching facts and problems by the case-by-case method is met better by formal lectures and textbooks than by class discussion. This formal approach is reinforced in countries where published reports of local court decisions are scanty.

Much has happened in recent years, however, to diminish these contrasts. Law schools everywhere are seeking a better balance between theory and practice. American law professors increasingly consider the case method only one of several useful teaching techniques, while continental law faculties, particularly since the student unrest of the 1960s, have instituted various reforms designed to provide students with more opportunities for practical exercises and classroom discussion. The main obstacle faced by continental law faculties engaged in these efforts, however, is the unfavourable ratio of teachers to students.

EXAMINATIONS AND QUALIFICATIONS

The process of selecting members of the legal profession begins in the universities and law schools and continues afterward in the form of professional entrance requirements.

School examinations. In the United States, Great Britain, and other common-law countries students are generally required to pass an examination in each subject. Four or five subjects are studied simultaneously during the academic term, and students must take examinations in all of them at the end of the term or year. In France and Italy, too, students are required to pass a certain number of examinations in various subject matter areas in order to qualify for a degree.

In some continental European countries more comprehensive examinations are the rule. In West Germany the course work for the university law degree normally takes about five years, with a single comprehensive examination at the end of those five years (the First State Examination). Students are admitted to this examination if they produce certificates of satisfactory work in each subject, in a jurisprudence seminar, and in a course on economics and finance. The Netherlands has an intermediate system: the course for a first degree in law lasts five years, with an examination at the end of the second year and another at the end of the fifth. The Soviet Union combines the system of examinations in each course with a comprehensive examination at the end of the five-year period of study.

The method of subject-by-subject examination is less taxing on the memory than the system of comprehensive examination. It may well enable the student to do more detailed work on the problems of each subject. It has the disadvantage of encouraging him to think in terms of separate subjects, whereas the comprehensive examination leads him to consider legal problems in all their aspects. Being aware of the dangers of compartmentalized thinking, some law schools in the common-law world have introduced into their curricula "general" subjects, such as "common law," in place of separate courses in contract and tort, or they require the student to write papers about issues that relate to several of the subjects studied.

No formal test is wholly satisfactory as a method of screening potential lawyers. The type used most widely, in which students write answers to questions in an examination hall, has been criticized for placing too much emphasis on memory. This criticism is met to some extent in many universities by allowing candidates to consult books and reference materials during the examination, thus bringing the test a little closer to what a lawyer will do when confronted with a real problem. Another objection is that testing creates a situation of stress, in which a candidate does not necessarily demonstrate how he has benefited from his legal education, and also one in which the skill demonstrated in the examination hall is not all the skill required of a lawyer. In particular, the examination does not test capacity for patient research or the capacity for oral argument, which requires theses and oral examinations. Examinations to be done outside orthodox examination halls have thus been proposed.

Some universities in the United States, Great Britain, and the Commonwealth countries require one or more long essays or a short thesis or research paper as part of the work for a first degree in law (as opposed to the more substantial dissertation, or thesis, for a postgraduate law degree). This is commonly written during the final year with no restriction on the resources employed. A thesis in the last year of study is required in Soviet law schools and in some civil-law countries. Credit is also sometimes given for articles or notes published by students in law reviews. Such student publishing is more common in the

United States than elsewhere, partly because most U.S. law schools have their own legal journals and partly because American law students are nearly always college graduates. Such student work also enhances prospects of employment, particularly if the student becomes an editor of the law review.

Oral examinations are the rule in some countries, such as Italy and the Soviet Union. In the United States oral examinations are rare. French universities typically use both written and oral examinations. Some British and overseas Commonwealth universities hold oral examinations to confirm or resolve doubtful results on written papers or as a prerequisite to the award of first class honours. In Italy, where a law student must present a thesis after passing his other examinations, the thesis must be orally defended before examiners. The West German law student, after passing his written examination, has an oral one. In Japan, for professional qualification at the Legal Training and Research Institute (see below), there is an oral examination in each of the compulsory subjects after the written examination has been passed.

Qualifications for practice. *Common-law countries.* In England and Wales practicing lawyers must be either barristers (advocates and consultants) or solicitors (general legal advisers dealing with all kinds of legal business out of court and advocates in some of the lower courts). The former are organized in four Inns of Court (Lincoln's Inn, Inner Temple, Middle Temple, Gray's Inn) under the discipline of the Senate of the Inns of Court; the latter are under the jurisdiction of the Law Society. It is not necessary to hold a law degree or any university degree to qualify for the profession of law, but such a degree is usual. To become a barrister a candidate must pass a two-part examination in legal subjects, but university graduates may obtain partial or total exemption from the first part, depending on their degrees. A barrister's preparation also includes practical courses and a period of pupilage administered under the authority of the Senate of the Inns of Court. A barrister may not practice at all until he has undergone six months of pupilage in chambers and may not practice independently until he has been a pupil for a year. Pupilage causes some difficulty, partly because of the cost but mainly because of the increasing shortage of places in chambers. To qualify as a solicitor, the normal course is to serve as an articled clerk for two years and also pass law examinations in two parts. In Scotland and Ireland (both the republic of Ireland and Northern Ireland) there are similar requirements, though the arrangements differ in detail.

In the United States admission to the bar qualifies one for all types of legal work. The only formal requirements are the passing of state bar examinations after graduating from a law school; in a few states the law degree alone is sufficient.

In both England and the United States, as in many other common-law countries, becoming a judge or magistrate is a promotion (by appointment or election) from the ranks of the bar, and there is no special training for the exercise of judicial functions. But in some other common-law countries, especially in Africa and Asia, a newly qualified lawyer may enter the government legal service and find himself appointed in a short time to a junior magistracy. Even in these countries there is generally no special training for the job of adjudicating.

Civil-law countries. In continental European countries the qualifications to practice law typically depend on which of the various branches of the profession the university law graduate wishes to enter. Some countries place more emphasis on apprenticeship and others on examination. In France, for example, a legal practitioner may be an advocate, an *avoué,* a notary, or a judge. Each receives a different training, but all normally have gone through third- and fourth-year law degree courses. The advocate (roughly corresponding to the English barrister) must pass a bar examination and then serve as a probationary lawyer for three years, during which he takes further course work as well as acquiring practical experience. The *avoué* (something of a cross between a junior barrister and a senior solicitor) serves a period of articled clerkship and undergoes a professional examination by practicing lawyers. The notary (who does the noncontentious work performed in England by a solicitor) need not be a university graduate and can be a product of a professional school. His period of training lasts at least three years in a notary's office. He also takes a professional examination and if successful must wait for a vacancy since there is a limited number of notarial offices established by law.

In West Germany the graduate in law who seeks a legal career must embark upon a period of practical training as a *Referendar.* This is a uniform program involving two years of practical work in the courts, in the office of a lawyer in private practice, in the office of a public prosecutor, in the civil service, and sometimes in the legal department of a commercial concern. Upon its completion, he must pass a state examination (*Assessorexamen*).

A somewhat similar procedure is followed in Japan. Law graduates who seek careers as judges, public procurators, or lawyers in private practice must (with the exception of summary court magistrates and assistant procurators) pass the National Law Examination for entrance to the Legal Training and Research Institute. This is an organ of the Supreme Court. Like his German counterpart, the *Referendar* training to become an *Assessor,* the Japanese student at the institute is paid by the state. The period of training at the institute lasts two years. The bulk of the work consists of practical exercises and discussions, lectures on legal topics, and visits to institutions of concern to lawyers (such as prisons). The training is uniform, leads to a single examination, and qualifies the graduate for any branch of legal practice.

In some countries, such as France and Spain, there are special schools for training judges. In others, such as West Germany and the Nordic countries, judicial training is acquired in the post-law school practical internship period. In West Germany, for example, a law graduate may be appointed to a lower court after completing the *Referendarzeit* and passing the Second State Examination. After serving a three-year probationary period, he becomes eligible for an appointment for life. In France the first step to becoming a judge is to pass an annual competitive examination for which students prepare by taking a special program in their last year of law studies. Successful candidates then must undergo 28 months of training consisting of a period of formal study at the National School of the Judiciary in Bordeaux, followed by a series of short practical internships in such settings as police departments, law offices, prisons, and the Ministry of Justice in Paris. This training culminates in a judicial apprenticeship, during which the future judge participates on a daily basis in all the activities of a variety of courts. Upon completion of their training period, the students are ranked on the basis of their grades and the evaluations of supervisors and are then assigned to their first positions in the judicial system. Since the administrative law courts in France are not part of the judiciary but rather of the administration, most judges for these courts are drawn not from the lawyers trained in the National School of the Judiciary but from the civil servants trained in the National School of Administration.

LEVELS OF STUDY

Law degrees are undergraduate degrees in most countries. The student embarks upon the study of law at a university at about the age of 18. In France the universities offer a course of two years in duration that may be taken by anyone who has completed his secondary education. High marks in this entitle the candidate to enroll for the *licence-en-droit,* which is given at the end of the third year of study. Successful completion of a fourth year leads to a *maîtrise-en-droit,* which for all practical purposes has become the basic French law degree.

In the United States, by contrast, most law schools require the entrant to be a university graduate. Consequently, the U.S. student of law is generally in his early 20s. Other countries where legal education is organized on a graduate level include India and Pakistan.

University law schools in many countries accept all candidates with a certain level of prelegal education. One drawback of this open admissions system is a substantial failure rate in examinations. In countries where candidates are screened before admission to law school there is less attrition. In England, for example, each university imposes a quota on entry to its law school and selects among candidates on the basis, usually, of academic performance. In the United States candidates are selected on the basis of academic performance and the results of a test designed to demonstrate aptitude for the study of law. In both the United States and England entry requirements vary according to the prestige of the law school.

Most countries also provide for higher degrees in law. In common-law countries there is usually a series of steps, ascending through a degree of master of laws to a doctorate or senior doctorate. In civil-law countries it is normal to go straight from a first degree to a doctorate. Master's degrees are, as a rule, based on advanced examination after courses of instruction, though sometimes they are awarded for research or for a combination of examination and dissertation. Doctorates are awarded for theses expounding the results of original research and senior doctorates for published contributions to scholarship in the subject. In many countries there are also specialized postgraduate diplomas or certificates in particular subjects.

TRENDS IN LEGAL EDUCATION

Modern legal education is expanding both in quantity and scope, and formal university legal education has become dominant everywhere. The opportunities for university and professional education in law increased greatly after World War II. In England and Wales, for example, where before the war only the universities of Cambridge, Oxford, and London produced significant numbers of law graduates, there are now more than 50 academic law departments. A second wave of expansion took place, starting in the 1960s, when the numbers of law students, law instructors, and institutions teaching law grew dramatically almost everywhere. A particularly noteworthy development has been the increase in women law students, once a rarity but now constituting 30 to 40 percent of law students in most countries. The number of students enrolled in accredited law schools in the United States tripled between 1961 and 1980; thereafter demand for legal education began to level off and decline somewhat. A large increase in the teaching of law has occurred in Africa, where newly independent countries have established universities and professional schools concentrating on local laws and practice. Many governments have made provision, or greater provision, for the financial support of students, and legal education has been opened to a larger cross section of society in many places. The children of middle-class parents nevertheless continue to predominate in the field.

Since the late 1960s universities in several civil-law countries have departed from the rigidity of prescribed syllabi to allow a greater range of student choice in selecting subjects. In most countries, more attention is paid than formerly to foreign legal systems, transnational law, and to comparative law. In some countries nonlegal subjects have long been part of the syllabus; in others where law as a first-degree specialization has hitherto comprised only law studies, there has been a tendency to include nonlegal studies, joint courses in law and social sciences, or a more sociological approach to law.

Legal education has always had the problem of reconciling its aim of teaching law as one of the academic disciplines with its goal of preparing persons to become members of a profession. Most law schools are trying to find a middle path between being mere trade schools or citadels of pure theory. The criticism is often made that these efforts result in a type of education that is not practical enough to be really useful in resolving day-to-day legal problems but yet not as rigorously theoretical as a truly academic discipline ought to be. (L.A.S./M.A.Gl.)

BIBLIOGRAPHY

Legal profession: J.H. WIGMORE, *Panorama of the World's Legal Systems,* 3 vol. (1928, reissued in 1 vol., 1936), a comparative historical survey; DERK BODDE and CLARENCE MORRIS, *Law in Imperial China* (1967, reprinted 1973), a comprehensive account closely based on original sources; ROBERT J. BONNER, *Lawyers and Litigants in Ancient Athens: The Genesis of the Legal Profession* (1927, reprinted 1969); FRITZ SCHULZ, *History of Roman Legal Science* (1946, reprinted 1967), one of the few books on Roman law concentrating on the role of lawyers; HAROLD DEXTER HAZELTINE, "Roman and Canon Law in the Middle Ages," in *The Cambridge Medieval History,* vol. 5, ch. 21, pp. 697–764 (1926, reprinted 1968), a classic account; HERMAN J. COHEN, *History of the English Bar and Attornatus to 1450* (1929, reprinted 1967), a classic account of the history of the legal profession in the Middle Ages; MICHAEL BIRKS, *Gentlemen of the Law* (1960), a definitive work on the English solicitor; R.E. MEGARRY, *Lawyer and Litigant in England* (1962); and BRIAN ABEL-SMITH and ROBERT STEVENS, *Lawyers and the Courts: A Sociological Study of the English Legal System, 1750–1965* (1967), two opposing views of the legal profession in England; SIR FRED PHILLIPS, *The Evolving Legal Profession in the Commonwealth* (1978), a comparative survey of the legal profession, ethics, and education in countries with legal systems modeled on the English; DIETRICH RÜSCHEMEYER, *Lawyers and Their Society* (1973), a comparative study of the profession in West Germany and the United States; MARY ANN GLENDON, MICHAEL WALLACE GORDON, and CHRISTOPHER OSAKWE, *Comparative Legal Traditions* (1985), an introduction to the study of Romano-Germanic, English, and socialist law with material on the legal profession in those systems; C.J. DIAS *et al.* (eds.), *Lawyers in the Third World: Comparative and Developmental Perspectives* (1981); HIDEO TANAKA (ed.), *The Japanese Legal System* (1976, reprinted 1982).

Legal ethics: On the social nature and function of professional ethics, see ROBERT M. MACIVER, "The Social Significance of Professional Ethics," *Ann. Am. Acad. Polit. Soc. Sci.,* 297:118–124 (January 1955); and EMILE DURKHEIM, *Professional Ethics and Civic Morals,* trans. by CORNELIA BROOKFIELD (1957, reprinted 1983; orginally published in French, 1950). The leading texts on the ethics of the legal profession in the United States are HENRY S. DRINKER, *Legal Ethics* (1953, reprinted 1980); and L. RAY PATTERSON, *Legal Ethics: The Law of Professional Responsibility,* 2nd ed. (1984). The legal ethics of the Canadian legal profession are dealt with in MARK M. ORKIN, *Legal Ethics: A Study of Professional Conduct* (1957). For the English legal profession, see SIR THOMAS LUND, *Professional Ethics* (1970); and SIR WILLIAM BOULTON, *A Guide to Conduct and Etiquette at the Bar of England and Wales,* 6th ed. (1975). Information on legal ethics in other countries may be found in PIERRE G. LEPAULLE, "Law Practice in France," *Columbia Law Review,* 50:945–958 (1950); MAURO CAPPELLETTI, JOHN HENRY MERRYMAN, and JOSEPH M. PERILLO, *The Italian Legal System* (1967); TAKAAKI HATTORI, "The Legal Profession in Japan: Its Historical Development and Present State," in ARTHUR T. VON MEHREN (ed.), *Law in Japan* (1963); and K. OHIRA and G.N. STEVENS, "Admission to the Bar, Disbarment and Disqualification of Lawyers in Japan and the United States: A Comparative Study," *Washington Law Review,* 38:22–27 (1963). On the subject of contingent fees, see F.B. MACKINNON, *Contingent Fees for Legal Services: A Study of Professional Economics and Responsibilities* (1964). Specialization in the U.S. legal profession is discussed in BARLOW F. CHRISTENSEN, *Specialization* (1967). The same author discusses group practice in *Group Legal Services* (1967); his *Lawyers for People of Moderate Means: Some Problems of Availability of Legal Services* (1970), covers the subject of fees, specialization, group services, and advertising and solicitation.

Legal education: For England and Wales, much interesting information may be found in the GREAT BRITAIN. COMMITTEE ON LEGAL EDUCATION, *Report* (1971), which also contains a survey of legal education in Australia, Canada, France, West Germany, India, Italy, The Netherlands, New Zealand, Nigeria, Scotland, South Africa, Sweden, and the United States. In Great Britain the journal *Legal Studies* (three times a year) publishes articles on legal education, chiefly but not exclusively relating to the United Kingdom; a similar periodical in the United States is the *Journal of Legal Education* (quarterly). ROBERT STEVENS, *Law School: Legal Education in America from the 1850s to the 1980s* (1983), is a comprehensive treatment of the subject. Legal education in parts of Asia is discussed in REGIONAL CONFERENCE ON LEGAL EDUCATION, 1962, SINGAPORE, *Report,* ed. by S.P. KHETARPAL (1964); HAKARU ABE, "Education of the Legal Profession in Japan," in ARTHUR T. VON MEHREN (ed.), *op. cit.;* and JAY MURPHY, *Legal Education in a Developing Nation: The Korea Experience* (1967). For Africa, see JOHN S. BAINBRIDGE, *The Study and Teaching of Law in Africa* (1972).

(G.S./M.E.P./L.A.S./M.A.Gl.)

Animal Learning

That animals can learn seems to go without saying. The cat that runs to its food dish when it hears the sound of the cupboard opening; the rat that solves a maze in the laboratory; the bird that acquires the song of its species—these and many other common examples demonstrate that animals can learn. Yet what is meant by saying that animals can learn? What, in other words, is learning? This question proves exceedingly difficult to answer, and, in fact, some theorists propose that no single, all-encompassing definition of learning is possible. Moreover, a moment's reflection shows that there are different kinds of learning. The learning of number concepts, for example, surely seems to be of a different nature than the learning of the association between the sound of a cupboard door and the receipt of food. To explore animal learning, then, this article first considers what learning is and is not and then examines in detail some of the specialized types of learning that occur in animals.

The article is divided into the following sections:

The general nature of learning

Many animals live out their lives following fixed and apparently unvarying routines. Among numerous species of solitary insects, for example, the life cycle consists of the following unvarying events: the females lay their eggs on a particular plant or captured prey; the newly hatched larvae immediately start eating and then follow a standard sequence of developmental stages; the adults recognize appropriate mates by a set of fixed signs, perform a fixed sequence of mating responses, provision their eggs with suitable nourishment, and finally die before the next generation hatches. The same unchanging sequence is repeated generation after generation. And it is, of course, eminently successful. The same set of responses is invariably elicited by the same set of stimuli, because those responses were, and continue to be, adaptive. Where circumstances do not change, there is little need for an animal's behaviour to change. Even many aspects of the behaviour of mammals show a similar fixity. A dog withdraws its foot if it is pricked and a young child his hand if burned; both people and rabbits blink whenever an object is moved rapidly toward their eyes; the feeding behaviour of young infants of virtually all mammalian species consists of sucking elicited by contact with the lips.

Whenever the same response is always appropriate in a particular circumstance, there is little reason why an animal should need to learn what to do in that circumstance. But the world is not always so stable a place. The food supply that was plentiful yesterday may be exhausted today, and the foraging animal that always returns to the same spot will starve to death. Moreover, a particular food supply may be temporarily depleted but will be replenished if left long enough; the successful forager needs to remember where the supply was and when it was last visited, so as to time a return to advantage. In other words, circumstances may change, and the same response is not always appropriate to the same stimuli. Knowing what behaviour is appropriate may depend, therefore, on keeping track of past events.

POSSIBLE EXPLANATIONS OF BEHAVIORAL CHANGES

If an animal's behaviour toward a particular stimulus changes, one must look for an explanation of that change. One possible explanation is that the change is due to learning, but there are numerous other possibilities. If a definition of learning is to be provided, that definition must specify when to attribute the change to learning, and when to other causes.

At least two other major causes of behavioral change have been widely recognized. The first of these is motivation. A laboratory rat may pick up, chew, and swallow a pellet of food at one moment; half an hour later, after having eaten 20 grams of food, the rat will simply ignore any further pellets offered. Similarly, a male rat may mount and copulate with a receptive female introduced into his cage, but he will not repeat this pattern of behaviour endlessly even if offered the opportunity to do so. Some male territorial birds, such as chaffinches, will feed amicably beside other males at certain times of day or certain seasons of the year, but at other times they will launch an attack on any intruding male. In all these cases, it is more reasonable to attribute the change in behaviour not to anything the animal has learned but rather to a change in the creature's motivational state.

Changes due to motivation

It should not be thought, however, that just because all of these examples can be attributed to a single item—*i.e.,* motivation—that their detailed explanation will always be the same. The analysis of motivation is itself a large field of study, and it has proved to be more profitable to concentrate on the specific explanation of individual cases of changes in behaviour rather than to search for broad explanatory principles that end up being nearly vacuous. Nonetheless, it does seem possible to draw a contrast between motivational explanations for such changes and those that appeal to learning.

A second broad class of changes in behaviour can be attributed to maturation. We are inclined to ascribe the unfolding pattern of behaviour that emerges over the first few weeks of life to this ill-defined process. Newborn rat pups, for example, are relatively helpless; their eyes do not open for about two weeks, and their main sources of sensory input are probably touch and smell. As their sensory apparatus matures, the pups exhibit changed behavioral responses. The other obvious instance of a maturational change in behaviour is that which comes with sexual maturity: sexually mature adults of most species behave toward one another in ways quite different from those of

Changes due to maturation

younger members of the species. It is not only courtship and mating behaviour that change with sexual maturity; for instance, male puppies urinate in the same way as females, by squatting, and it is the onset of sexual maturity that produces the adult pattern of cocking a hind leg.

The concept of maturation is probably no better defined than that of motivation, and it is equally important to stress that it must cover a number of different processes. And, as with motivation, it is more profitable to analyze each case in detail, in order to uncover the precise mechanisms involved, than it is simply to label a change as an example of maturation. Indeed, at the level of physiological process, it seems probable that both motivational and maturational changes are often due to alterations in the hormonal state of the animal, and the distinction between the two is largely one between the unidirectional nature of the change in the case of maturation contrasted with the cyclical change common to short-term motivational states.

But how are these changes discerned from those that might be ascribed to learning? In many cases, of course, the answer is because a precise causal explanation has been provided: a great deal is known, at a physiological level, about the changes in brain and body associated with the motivational states of hunger and thirst. Even without any such detailed knowledge of the underlying mechanisms, it is possible to insist that certain changes in behaviour be attributed to motivation rather than to learning if the opportunity to learn anything relevant was lacking and the opportunity for a motivational change was present. If, for example, an animal that has been deprived of food for a long time behaves in one way toward food-related stimuli, but some hours later, after having been given ample opportunity to eat, it behaves differently toward those stimuli, the obvious interpretation is a motivational one. This interpretation would be strengthened if the animal had not come into contact with these stimuli during the intervening period and had been given, as far as one could judge, no other opportunity to learn anything about them. Learning, in other words, depends on certain kinds of opportunity, and a definition of learning may well turn out to be no more than a specification of the particular set of opportunities and experiences that produce it.

CIRCUMSTANCES THAT PRODUCE LEARNING

A particular change in behaviour is attributed to learning, then, because it is possible to specify the set of circumstances that produced it. What are those circumstances?

The role of practice

It is common to claim that learning depends on practice. (An older generation of experimental psychologists would have claimed that it depended on "reinforced" practice.) This definition can be misleading, however, if it causes one to attribute to learning all behavioral changes that follow what appears to be practice. In other words, it is not enough to show that an animal appeared to engage in practice and its behaviour subsequently changed. A temporal correlation of this sort does not establish a causal connection. Young birds, for example, are unable to fly, and their first attempts at flight are clumsy and ill-coordinated. Casual observation suggests that young birds improve with practice, gradually perfecting the set of skills they display as adults, but experimental analysis suggests that this practice may be unnecessary. Young birds have been brought up under restricted conditions that completely prevented their flying. When released at the age at which normally reared birds fly proficiently, the experimental subjects flew—without practice—as successfully as those that had spent their time in trial flight. The development of the skill appears to depend more on the maturation of strength and agility than on specific practice.

The notion that learning depends on practice also seems unduly restrictive and is, perhaps, an unnecessary legacy of an earlier version of behaviourism. It is not obvious that an animal should actually have to engage in a particular form of behaviour in order that this pattern of behaviour should be affected by learning. In many cases, indeed, no such practice seems necessary. The young of many songbirds must, it is quite clear, learn their species-typical song. There are several aspects to this learning process, one of which may indeed involve practicing the song at the beginning of the young bird's second season. But another critical aspect is simply exposure to the adult song at some point during the autumn of the young bird's first year, at a time when the young bird does not practice singing at all. Deprived of such experience, chaffinches and song sparrows produce an extremely impoverished version of the adult song; some finches may develop a song more characteristic of another species if that is what they heard during this period of their life. There are numerous other examples where learning appears to depend more on the opportunity to observe than on the opportunity for practice.

Broad versus narrow definitions of learning

This suggests that the definition of learning will have to refer to changes in behaviour that are attributable to particular kinds of experience. The danger now is that, as with motivation and maturation, the definition of learning will be so broad and vague as to be useless. As in those cases, it may be more profitable to concentrate on more detailed analysis of particular instances of learning. Such analysis has, for example, led to widespread agreement on the definition of classical conditioning, a particular type of learning whose study was pioneered by the Russian physiologist Ivan Petrovich Pavlov. In a typical experiment on classical conditioning, an experimenter might arrange a correlation between the ringing of a bell and the delivery of food to an animal. The animal predictably learns to direct food-related activity toward the sound of the bell. Analyses of such experiments have led to the definition of classical conditioning as a type of learning that occurs when there is a correlation between two stimuli and the animal's behaviour toward one of these stimuli changes in a predictable manner determined by the nature of the other. This definition, which will be expanded later in this article, is useful because it specifies both the circumstances responsible for learning (a temporal correlation between two stimuli) and the general way in which experience of those circumstances changes behaviour (the animal starts directing toward one stimulus responses that are related to those normally directed toward the other). Experimental psychologists and ethologists, however, have devised a tremendous range of procedures for studying learning in animals. The range and variety are such that it may be well-nigh impossible to formulate a meaningful definition of the circumstances that produce learning, for the definition either will be so restrictive that it clearly applies to only a fraction of the cases that should be regarded as instances of learning, or it will be so broad that it says nothing.

Rather than pursue any further the attempt to find an all-embracing, single definition of learning, it seems more useful to provide narrower definitions for particular cases, along the lines suggested above for classical conditioning. One consequence of this approach is that it may encourage the belief that learning consists of a large number of distinct processes that have nothing in common with one another. It is, of course, an open question as to whether this is true: it is certainly possible that, just as with the concept of motivation, the layman's concept of learning encompasses a large number of different cases whose underlying mechanisms are quite distinct. It is important not to prejudge this issue. Insistence on a single, global definition may well tend toward just such prejudgment by encouraging the belief that learning is a single, common process. To start by drawing some distinctions between types of learning does not rule out the possibility of seeing whether the various cases studied do have anything in common.

The interplay of theory and definition

In the final analysis, as is true of all scientific definitions, the definition of learning is a matter of theory. It has been said that a good scientific definition is the end product of good theory and experiment, not the starting point. Thus, there is a single process of learning if it turns out to be possible to devise a single theory that adequately accounts for the variety of cases in which learning is assumed to occur. Superficial appearances may be deceptive: just because the circumstances that produce learning in two cases, along with the consequences of that learning, appear quite different, it does not follow that the processes underlying learning are different. For instance, the phenomenon of filial imprinting, first seriously analyzed by the Austrian ethologist Konrad Lorenz, appears to be a highly specialized form of learning in which a newborn animal (*e.g.,* a chick, duckling, or gosling) rapidly learns to follow the first salient, moving object it sees. Normally this object will be the mother, but Lorenz discovered that the range of potential imprinting objects is large, extending from Lorenz himself to a bright red ball. There is no question but that some process of learning occurs here, and Lorenz assumed it to be highly specialized. Yet one theory seeks to explain imprinting in terms of simple classical conditioning. Whether or not the account of imprinting provided by this theory is correct, the point is made that how learning is defined and whether it is defined as a single, monolithic process or as many specialized processes are, in the end, questions of theory.

Types of learning

SIMPLE NONASSOCIATIVE LEARNING

When experimental psychologists speak of nonassociative learning, they are referring to those instances in which an animal's behaviour toward a stimulus changes in the absence of any apparent associated stimulus or event (such as a reward or punishment). Studies have identified two major forms of simple nonassociative learning, which are to some extent mirror images of one another: habituation and sensitization.

Habituation. A classic example of habituation is the following observation on the snail *Helix albolabris.* If the snail is moving along a wooden surface, it will immediately withdraw into its shell if the experimenter taps on the surface. It emerges after a pause, only to withdraw again if the tap is repeated. But continued repetition of the same tapping at regular intervals elicits a briefer and more perfunctory withdrawal response. Eventually, the stimulus, which initially elicited a clear-cut, immediate response, has no detectable effect on the snail's behaviour. Habituation has occurred.

Habituation can be defined in behavioral terms as a decline in responding to a repeatedly presented stimulus. As such, it is a very widespread phenomenon, one that can be observed in animals ranging from single-celled protozoans to humans. Most animals behave differently to novel and familiar stimuli: the former sometimes elicit startle responses, sometimes investigatory or exploratory responses; the latter often apparently are ignored. The suggestion that habituation is a simple form of learning, however, implies that it can be distinguished from some even simpler potential causes of this sort of change in behaviour. One reason why an animal might stop responding to a stimulus is that it no longer detects the stimulus; *i.e.,* some form of sensory adaptation might have occurred. Another potential cause is fatigue: perhaps some temporary refractory state is produced by repeated elicitation of the same response, making it impossible to perform that response again. Whether or not one would want to call either of these processes a form of learning is doubtful. But both behavioral and physiological evidence establishes that habituation cannot be explained in these terms.

Evidence that habituation is a form of learning

The critical behavioral evidence is that habituation can be disrupted by almost any change in the experimental conditions. If repeated presentation of one stimulus leads to habituation of a response, the same response can still be elicited by a different stimulus. Even if the experimenter presents a novel stimulus that does not itself elicit the response in question, its presentation may restore the response on the next trial in which the originally habituated stimulus is presented. This latter observation, usually referred to as an instance of dishabituation, seems to rule out any simple sensory adaptation; both observations rule out simple effector fatigue.

Neurophysiological analysis of habituation in various mollusks—for example, in the sea snail *Aplysia*—has confirmed that habituation need not depend on changes in the activity of sensory or motor neurons. In the case of *Aplysia,* researchers have studied the gill withdrawal reflex, a response that rapidly habituates to repeated stimulation of the snail's siphon or mantle shelf. But habituation still occurs even if it is elicited by direct, electrical stimulation of the motor nerve, bypassing the sensory receptors completely; and recording from the sensory nerve during normal habituation reveals no decline in its level of activity. These observations eliminate sensory adaptation as a possible cause of the animal's having ceased to respond to the stimulus. Effector fatigue can be ruled out by showing that direct stimulation of the motor neurons controlling the withdrawal response can still elicit a perfectly normal reaction even after the response has completely habituated. Research shows that habituation in *Aplysia* depends on changes in the activity of more central neurons. Repeated tactile stimulation of the siphon, leading to habituation of the withdrawal response, causes changes in the activity of the motor neurons innervating the response. Specifically, these motor neurons show a decline in excitatory postsynaptic potential, which is the electrical change that enables the nerve impulse to cross the gap (synaptic cleft) that separates one neuron in the pathway from the next. The decline in excitatory postsynaptic potential short-circuits the response. Moreover, the presentation of a novel stimulus, sufficient to dishabituate the behavioral response, restores the postsynaptic potential.

Habituation occurs even in animals without a central nervous system—probably in single-celled protozoans; certainly in animals such as the coelenterate *Hydra,* which have a diffuse nerve net and do not appear to be capable of associative learning. Among mammals, habituation of certain reflex responses can be observed even in "spinal" subjects, that is, those whose spinal cord has been severed from the brain. There can be little doubt, then, that habituation is not only widespread, but that it also can be a relatively simple phenomenon. There is, however, no guarantee that it is the same phenomenon wherever it appears. The waning response to a repeatedly presented stimulus admits of a number of different explanations. In principle, as we have already seen, it might be due to sensory adaptation, effector fatigue, or a more central neural change. These distinctions make rather little sense in the case of a single-celled animal. And one should not necessarily expect the habituation observed in a spinal mammal to involve precisely the same mechanisms as those responsible for comparable behavioral effects in an intact animal. Some psychologists have proposed theories of habituation that appeal to processes of classical conditioning. Such a theory is not likely to apply to the habituation observed in an animal that shows no capacity for classical conditioning.

Habituation is usually, as here, classified as an instance of simple, nonassociative learning. It is supposedly nonassociative because all that happens in the course of habituation is that a stimulus is repeatedly presented and the animal's behaviour changes; there is, on the face of it, no other event with which the stimulus can be associated. Habituation must therefore, it appears, be understood by reference to some change in the pathway between stimulus and response, and the work with *Aplysia* and other mollusks shows how this analysis may proceed at the physiological level. But if habituation is not always the same phenomenon, it is possible that different processes may underlie the habituation of the startle response to a loud noise in an intact mammal. And despite appearances to the contrary, those processes may involve some associative learning. One suggestion is that novel stimuli elicit a biphasic response: an initial increase in startle responses, which include components of emotion or anxiety, followed by a rebound in the opposite direction. Habituation occurs when the latter, rebound response becomes conditioned to the stimulus, occurring sooner and sooner with each repetition of the stimulus and thereby damping down and eventually canceling out the initial reaction. An alternative possibility is that long-term habituation depends on associating the repeatedly presented stimulus with the context in which it occurs, a suggestion that would explain why presentation of the stimulus in a different context sometimes leads to dishabituation.

Possible associative bases for habituation

The generality of habituation implies that this behavioral phenomenon has considerable adaptive significance; if true, it would be quite reasonable to expect that a number

Adaptive value of habituation

of different mechanisms might have evolved to produce the behavioral result. The adaptive value of habituation is not difficult to see. A novel stimulus may signify danger, and an animal should react to this stimulus either by withdrawing or at least by orienting toward it to see what will happen next. But if the same stimulus occurs again with no further consequence, it is probably safe: regular repetition of the same stimulus implies that it is part of the background, such as the waving of a branch in the wind or the shadow caused by a piece of seaweed floating with the waves. If the stimulus is not dangerous, time should not be wasted on it. Withdrawal, especially in the case of a snail into its shell, is a time-consuming effort, incompatible with such vital activities as searching for food. If it is important, therefore, for animals to be wary of novel stimuli, it is equally important that they should discriminate the novel and potentially dangerous from the familiar and probably safe.

Sensitization. The effect of habituation is to eliminate unnecessary responses, but the main function of learning has usually been thought to be the production of new responses. Traditional psychological theories of learning have assumed that the learning of new patterns of behaviour comes about through the association of a new response with a particular stimulus. Consequently, psychologists usually have either ignored the possibility that nonassociative processes might be sufficient to increase the probability of a new response or regarded it as a nuisance that interferes with the measurement of associative changes. They have rarely treated it as a subject worthy of study in its own right.

This is unfortunate, for the nonassociative phenomenon of sensitization is probably fairly widespread, and it provides a simple means of acquiring adaptive behaviour. Sensitization is said to occur when the repeated presentation of a particular significant stimulus (such as food or electric shock) lowers the threshold for the elicitation of appropriate behaviour to the point where a second stimulus, not normally capable of calling forth that behaviour, now does so. A typical example is provided by the behaviour of the marine worm *Nereis.* If the worm is kept in a small tube and fed at regular intervals, it becomes progressively more likely to respond to any novel stimulus, such as a change in illumination, by exploratory, food-seeking movements toward the open end of the tube. If, on the other hand, the worm receives mild electric shocks at regular intervals, it becomes progressively more likely to respond to a novel stimulus by withdrawal.

Relationship of sensitization to habituation

The first point to note about sensitization is its relationship to habituation. Habituation refers to a decline in the probability of responding to a repeatedly presented stimulus. Sensitization, by contrast, refers to an increase in the probability that behaviour appropriate to a repeatedly presented stimulus will occur, even in response to another stimulus. Although these two outcomes cannot be observed simultaneously, it is quite possible that the same operation—repeated presentation of a stimulus—can simultaneously engage two different processes, one causing a decline in the probability of responding to that stimulus, the other causing an increase. Experimental analysis suggests that both processes are real and may be engaged in the same experiment, so that the observed change in behaviour actually results from a mixture of the two. Typically, the process of habituation wins out, and what is observed is an overall decline in responding. But a common finding in habituation experiments is that responding initially increases before declining; the implication is that the initial presentations of a stimulus result in more sensitization than habituation, while further presentations produce more habituation than sensitization. A second factor influencing the relative importance of the two processes is the intensity or significance of the stimulus. A weak stimulus, or one with little intrinsic biological significance, will show relatively rapid habituation and little or no initial sensitization. A stronger stimulus, especially one, such as food or shock, that has substantial significance to the animal, may show marked sensitization and relatively little habituation.

The second point about sensitization is that it may mimic the effect of associative learning or conditioning. As has been mentioned, in a classical conditioning experiment a neutral stimulus, such as a change in illumination, is paired with the delivery of a significant stimulus, such as food or shock. Repeated pairing causes the neutral stimulus to elicit responses initially called forth by the significant stimulus; for example, a change in illumination that has been associated with an electric shock would come to elicit retreat or withdrawal. But in the case of the worm *Nereis,* experiments demonstrate that the light would come to elicit this change in behaviour whether or not it had been paired with shock: all that is needed is sufficient exposure to the shock. To attribute the change in behaviour toward the light to its association with food or shock, one must show that this change is greater than that which would have resulted from sensitization alone.

Relationship of sensitization to associative learning

The physiological processes underlying sensitization, like those underlying habituation, have been analyzed in experiments on such invertebrate species as *Aplysia.* Not surprisingly, the mechanisms involved appear to mirror one another. Whereas habituation is correlated with a decline in postsynaptic potentials, sensitization is correlated with an increase in the magnitude of postsynaptic potentials at the same locus.

Adaptive value of sensitization

Although sensitization has often been treated as a nuisance whose effects must be controlled in studies of habituation or associative learning, it remains a process worthy of study in its own right, for the behavioral changes it produces can have significant adaptive value. Without requiring the presumably more complex neural machinery necessary to subserve associative learning, sensitization enables animals to respond to local variations in the occurrence of significant events. If an animal's sources of food tend to occur together (that is, they are not distributed randomly in time or space), then it pays that animal, having once found food, to continue to behave in a food-gathering manner. Conversely, the animal that is increasingly wary after exposure to danger will have a better chance of evading a lurking predator. Sensitization thus enables an animal to take advantage of statistical regularities in the occurrence of significant events, without requiring it to detect other events that predict the significant ones. No doubt, further advantage accrues to the animal that can perform such calculations, for associative learning provides a powerful means of predicting the future. But there can be equally little doubt that such a process requires a more elaborate nervous system.

ASSOCIATIVE LEARNING: CONDITIONING

The study of animal learning in the laboratory has long been dominated by experiments on conditioning. This domination has been resisted by critics, who complain that conditioning experiments are narrow, artificial, and trivial, and, as such, miss the point of what animals are adapted to learn. From the critics' point of view, one unfortunate effect of their attacks has been the progressive refinement and elaboration of the theory of conditioning to the point where it can often explain the exceptions to which they drew attention. This is not to insist that associative learning is the sole, or even the most important, form of learning in vertebrates, but rather to introduce the idea that the processes underlying conditioning may be more interesting than older theories and an earlier generation of textbooks suggested.

Classical and instrumental conditioning. Pavlov was not the first scientist to study learning in animals, but he was the first to do so in an orderly and systematic way, using a standard series of techniques and a standard terminology to describe his experiments and their results. In the course of his work on the digestive system of the dog, Pavlov had found that salivary secretion was elicited not only by placing food in the dog's mouth but also by the sight and smell of food and even by the sight and sound of the technician who usually provided the food. Anyone who has prepared food for his pet dog will not be surprised by Pavlov's discovery: in a dozen different ways, including excited panting and jumping, as well as profuse salivation, the dog shows that it recognizes the familiar precursors of the daily meal. For Pavlov, at first, these "psychic secretions" merely interfered with the planned study of the digestive system. But he then saw that he had a tool for the objective study of something even more interesting: how animals learn. From about 1898 until 1930, Pavlov occupied himself with the study of this subject.

Pavlov's experiments

Pavlov's experiments on conditioning employed a standard, simple procedure. A hungry dog was restrained on a stand and every few minutes was given some dry meat powder, an event signaled by an arbitrary stimulus, such as the ticking of a metronome. The food itself elicited copious salivation, but, after a few trials, the ticking of the metronome, which regularly preceded the delivery of food, also elicited salivation. In Pavlov's terminology, the food is an unconditional stimulus, because it invariably (unconditionally) elicits salivation, which is termed an unconditional response. The ticking of the metronome is a conditional stimulus, because its ability to elicit salivation (now a conditional response when it occurs in reaction to the conditional stimulus alone) is conditional on a particular set of experiences. The elicitation of the conditional response by the conditional stimulus is termed a conditional reflex, the occurrence of which is reinforced by the presentation of the unconditional stimulus (food). In the absence of food, repeated presentation of the conditional stimulus alone will result in the gradual disappearance, or extinction, of its conditional response. In translation from the Russian, the terms "conditional" and "unconditional" became "conditioned" and "unconditioned," and the verb "to condition" was soon introduced to describe the experimental activity.

Thorndike's experiments

To the American psychologist Edward L. Thorndike must go the credit for initiating the study of instrumental conditioning. Thorndike began his studies as a young research student, at about the time that Pavlov—already 50 years old and with an eminent body of research behind him—was starting his work on classical conditioning. Thorndike's typical experiment involved placing a cat inside a "puzzle box," an apparatus from which the animal could escape and obtain food only by pressing a panel, opening a catch, or pulling on a loop of string. Thorndike measured the speed with which the cat gained its release from the box on successive trials. He observed that on early trials the animal would behave aimlessly or even frantically, stumbling on the correct response purely by chance; with repeated trials, however, the cat eventually would execute this response efficiently within a few seconds of being placed in the box.

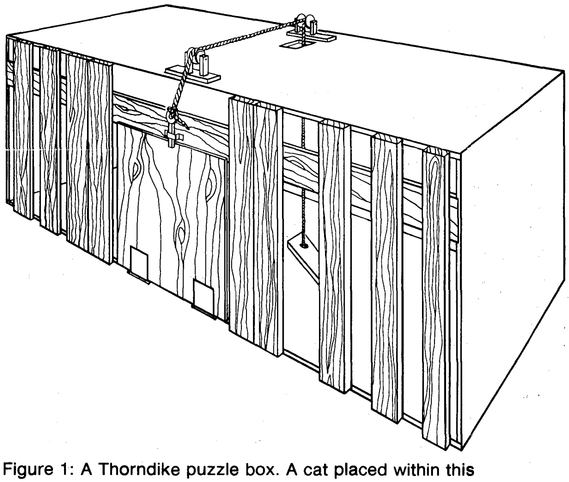


Figure 1: A Thorndike puzzle box. A cat placed within this apparatus could obtain its release by pressing a pedal inside the box.

Skinner's experiments

Thorndike's procedures were greatly refined by another U.S. psychologist, B.F. Skinner. Skinner delivered food to the animal inside the box via some automatic delivery device and could thus record the probability or rate at which the animal performed the designated response over long periods of time without having to handle the animal. He also adopted some of Pavlov's terminology, referring to his procedure as instrumental, or operant, conditioning; to the food reward as a reinforcer of conditioning; and to the decline in responding when the reward was no longer available as extinction. In Skinner's original experiments, a laboratory rat had to press a small lever protruding from one wall of the box in order to obtain a pellet of food. Subsequently, the "Skinner box" was adapted for use with pigeons, who were required to peck at a small, illuminated disk on one wall of the box in order to obtain some grain.

In experiments on both classical conditioning and instrumental conditioning, the experimenter arranges a temporal relation between two events. In Pavlov's experiment the food was always preceded by the conditional stimulus; in Skinner's original experiment the delivery of food was always preceded by the rat's pressing the lever. Conditioning, or associative learning, is inferred if the animal's behaviour changes in certain ways and if that change can be attributed to the temporal relationship between these events. If the dog started salivating to the ticking of a metronome just because it had recently received food, rather than because the delivery of food had been signaled by the metronome, this should be regarded as an instance of sensitization rather than associative learning. One of the simplest ways of establishing that the change in behaviour results from the temporal relationship between the conditional stimulus and the unconditional stimulus in a classical experiment, or between the response and the reinforcer in an instrumental case, is to impose a delay between the two. A gap of even a few seconds between the rat's pressing the lever and the delivery of food will seriously interfere with the animal's ability to learn the connection. And although in some classical experiments evidence of conditioning can be found in spite of relatively long gaps between the conditional stimulus and the unconditional stimulus, increasing this interval beyond a certain point invariably causes a decline in conditioning.

Laws of associative learning. The temporal relation between the conditional stimulus and the unconditional stimulus, or between the response and the reinforcer, was for a long time regarded as the primary determinant of conditioning. Conditioning is certainly a matter of associating temporally related events, but temporal contiguity is only one of several factors—and probably not the most important—that influences conditioning. A variety of experiments have shown that classical conditioning will occur only if the conditioned stimulus is the best predictor of the occurrence of the unconditional stimulus. In other words, it is the correlation between two events, just as much as their temporal contiguity, that establishes an association between them. A pigeon, for example, will learn by classical conditioning to peck an illuminated disk in a Skinner box if, whenever the disk is illuminated, food is delivered. This temporal relationship between the light and food can be preserved intact, but if the experimenter now arranges that food is equally available at other times (when the light is not on), the pigeon will not peck at the illuminated disk. Delivering food at other times destroys the correlation between light and food (although leaving the temporal relationship untouched) and abolishes conditioning.

Conditioning occurs to the best predictor

Although some conditioning will occur when the conditional stimulus is not perfectly correlated with the delivery of food (perhaps because on a proportion of trials the conditional stimulus is presented alone without food) or when the temporal relationship is less than perfect (there is a gap between the conditional stimulus and the delivery of food), this conditioning is abolished if the experimenter ensures that there is some better predictor always available. If a dog is conditioned to the ticking of a metronome paired with the delivery of food, the animal will salivate in response to the metronome even if the food is presented in no more than 50 percent of the trials. If, however, a light is illuminated on those trials when the metronome is accompanied by food, and not on the remaining 50 percent of the trials, the dog will become conditioned to the light and not to the metronome. Similarly, a pigeon will learn to peck at a disk illuminated with red light even if a gap of several seconds separates this response from the delivery of food. But if, during this interval, after the red light has been turned off and before food is delivered,

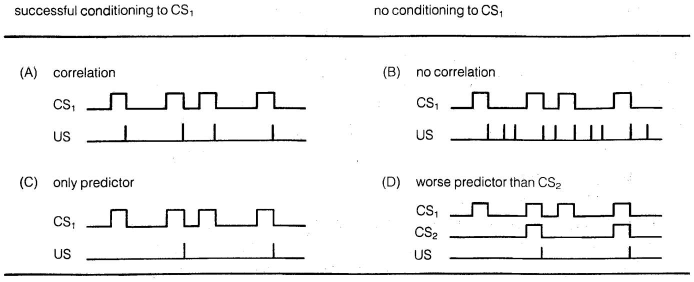

successful conditioning to CS_1
no conditioning to CS_1
(A) correlation
CS_1
US
(B) no correlation
CS_1
US
(C) only predictor
CS_1
US
(D) worse predictor than CS_2
CS_1
CS_2
US

Figure 2: Effects on conditioning of various relationships between a conditioned stimulus (CS$_1$) and an unconditioned stimulus (US). (A) and (B) show that temporal contiguity of CS$_1$ and the US is not by itself sufficient to ensure conditioning; (C) and (D) demonstrate that conditioning occurs to the best predictor of the US (see text).

a green light is turned on, the pigeon will never learn to peck at the red light. It is as though the pigeon attributes the occurrence of food to the most recent potential cause (now the green light rather than the red), and the dog attributes food to the stimulus best correlated with its delivery (the light rather than the metronome). Conditioning, in other words, occurs selectively to better predictors of reinforcement at the expense of worse predictors. This same principle explains the earlier observation of the role of correlation in general. The pigeon will not associate the illumination of the disk with food if food is equally probable both when the light is on and when it is switched off; from the pigeon's point of view, food occurs whenever the animal is placed in the Skinner box. The illumination of the light signals no increase in the probability of food, and the best predictor of food is the mere fact of being in the Skinner box.

Built-in biases toward certain associations

Temporal contiguity, therefore, is not necessarily the most important factor in successful conditioning. Moreover, there is yet another factor that should be stressed. It will hardly have escaped the reader's attention that there is an astonishing artificiality to the typical conditioning experiment conducted by Pavlov or Skinner. An animal is placed in a bare, confined space; lights are flashed on and off; the animal is permitted to operate some mechanical contrivance; some meat powder or a pellet of food is delivered. How could one possibly suppose that the ways in which animals learn anything of importance in the real world will be illuminated by this contrived and restrictive kind of experiment? This question raises large issues, some of which will recur at later points in this article. But one point should be acknowledged right away: the more restricted the range of experimental manipulations employed, the greater the chance that the investigator will completely miss important principles. Experiments with lights and metronomes failed to reveal the following important principle of conditioning: animals appear to have built-in biases toward associating some classes of stimuli with certain classes of consequences. The most dramatic instance of this principle is provided by conditioned food aversions. If rats eat some novel-flavoured substance and shortly thereafter are made mildly ill (for example, by an injection of a drug such as apomorphine or lithium chloride), they afterward will show a marked aversion to the novel food. Because they will show an aversion even though an interval of several minutes, or sometimes even hours, intervenes between eating the food and the onset of the illness, there has been some question as to whether this should be regarded as an instance of conditioning at all. But the parallels between food aversions and other forms of conditioning are so extensive that it is hard to believe that some common processes are not involved. And there is no question but that the length of the interval is important; other things being equal, rats will form a stronger aversion to a food they have eaten recently than to one they have eaten several hours earlier.

The most interesting feature of such aversions is that they are, by and large, confined to foods. If rats suffer the unpleasant experience of being made ill, they are not likely to show an aversion to anything other than a novel-tasting food or drink they have recently ingested. As in other forms of conditioning, the novelty of the potential conditional stimulus is important. Rats will not show any marked aversion to a thoroughly familiar diet unless the experience of illness is repeatedly induced shortly after eating the daily ration, just as, in Pavlov's experiments, conditioning will proceed only slowly to the ticking of a metronome if the dog has heard this sound repeatedly before. The more striking restriction, however, is that it is the taste of the food or drink that is associated with illness. If rats drink plain tap water before being made ill, they will show little aversion to tap water (since there is no novelty here). But even if a novel buzzer is sounded while they are drinking and they are then made ill, they will not associate the buzzer with the illness. This is certainly not because rats are unable to associate the buzzer with an aversive consequence. If drinking water while the buzzer is sounded produces a mild electric shock, rats will rapidly learn to stop drinking whenever they hear the buzzer. In this case it is the flavour of the water that rats find difficult to associate with the shock; punishing rats with a mild shock whenever they drink sugar-flavoured water has little effect on their tendency to drink sugar-flavoured water. The flavour of food or drink is readily associated with subsequent illness, but only poorly associated with other painful consequences. Conversely, an external stimulus such as a buzzer or flashing light, which is readily established as a signal for shock, is only with great difficulty associated with illness. These relationships are summarized in the Table.

Occurrence of Conditioning to Certain Combinations of Conditional Stimulus and Unconditional Stimulus

conditional stimulus	unconditional stimulus	outcome
Water with novel flavour	Lithium-induced illness	Conditioned aversion to flavour
Water with novel flavour	Mild shock	No aversion to flavour
Water plus buzzer	Lithium-induced illness	No aversion to buzzer
Water plus buzzer	Mild shock	Conditioned aversion to buzzer

Adaptive value of biases

The full explanation of this finding remains uncertain. It is known that even very young rats show such selectivity, so it cannot depend solely on any prior experience. What is easy to see is that this behaviour makes biological sense. Internal malaise, such as that caused in the psychologist's experiment by an injection of lithium, will in the real world usually be a consequence of eating spoiled or poisonous food or of drinking tainted water. The most reliable sign of such food or drink will be its taste, and animals predisposed to associate the taste of what they have ingested with subsequent illness are likely to be better equipped to avoid potentially harmful food in the future. On the other hand, painful injury, mimicked in the laboratory by a brief electric shock, is hardly likely to be a consequence of eating food of a particular flavour; it will usually be caused by external circumstances, such as contact with a sharp or very hot object or a narrow escape from a predator. The natural suggestion is that the function of conditioning is to enable animals to find out what causes certain events of biological significance. If this is so, a built-in bias toward associating certain classes of events together makes adaptive sense. Conditioning is not just a matter of associating two events because one happens to follow the other; it is more profitably seen as the process whereby animals discover the most probable causes of events of consequence to themselves.

Laws of performance. Conditioning could have no function at all, however, if it did not involve changes in an animal's behaviour. Nor could scientists infer that conditioning has occurred unless they could observe, at some point, a change in an animal's behaviour attributable to certain conjunctions of events. So, although conditioning may involve the formation of associations between events or the attribution of particular events to their most probable antecedent causes, it must also include some mechanisms for translating these associations into changes in behaviour.

Stimulus–response theory

For an earlier generation of behaviourists, the fundamental fact about conditioning was precisely that it changed behaviour, and the theories they advanced were determined by this fact. The description of conditioning as the establishment of a new response to a stimulus that had not previously elicited that response naturally suggested that conditioning was a matter of forming new stimulus–response connections. This conceptualization led to the development of the stimulus–response theory, variations of which long provided the dominant account of conditioning. One version of the stimulus–response theory suggested that the mere occurrence of a new response to a given stimulus, as when Pavlov's dog started salivating shortly after the metronome had started ticking, is in itself sufficient to strengthen the connection between the two. Thorndike, however, argued that the probability that a particular stimulus will repeatedly elicit a particular response depends on the perceived consequences of this response. According to this view, new stimulus–response connections are strengthened only if the response is followed by certain kinds of consequences.

Thorndike's law of effect

There are several questions raised here, and it is important to keep them distinct. One is whether responses are sometimes (or even always) modified by their consequences. Although denied by some theorists, their denial seems distinctly paradoxical. A rat whose presses on a lever are followed by the delivery of a food pellet will press the lever again; if the only consequence of pressing the lever is the delivery of a painful shock, the rat will desist from this action. Thorndike's law of effect—which stated that a behaviour followed by a satisfactory result was most likely to become an established response to a particular stimulus—was intended to summarize these observations, and it is surely an inescapable feature of understanding how and why humans and other animals behave. In keeping with this understanding, parents reward children for good behaviour and punish them for bad. When this fails to produce the desired behaviour, we are inclined to argue that the child is finding other sources of reward or does not find the intended punishment particularly unpleasant, or that the parents' behaviour is hopelessly inconsistent. We are far less likely to question the assumption that, other things being equal, people (and other animals) repeat actions that have desirable consequences and avoid repeating those that have undesirable consequences.

Thorndike's law of effect was, however, also a theory of how reward and punishment modify behaviour. This theory, which states that behaviour normally is modified by changing the strength of stimulus–response connections, finds less general acceptance today. A simple experiment suggests one reason for this. A rat is trained to press a lever in a Skinner box, being rewarded with a small quantity of sucrose solution for each press of the lever. Once the response has been established, the rat is removed from the Skinner box. The next day, while in its home cage, the animal is given sucrose solution to drink and shortly thereafter is made ill by an injection of lithium. Once this treatment has established a strong aversion to the sucrose, the rat is returned to the Skinner box, where, despite the opportunity to do so, the animal does not press the lever again. The result is hardly surprising: there is no reason to expect the rat to perform a response whose sole consequence is the delivery of the now aversive sucrose solution. But this behaviour cannot be explained by Thorndike's theory, for according to Thorndike all that the rat learned in the first stage of the experiment was a new stimulus–response habit; stimuli from the Skinner box should, by Thorndike's reasoning, now elicit the response of pressing the lever. Thorndike's stimulus–response theory credits the rat with no acquired knowledge of the connection between pressing the lever and obtaining sucrose; the function of sucrose is merely to strengthen the stimulus–response connection.

That responses are modified by their consequences, therefore, need not call for Thorndike's theoretical account of this fact. It is probably more reasonable to suppose that animals learn about the relationship between their actions and consequences (just as they can also learn about the relationship between any other classes of events), and that they then modify their actions in accordance with the current value of these consequences. The next question to consider is whether this is an entirely general principle of performance, or whether it applies only to some classes of response in some kinds of situations. Why, for example, does Pavlov's dog start salivating to the ticking of the metronome? Is it because the response of salivating is followed by a rewarding consequence? The response is, at first, elicited by the sight of food and is shortly followed by the rewarding consequence of chewing and swallowing the food. But another simple experiment suggests that salivating to the metronome is not strengthened because it is followed by food. The experimenter can turn on the metronome for five seconds on each trial, at the end of which time the dog receives food—but only if it did not start salivating before the arrival of food. Now the response of salivating to the metronome is followed by an undesirable consequence, the cancellation of the food that would otherwise have been delivered on that trial, but the dog still cannot help salivating (at least sometimes) to the metronome. The implication is that salivating is not a response modified by its consequences, but one reflexly elicited by food and also by any stimulus associated with food. Voluntary responses can be modified by their consequences; involuntary responses (such as blushing when a person is embarrassed or the release of adrenalin when a person is angry or afraid) cannot. The reason Pavlov's dog starts salivating to the metronome is, just as Pavlov himself supposed, that the association between metronome and food means that the metronome can substitute for food. To put it another way, the metronome now produces activity in neural centres normally responsive to the delivery of food, activity that is reflexly connected to the salivary response.

It should not be thought that only autonomic, glandular responses are involuntary in this sense. If a small light is always illuminated for five seconds before the delivery of food to a hungry pigeon, the pigeon will learn, by classical conditioning, to approach and peck at the light. Exactly the same experiment as that described above can be undertaken, with food delivered only on those trials when the pigeon does not approach and peck the light during the initial five seconds. The pigeon cannot help doing so. Pavlovian conditioning appears to be a widespread phenomenon, applying to a relatively wide range of responses.

Functions of conditioning. The behaviour of the dog and pigeon in the above experiments seems maladaptive, precisely because it violates the law of effect. If the way to obtain food is to refrain from performing a particular response, then that is what the law of effect says the animal should do. The law of effect makes obvious adaptive sense; several writers, indeed, have pointed to the analogy between the law of effect and natural selection. Just as natural selection favours those variations that happen to increase fitness, so the law of effect selects those responses that happen to be followed by certain consequences.

Adaptive value of Pavlovian conditioning

The fact that Pavlovian conditioning may result in apparently maladaptive behaviour in the artificial confines of the experimental psychologist's laboratory, however, does not mean that it is not adaptive in the real world. The pigeon's behaviour provides a clue. In a normal classical conditioning experiment, where the illumination of a small light regularly precedes the delivery of food, the pigeon will rapidly learn to approach and direct pecks at the light. Approach and pecking are food-related activities: what is happening is that a simple process of Pavlovian conditioning is ensuring that responses related to food are being elicited by stimuli associated with food. It is not difficult to appreciate the adaptive significance of a process that results in animals approaching places where they have found food in the past, or in learning that a particular novel object is in fact an example of food, and directing food-related activity toward these stimuli in the future.

Pavlovian conditioning also affects other significant behaviours. For example, it probably provides the basic process by which animals learn to avoid poisonous foods. If a novel food is associated with illness, its taste will elicit responses of disgust or nausea, ensuring that the substance will subsequently be rejected after the first taste.

In territorial birds and fish, aggressive displays and attacks can become conditioned to stimuli that regularly precede the appearance of a rival male. A male already primed to threaten and attack an intruder, because he has learned that certain signs herald the appearance of the intruder, should be more successful in defense of his territory than the male that is unprepared. Experimental analysis has, in fact, nicely confirmed this expectation. In general, any pattern of defensive behaviour that is adaptive in response to an intruder or predator—such as displaying or fighting, fleeing or taking other evasive action, or freezing into immobility or feigning death—will be even more adaptive if performed in advance, at the first reliable signal of the predator's or intruder's appearance.

Adaptive value of instrumental conditioning

The process of Pavlovian conditioning thus often enables animals to behave appropriately in anticipation of events of biological significance, without involving any direct modification of that behaviour by its success or failure. But further modification must sometimes be of further advantage. For instance, it is not always enough just to approach a stimulus associated with food; if that stimulus is a prey species, it may take evasive action that will require much more elaborate behaviour on the part of the predator. This can be seen in the feeding behaviour of the oystercatchers, a group of birds that eat bivalve mollusks. Oystercatchers first catch their pray by probing down the hole made by the bivalve in the mud; the sight of the hole must be rapidly established as a conditional stimulus for food. But the birds must then perform a complex series of actions to get at the mollusk's flesh, and this skilled sequence of responses also must be learned, presumably in accordance with the law of effect. Similarly, many animals have a wide range of defensive behaviour patterns; in the laboratory, at least, which one eventually predominates in any given situation normally depends on which one successfully enables the animal to escape or to avoid aversive consequences. In all these cases, it appears that instrumental conditioning serves to modify, via the law of effect, initial responses that owed their origin to Pavlovian conditioning.

The adaptive value of instrumental conditioning is an area of research that has seen some fruitful collaboration among experimental psychologists, ethologists, and behavioral ecologists. From ecology has come the "optimal foraging theory," the idea that efficient foraging behaviour should maximize an animal's net rate of food intake. From ethology and experimental psychology has come the idea that an animal's instrumental behaviour in any given situation is a product of competition between various possible activities, a competition whose resolution depends on weighing the costs and benefits of increasing one activity at the expense of another. Both in the laboratory and in more natural settings, for example, the proportion of time spent searching for one kind of food depends not only on the probability of finding that food and on its value when found but also on the probability of the animal finding an alternative food if it looks elsewhere. There is also abundant evidence that animals improve their foraging efficiency with practice; this clearly must depend on learning which stimuli signal the availability of which kinds of food, the most efficient way of taking a given food, and the most effective distribution of time between alternatives.

SPATIAL LEARNING

One of the major problems many animals must confront is how to find their way around their world—for example, to know where a particular resource is and how to get to it from their present location, or what is a safe route home to avoid a predator. Such spatial learning may cover only the highly restricted confines of an animal's home range or territory, or it may embrace a migration route of several hundreds or even thousands of miles. Although some forms of navigational behaviour may be explicable in relatively simple terms, not necessarily requiring appeal to processes more complex than those of simple conditioning, others suggest some quite new principles.

Maze learning. In the psychologist's laboratory, the primary method of studying spatial learning has been to put a rat in a maze and watch how it finds its way to the goal box, where it is fed. As befits the analytic (some would say sterile) approach so popular in experimental psychology, the elaborate and complex mazes used in earlier studies (the very first published experiment used a scaled-down replica of the maze at Hampton Court, London) soon gave way to something very much simpler, a T-maze or Y-maze. A rat placed at the end of one arm must run to the central choice-point, from where it has to enter one of the two remaining arms. Although extremely simple, even this apparatus allows for a number of possible modes of solution. One possibility is that the rat learns to execute a particular response, a left turn or a right turn, at the choice-point, because that response is followed by food. A second possible solution is that the rat learns that the two alternative arms differ in some particular way and further learns to associate one of the arms with food and hence to choose it. The third and most interesting possibility is that the rat learns to define the rewarded arm not in terms of its own intrinsic characteristics but by its spatial relationship to an array of landmarks outside the maze. Thus the rat might learn that the correct arm is the one pointing to the left of a window and away from a table with a lamp on it. Experiments show that whenever such landmarks are available, this third solution mode is the one used.

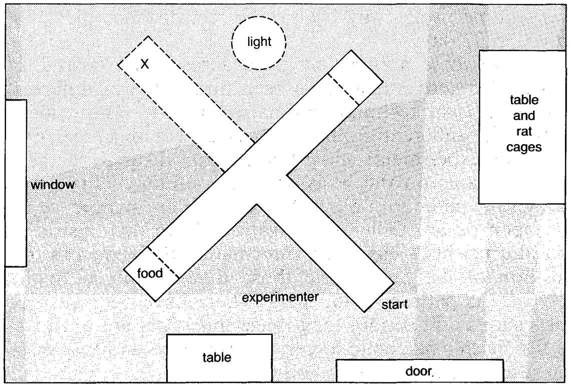


Figure 3: Diagram of a T-maze. Once a rat has learned to run to the arm that contains food, the maze can be flipped (so that the start-arm is in the position marked X) to determine if the rat has learned to reach the goal by making a left turn. If the rat turns right on this trial, the experimenter can rotate the maze 180° to see if the animal has learned to identify the goal-arm in terms of its intrinsic characteristics or, as is usually the case, by its relation to landmarks outside the maze.

Evidence of spatial learning in rats

Perhaps the most convincing demonstration that rats can find their way to a particular location—one defined solely in terms of its spatial relation to various external landmarks—has been provided by experiments in which the animals are placed in a large circular tank of water and must swim to a transparent platform submerged somewhere in the middle of the tank. They can rapidly learn to do this, regardless of where they are initially put into the tank and even though the platform itself is invisible. (The invisibility of the platform is shown by the following: if the platform is moved, the rat will swim straight past it, heading instead toward the position it used to occupy.)

Rats in these experiments are not simply approaching a single landmark; they locate their goal by reference to its spatial relationship with a whole series of landmarks, no one of which is necessary. This can be established by using half a dozen arbitrary but easily identified objects as landmarks during maze training. Removal of any one or two of them in no way disrupts the rat's behaviour. If all the landmarks are systematically rotated around the room, the rat will identify a new arm of the maze as correct (the one that has the same relationship to the landmarks as the initially correct arm). If, however, the landmarks are rearranged in such a way as to destroy their original spatial relationship to one another, the rat does not know which arm to choose.

The processes involved in this sort of learning are not well understood. Some psychologists have been sufficiently impressed by the rat's flexibility in these experiments to argue that the animal is constructing a map of its environment—not, obviously, a written map but an internal, maplike representation that encodes a complete set of spatial relationships between major landmarks. The best evidence for such a maplike representation would be if a rat could take an unfamiliar route when its original route to a goal is blocked. Unfortunately, there is little evidence of such performance in rats, except in the not especially critical case where the goal, or a stimulus very close to it, is clearly visible from the choice-point. On the other hand, studies of long-range navigation have shown that some animals can do just this.

Navigation. Salmon return from the ocean to spawn in the stream in which they were hatched; swallows return to the same nest sites in northern Europe each spring from wintering in southern Africa. These and other examples of large-scale migrations have long fascinated students of animal behaviour, and experimental intervention has produced some remarkable results. A Manx shearwater was taken in an airplane from its breeding site on the island of Skokholm, off south Wales, to Boston, Mass. It returned to Skokholm within 13 days of being released in Boston; the direct distance between these two points is 3,050 miles, which implies (assuming that the bird did not fly at night) a minimum average speed in excess of 20 miles per hour. An albatross flew from a release site in the Philippines to its home in Midway Island, a direct distance of 4,120 miles, in 32 days.

Navigational cues

How do these animals navigate across such great distances? Numerous cues have been implicated in different instances. Near to home, animals probably rely on local cues quite different from those used at a distance. For example, experiments show that salmon distinguish their home streams on the basis of smell, although this sense can hardly come into play while the fish are swimming in the open ocean. Other investigations have demonstrated that diurnal birds use visual information derived from the position of the Sun, while those that migrate at night rely on the pattern of the stars. There have been several suggestions that certain long-range migrators are sensitive to the Earth's magnetic forces; sensitivity to auditory cues has also been suggested in some cases.

The most intensive analysis of long-range navigation has been undertaken with homing pigeons. These birds are trained by being released from sites progressively further from their home loft. Just what the pigeons learn on these training flights is not entirely clear. In part, they obviously learn the visual landmarks immediately surrounding the home loft, but experimental evidence suggests that they use such landmarks only very close to home. Once some training has been given, however, a pigeon can be taken 100 miles or more in any direction from home, and it will, within a few minutes of its release, start flying in a homeward direction.

One general class of theory on homing behaviour postulates that the pigeon detects a discrepancy between a particular set of stimuli observed at the release site and its stored knowledge of what that set of stimuli should be like at home, and it then flies in such a direction as to reduce this discrepancy. Different versions of this theory appeal to different sets of stimuli that might be used to guide the pigeon home. At one time, a popular idea was that the pigeon used the Sun's height in the sky in combination with an internal clock. At any given season and time of day, the Sun's height in the sky—and, by extrapolation from its current rate of climb, its maximum height—are unique to a single place (in this case, the pigeon's home). Assuming that the pigeon's home loft and the release site are both in the Northern Hemisphere, then if the Sun's maximum height is lower at the release site, the release site is north of home; if higher, then the release site is south of home. If the Sun will reach its maximum height later than at home, the release site is west of home; if earlier, the site is east of home. If released at noon at a site in the Northern Hemisphere 200 miles northeast of home, the pigeon must fly so as to raise the maximum height of the Sun (*i.e.,* south), and so as to stop the Sun falling (*i.e.,* west).

This explanation is immensely ingenious and, although calling for some astonishingly fine sensory discriminations on the part of the pigeon, not impossible in principle. Unfortunately, it is probably wrong. Two critical experiments have produced results quite at variance with its predictions. The first suggested that pigeons do not rely on the height of the Sun to navigate at all. In this experiment, the pigeons were confined to a laboratory from which they could see the Sun for only a relatively short time around noon each day, and the apparent height of the Sun above the horizon was raised or lowered by allowing the birds to view the Sun only through a complex series of mirrors. This should have had drastic effects on their perception of the true position of their home; for instance, an increase of 70′ in the apparent height of the Sun at noon would correspond to an 80-mile southward relocation of the home. The pigeons were then taken from home and released 40 miles south, where they saw the real Sun for the first time in several weeks. If the Sun's height was indeed a critical stimulus in navigation, the pigeons would be expected to fly south rather than north. In fact, they correctly flew north.

Experiments on navigation in homing pigeons

The second experiment involved shifting the birds' internal clock, by confining them indoors and exposing them to a new light–dark cycle. Independent observations had shown that this procedure is entirely successful: if a bird is confined indoors for a few weeks with the lights switched on every day at midnight and switched off at noon, its clock soon will be entrained on this new cycle, so that 6:00 AM is regarded as the middle of the day. In the critical experiment, the bird was taken out of the laboratory and released at 6:00 AM (true time) from a site 50 miles south of home. The Sun, now seen for the first time in several weeks, was just rising; but, according to the pigeon's internal clock, the time at home was noon. This implied that the release site was a long way west of home, and if the pigeon were using the height of the Sun as a cue to guide it home, it should have flown east. In fact, the pigeon flew west.

Use of the Sun as a compass

The result of the second experiment indicates that the pigeon was using the position of the Sun in the sky, and that the clock shift had been effective (for the pigeon was not flying in the direction of home). This is readily explained by the hypothesis that the pigeon used the Sun as a compass. If we allow, for the sake of argument, that the pigeon knew that the release site was south of home, then it should have tried to fly north. In the Northern Hemisphere at noon, the Sun is due south; therefore, the pigeon—whose internal clock said it was noon—should fly away from the Sun. But although the pigeon's shifted clock said that it was noon, the true local time was 6:00 AM, and the Sun was in the east. Flying away from the Sun, the pigeon flew west. This experiment then suggests first, that the pigeon was not using the height of the Sun at all; second, that it used the Sun's horizontal position, or azimuth, to provide a compass bearing; and, third, and most important, that the pigeon had some other map that told it that the release site was south of home. In general, a compass is of no use without a map.

The role of experience in establishing an internal map

The basis for the map component of the pigeon's navigational skill remains extremely obscure. There is evidence from studies of many migratory birds that the compass component is in some sense innate, but that a map of the relative positions of the summer and winter habitats and of other places in between (or even not in between) develops only with the experience of migration. For example, starlings that breed around the Baltic Sea fly southwest in autumn to winter in southern England, northern France, and Belgium. When captured during this autumn migration and released in Switzerland (some 500 miles south of their normal route), experienced, adult birds flew back to northern France and Belgium—even though they had presumably never flown over any part of this route before. Young birds, however, for whom this was the first migration, flew southwest from Switzerland and ended up in southern France or northern Spain. They clearly had a compass that told them which direction was southwest; what they lacked was any knowledge of the spatial relationship between their present location in Switzerland and their goal in northern France.

PERCEPTUAL LEARNING

According to Thorndike's stimulus–response theory, learning, which is reducible to the strengthening and weakening of the tendency to perform a particular response in the presence of a particular stimulus, occurs only when that response is performed; learning, in other words, depends on trial and error. Even in the realm of simple conditioning, there are good reasons to question this restriction. Conditioning is better conceptualized as the acquisition of knowledge about temporal relationships between events rather than as the acquisition of behaviour. Spatial learning seems to be a matter of learning about spatial relationships between objects and places in one's environment and, apparently, the construction of some sort of map that will subsequently permit the animal to perform a new sequence of actions across unknown territory. This section considers other examples of learning, in which at least part of what an animal appears to acquire is the recognition of a more or less complex set of stimuli that subsequently can be used to guide its actions.

Imitation and observational learning. One reason why Thorndike adopted such a narrow, behavioral view of learning was that he looked for evidence of other forms of learning without success. Having taught one cat to escape from the puzzle box by operating a latch, he looked to see whether a second cat would acquire the correct solution simply by watching the first. A series of such experiments produced uniformly negative results, and Thorndike concluded that trial and error was the only form of learning available to animals other than humans.

Why Thorndike should have been so unsuccessful is something of a mystery, for later experiments have established quite convincingly that animals can often benefit from watching another member of their species perform a particular task. Casual observation in natural settings, for instance, reveals that young chimpanzees intently watch their elders perform intricate tasks; this certainly suggests that learning by observation is very common in some species.

Experimental analysis has revealed a number of important distinctions concerning the role of observation in behaviour. For example, domestic chickens that have eaten to satiation a particular source of food will start eating again if they observe other chickens feeding. Although the observation of conspecifics engaged in a particular activity has clearly affected the tendency of the satiated chicken to engage in that activity, it is not clear what they might have learned from this observation. They already know how to peck, and they already know that the grain before them is palatable food. It is probably more appropriate to regard this as an instance of "social facilitation" and to say that one of the stimuli that elicits feeding in chickens is the sight of other chickens feeding.

The basic nature of observational learning

The example above demonstrates the minimum requirement for establishing that an animal has learned by observation: in the absence of the opportunity to observe another, the animal must have been unlikely to have performed a particular response, and the reason for this must reside in lack of knowledge. An artificial, laboratory example of observational learning would be to allow an observer rat to watch a demonstrator rat pressing a lever for food. If the observer has never before pressed a lever and, given the opportunity, now does so much more rapidly than another rat denied the opportunity to observe the demonstrator, surely some genuine observational learning has occurred. But even here it remains difficult to establish exactly what it is that the observer has learned by watching the demonstrator, and more elaborate experiments may be required to elucidate this. An experiment with two monkeys showed how this may be done. The monkeys took turns acting as demonstrator and observer. The demonstrator's task was to choose between two objects, one of which contained some hidden food. Since the objects were changed on each new trial for the demonstrator, there was no way for the animal to know which choice was correct, and it necessarily picked one at random. The observer, however, could watch the demonstrator's trial and thus could find out which of the two objects in a particular set was correct. Given an opportunity to choose between the two, the observer more often than not chose correctly. That the observer was not simply watching the demonstrator, but was in fact looking to see the outcome of the choice, is established by the finding that the observer performed somewhat more accurately on those trials when the demonstrator's choice was wrong than on those when it was right.

The distinction between observation and imitation

This last finding points to a further distinction, that between observing the actions of another and imitating those actions. In this particular experiment, the monkeys clearly were not imitating one another, or they would have copied each other's choices even when these were wrong. A demonstration of imitation is provided by the behaviour of oystercatchers feeding on mussels. Having found a mussel, an adult oystercatcher obtains the food from within either by inserting its beak in the right place and cutting the muscle that holds the shell together or by pecking a hole in the weakest point of the shell. Young birds develop the method employed by their parents, but experiments in which chicks were fostered by adults with a different habit from that of the natural parents have established that this behaviour is not genetically determined. Rather, the young birds imitate the actions they observe being performed by their foster parents.

The best known natural example of such imitation was provided by a troop of macaques in Japan. In order to lure the monkeys out of the forest and into the open, where their behaviour could be better studied, scientists routinely left sweet potatoes and wheat on the beach. The monkeys ate this food but clearly disliked the fact that it had become liberally mixed with sand. A young female member of the troop, however, discovered that sweet potatoes could readily be washed free of sand, and that a handful of wheat and sand could be thrown into a pool, where the sand would sink, leaving the wheat floating behind. Both customs spread through the troop, first to the immediate family and young companions of the original inventor, and last of all (an interesting touch) to the old, conservative males. Other examples of observational learning are readily apparent in the behaviour of animals in the field, but in many cases, as in some of the laboratory studies cited above, it remains difficult to elucidate just what it is that has been learned.

Song learning. A special case of observational learning is that of young birds acquiring their species-typical song. Numerous species of animals, including many birds, produce species-typical calls or other vocalizations as adults; in many cases, however, there is little evidence that learning plays any significant role in their development. In many species of crickets, for example, the song is stereotyped, and the pattern of neural activity that produces the song can be detected even in young animals who neither sing nor apparently react to the adult song. But in most songbirds, there is reason to believe that learning has a significant effect on the development of the adult song.

The role of observation

The interesting feature of this learning is that it sometimes occurs in two distinct phases separated by several months. The first of these can be regarded as purely observational learning, the second as the perfection of the song through practice (*i.e.,* as imitation of a model). Song sparrows, for example, do not develop a normal adult song unless they have the opportunity to hear the song during their first autumn. There is thus a sensitive period during which they must hear their species' song if they are to develop normally, but it is important to note that they do not themselves sing at all during the first autumn. It is not until the next spring that they start practicing the song. At this point, they do not need to hear other sparrows singing, but they do need to hear themselves. If the bird is deafened before it starts practicing, only a very crude song emerges. The implication is that, during exposure in the first autumn, the sparrow learns to identify the detailed song and establishes a template of it; the following spring, the sparrow starts singing and needs practice to match its output to the stored template.

The song sparrow provides an example of a particularly clear separation between observation and imitation. In other species, such as the chaffinch, the young bird learns from exposure to song in the first autumn, but refinement of the song is produced by further exposure to other chaffinches singing during the following spring. In yet others, such as indigo buntings, the adult bird learns its song from territorial neighbours. But even where there is no temporal separation between the two aspects of learning, it still seems valid to distinguish between the learning involved in establishing the template and that involved in perfecting the motor skill.

If song learning consists solely of the young bird learning to reproduce the adult, species-typical song, one might wonder why any learning should be necessary at all. Why should the song not develop simply through maturation, or, in other words, why is not the template, at least, genetically laid down in the bird's brain? In fact, studies indicate that a relatively crude template is innately determined in most species. There are very strict limits to the range of songs that a bird of one species can learn. Moreover, among chaffinches and certain other species, even if a young bird hears no song at all it will still develop a crude song that has recognizable features of the full, species-typical one. The degree of this innate specification varies widely from species to species: at one extreme are such birds as cuckoos, which develop a standard call with no prior exposure at all; at the other extreme are such birds as marsh warblers, which develop idiosyncratic songs picked up, it seems, from any other species they come in contact with during the sensitive period.

Value of learned songs

Species whose song acquisition involves a great deal of individual learning are probably those in which individual birds develop slightly different songs. In some species, such as song sparrows, there are recognizable local "dialects" that the young birds learn from adults living in the same region. In other species, there is even more variation between individuals. If one function of the song is to attract a mate, then an interplay is called for between a song that simply advertises the singer's species and one that establishes his individual identity. The importance of individual learning, then, depends on the role of the song in the mating patterns of the species.

Value of imprinting

Imprinting. The young of many species are born relatively helpless: in songbirds, rats, cats, dogs, and primates, the hatchling or newborn infant is wholly dependent on its parents. These are altricial species. In other species, such as domestic fowl, ducks, geese, ungulates, and guinea pigs, the hatchling or newborn is at a more advanced stage of development. These are precocial species, and their young are capable, among other things, of walking independently within a few minutes or hours of birth, and therefore of wandering away from their parents. Since mammals are dependent on their mothers for nourishment, and even birds are still dependent on parental guidance and protection, it is important that the precocial infant not get lost in this way. The phenomenon of filial imprinting ensures that, in normal circumstances, the precocial infant forms an attachment to its mother and never moves too far away.

Although imprinting was first studied by the Englishman Douglas Spalding in the 19th century, Konrad Lorenz is usually, and rightly, credited with having been the first not only to experiment on the phenomenon but also to study its wider implications. Lorenz found that a young duckling or gosling learns to follow the first conspicuous, moving object it sees within the first few days after hatching. In natural circumstances, this object would be the mother bird; but Lorenz discovered that he himself could serve as an adequate substitute, and that a young bird is apparently equally ready to follow a model of another species or a bright red ball. Lorenz also found that such imprinting affected not only the following response of the infant but also many aspects of the young bird's later behaviour, including its sexual preferences as an adult.

Imprinting, like song learning, involves a sensitive period during which the young animal must be exposed to a model, and the learning that occurs at this time may not affect behaviour until some later date. In other words, one can distinguish between a process of perceptual or observational learning, when the young animal is learning to identify the defining characteristics of the other animal or object to which it is exposed, and the way in which this observational learning later affects behaviour. In the case of song learning, observation establishes a template that the bird then learns to match. In the case of imprinting, observation establishes, in Lorenz' phrase, a model of a companion, to which the animal subsequently directs a variety of patterns of social behaviour.

With imprinting, as with song acquisition, one can ask why learning should be necessary at all. Would it not be safer to ensure that the young chick or lamb innately recognized its mother? There are, in fact, genetic constraints on the range of stimuli to which most precocial animals will imprint. A model of a Burmese jungle fowl (the species whose domestication produced domestic chickens) serves as a more effective imprinting object for a young chick than does a red ball; there is even evidence that imprinting in the latter case involves different neural circuits from those involved in imprinting to more natural stimuli. Nonetheless, it is clear that the innate constraints are not very tight and that a great deal of learning normally occurs. The most plausible explanation, as in the case of song learning, is that imprinting involves some measure of individual identification. Lorenz argued that one of the unique characteristics of imprinting was that it involved learning the characteristics of an entire species. It is true that imprinting results in the animal directing its social and mating behaviour toward other members of its own species, and not necessarily toward the particular individuals to which it was exposed when imprinting occurred. But learning usually involves some generalization to other instances, and there does not seem to be anything peculiar to imprinting here. The primary function of imprinting, however, is to enable the young animal to recognize its own mother from among the other adults of its species. This no doubt is particularly important in the case of such animals as sheep, which live in large flocks. Only learning could produce this result.

There is also an important element of individual recognition in at least some cases of imprinting's effects on sexual behaviour. Experiments with Japanese quail have shown that their sexual preferences as adults are influenced by the precise individuals to whom they are exposed at an earlier age. Their preferred mate is one like, but not too like, the individuals on whom they imprinted. The preference for some similarity presumably ensures that they attempt to mate with members of their own species. The preference for some difference is almost certainly a mechanism for reducing inbreeding, since young birds will normally imprint on their own immediate relatives.

The possible role of rewards in imprinting

The difference between imprinting and song learning lies in the consequences of observational learning. The effect of imprinting is the formation of various forms of social attachment. But what mechanism causes the young chick or duckling to follow its mother? Lorenz thought that imprinting was unrewarded, yet the tendency of a young bird to follow an object on which it has been imprinted in the laboratory can be enhanced by rewarding the bird with food. Rewards also occur outside the laboratory: the mother hen not only scratches up food for her young chicks, she also provides a source of warmth and comfort. Moreover, following is also rewarded by a reduction in anxiety. As chicks develop over the first few days of life, they show increasing fear of unfamiliar objects; they allay this anxiety by avoiding novel objects and approaching a familiar one. This latter object must be one to which they have already been exposed—in other words, one on which they have imprinted. Imprinting works because newly hatched birds do not show any fear of unfamiliar objects, perhaps because something can be unfamiliar only by contrast with something else that is familiar. On the contrary, the newly hatched birds are attracted toward salient objects, particularly ones that move. Once, however, a particular object has been established as familiar and its features identified, different objects will be discriminated from it. These will be perceived as relatively unfamiliar, and hence they will provoke anxiety and the attempt to get as close as possible to the more familiar object. The imprinting of the young bird on one object necessarily closes down the possibility of its imprinting on others, as these will always be relatively less familiar. Thus, there is normally a relatively restricted period in the first few hours or days of life during which imprinting can occur. The only way to prolong this period is to confine the newly hatched bird to a dark box where it is exposed to no stimuli; prevented from imprinting during this period of confinement, the bird imprints on the first salient object it sees after emerging.

COMPLEX PROBLEM SOLVING

Experimental psychologists who study conditioning are the intellectual heirs of the traditional associationist philosophers. Both believe that the complexity of the human or animal mind is more apparent than real—that complex ideas are built from simple ideas by associating simple elements into apparently more complex wholes. According to this perspective, the only relationship between these ideas is their association, and the determinants of these associations are themselves relatively simple and few in number. Neither conditioning theorists nor associationist philosophers, however, have lacked for critics, who claim that intelligent problem solving cannot be reduced to mere association. Although allowing that the behaviour of invertebrates, and perhaps that of birds and fish, may be understood in terms of instincts and simple forms of nonassociative and associative learning, these critics maintain that the human mind is an altogether more subtle affair, and that the behaviour of animals more closely related to man—notably apes and monkeys, and perhaps other mammals as well—will share more features in common with human behaviour than with that of earthworms, insects, and mollusks.

Differing views of animal intelligence

The idea that animals might differ in intelligence, with those more closely related to humans sharing more of their intellectual abilities, is commonly traced back to Charles Darwin. This is because the acceptance of Darwin's theory of evolution was at the expense of the ideas of the French philosopher René Descartes, who held that there is a rigid distinction between man, who has a soul and can think and speak rationally, and all other animals, who are mere automatons. The Cartesian view had, in fact, been challenged long before Darwin's time by those who believed (as seems obvious from even the most casual observations) that some animals are notably more complicated than others, in ways that probably include differences in behaviour and intelligence. It was, however, the publication of Darwin's *Descent of Man* (1871) that stimulated scientific interest in the question of mental continuity between man and other animals. Darwin's young colleague, George Romanes, compiled a systematic collection of stories and anecdotes about the behaviour of animals, upon which he built an elaborate theory of the evolution of intelligence. It was largely in reaction to this anecdotal tradition, with its uncritical acceptance of tales of astounding feats by pet cats and dogs, that Thorndike undertook his studies of learning under relatively well-controlled laboratory conditions. Thorndike's own conclusions, already noted above, were distinctly Cartesian: animals ranging from chickens to monkeys all learned in essentially the same way, by trial and error or simple instrumental conditioning. Unlike man, none could reason.

This controversy actually involves two questions, which are worth keeping apart. The first is whether theories of learning based on the results of, say, simple conditioning experiments are sufficient to explain all forms of learning and problem solving in animals. The second question is whether new and more complex processes operate only in some animals, that is to say, whether some animals are more intelligent than others. The distinction between these questions is not always easy to preserve, for they are clearly related, and an answer to one usually has implications for the other. The remainder of this article is organized around the first question; in cases where the behaviour of an animal does, in fact, seem to indicate that more complex processes are involved, the second question is also considered.

Discrimination of relational and abstract stimuli. Laboratory studies of habituation and conditioning usually employ very simple stimuli, such as lights, buzzers, and ticking metronomes in Pavlov's experiments. Some of the other examples of learning considered earlier have already suggested that animals can actually respond to additional, more complex stimuli. Even the solution of simple spatial discriminations in the laboratory requires the animal to learn about spatial relationships between different landmarks; migration or navigation over hundreds of miles demands abilities at least as complex as this. Song learning requires the young bird to discriminate between different sequences of subtly varying notes and calls, and the individual recognition involved in imprinting requires response to elaborate configurations of features.

Transposition of relational discriminations

Thus, one way in which a problem may become more difficult is if its solution depends on response to more subtle changes in stimuli. Numerous laboratory studies have examined the abilities of a variety of animals to perform such discriminations. The phenomenon of transposition, first studied in chicks by the Gestalt psychologist Wolfgang Köhler, suggests that animals may solve even simple discriminations in ways more complex than the experimenter had imagined. Köhler trained his chicks to perform simple discriminations—say, to choose a large white circle (five centimetres in diameter) in preference to a small white circle (three centimetres in diameter). He then sought to discover whether the animal was responding to the relationship between the two stimuli or to the absolute characteristics of the stimuli. In other words, had the chick learned to select the larger of the two circles, or had it learned to pick the five-centimetre circle? If the former were the case, Köhler reasoned that given the choice between the five-centimetre circle and an even larger one (eight centimetres in diameter), the animal should transpose the relationship and choose the larger circle. This was indeed the result, demonstrating that the animal was responding in terms of the relationship between stimuli rather than, or at least in addition to, their absolute properties.

Counting as an example of an abstract discrimination

Transposition experiments show that animals can respond to relationships between stimuli varying along a particular continuum of physical characteristics: size, brightness, hue, etc. Another question is whether animals can respond to an abstract property of a stimulus array, independent of the actual physical stimuli making up that array. In experiments on counting, the animal must choose between an array containing, say, five stimuli and one containing three. The actual stimuli in the array vary from trial to trial, in order to rule out the possibility that the animal is responding in terms of other features, such as differences in total area or brightness, between the arrays. Counting experiments have been tried on birds more frequently than on any other class of animal, and several species, notably ravens, rooks, and jackdaws, have solved this type of problem. This success may not be entirely by chance, for there is reason to believe that the stimulus that controls when a female bird stops laying eggs is something to do with the number of eggs already laid and in the nest. Chimpanzees, however, have been trained to label pictures of various objects (*e.g.*, spoons, shoes, padlocks, and balls) with the numeral specifying the number of objects in the picture. Moreover, rats and other standard laboratory animals have solved similarly abstract discriminations, for example, of temporal duration. A rat can learn to perform one response after a stimulus has been turned on for two seconds and a different response after the stimulus has been turned on for five seconds. The nature of the actual stimuli employed can vary without disrupting the rat's discrimination, suggesting that it is the duration of the stimuli to which the rat responds.

Concept learning

Concept learning makes up another class of discriminations that may be solved by the abstraction of a particular property or set of properties from a very wide array of individual stimuli. In a typical experiment, a pigeon is shown a large number of colour photographs of natural scenes: half of these contain, somewhere within the scene, all or part of a tree or group of trees; the other half contain no tree (although there might be flowers, a climbing rose, or other plants). Responding to the pictures of trees is rewarded, but responding to the remaining pictures is not. Pigeons rapidly learn the discrimination. In one sense, perhaps this is not surprising: birds that roost in trees, one is inclined to argue, must be able to recognize them. But pigeons can learn other discriminations with almost equal facility; for example, they can be trained to distinguish between underwater scenes containing a fish and similar views with no fish present. In such cases, the class of stimuli in question is one for which their evolutionary history can hardly have prepared pigeons. The question, of course, is how the pigeons solve such problems. Are they, in some sense, abstracting a conceptual rule for categorizing the world into classes of stimuli? Or are they responding to what is no doubt a very large number of particular features that differentiate trees or fish from other objects in the world?

Pigeons, in common with most birds, rely more heavily on vision, and certainly have better developed colour vision, than most mammals—with the exception of primates. There is evidence that monkeys can solve the concept discriminations that have been set to pigeons, but there is no evidence that other mammals can. For extensive comparative analysis, therefore, it is necessary to turn to different kinds of tasks. One that has been studied almost to excess is discrimination reversal. In reversal tasks, an animal is first trained on a simple discriminative problem: for example, to choose the left-hand arm of a T-maze, where it is rewarded, rather than the right arm, where it is not. Once the animal has solved the problem, the experimenter reverses the reward assignments, so that the food is now in the right arm rather than the left. Training continues until the animal has learned this reversal, whereupon the assignment of reward is switched back to the left arm. And so on. Rats trained on this series of reversals eventually become extremely adept at the task. Although the initial reversal causes considerable problems, with animals making many more errors than on the original discrimination, after a few more reversals these difficulties vanish. Eventually, rats solve each new reversal in fewer trials than they took to solve the original discrimination, often with no more than a single error.

The learning of discrimination reversals

Similarly efficient performance has been observed in a relatively wide range of mammals. More interesting was the early suggestion that the few species of fish (goldfish, African mouthbreeders, and Paradise fish) trained on similar problems showed no evidence of the increase in efficiency displayed by mammals. The fish would learn the first reversal slowly and laboriously, and the 20th reversal equally slowly. Subsequent experiments have established that this was an unfairly pessimistic assessment, for improvements in experimental techniques have been accompanied by a significant improvement in the fish's performance, a finding that highlights the extreme difficulty of assessing the relative efficiency of widely differing animals on supposedly the same task. Nevertheless, it remains doubtful that goldfish are as adept at reversal tasks as rats are.

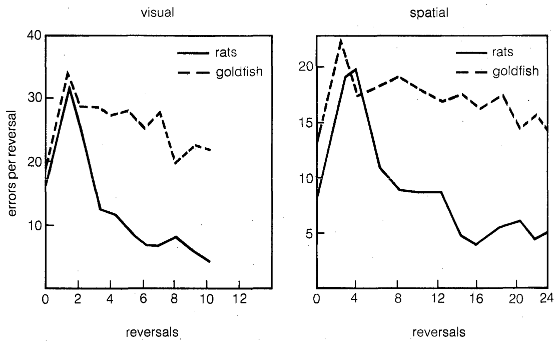


Figure 4: Learning of visual and spatial reversals by rats and goldfish. Although disrupted by the first few reversals of a discrimination, rats soon adjust and learn to reverse their choices with only a few errors. Goldfish also show some adjustment to repeated reversals (they learn later reversals significantly faster than earlier ones), but they do not seem to become as proficient as rats.

The theoretical question, however, is how rats attain such efficiency. What processes allow them eventually to learn the reversal of a discrimination faster than they originally learned the discrimination itself, and often with only a single error? The most plausible suggestion is that they develop a "win-stay, lose-shift" strategy. They learn, in other words, to characterize the alternatives between which they must choose not in terms of their physical features but in terms of whether or not they chose it on the previous trial. They then learn that, if the alternative they chose on the last trial was rewarded, choice of that alternative will be rewarded again on the current trial; while, if it was not, choice of the other alternative will now be rewarded. A variety of other experiments have shown that rats can rapidly learn to use the outcome of one trial to predict the outcome of the next, and hence keep track of regular sequential dependencies in the availability of food or other rewards.

The theoretical basis of reversal learning

Generalized rule learning. Second only to the reversal task in popularity as a tool for the comparative analysis of learning has been the learning set task. The latter is designed to measure the animal's ability "to learn to learn"—in other words, to discover whether after having learned a new behaviour the animal can then more readily learn other related behaviours. For example, an animal is trained on a simple discrimination between two objects, A and B. Once the problem has been solved, the experimenter substitutes a new pair of objects, C and D, for the original pair; when the animal has solved this new problem, yet another new pair, E and F, is substituted, and so on. Rhesus monkeys trained on such a series of problems become progressively more efficient at solving each new problem. Like rats trained on reversal tasks, the monkeys eventually solve each new problem after a single trial, choosing at random on the first trial with each new pair of stimuli but thereafter selecting with essentially perfect accuracy.

Learning set tasks

Performance on learning sets, as on reversals, was once thought to discriminate between more intelligent and less intelligent animals. Apes and rhesus monkeys were extremely efficient at such tasks, more so even than New World monkeys, who were, in turn, more efficient than any nonprimate mammals. Again, however, there are grave difficulties in the way of making valid comparisons. Primates have better developed visual systems than most other mammals, so it is not surprising that they should be better at solving a series of visual discrimination problems. Even the difference in performance between rhesus and cebus monkeys (Old World versus New World monkeys) turns out to be attributable to differences in colour vision more than anything else. Rats appear to solve learning set tasks very efficiently if olfactory stimuli are used.

Nevertheless, there may be important intellectual differences also underlying the differences in performance. One reason for thinking so arises from consideration of the processes probably involved in mastering learning sets. The win-stay, lose-shift strategy that explains the progressive improvement in reversal learning can also explain the same improvement in the learning set task—but only if the animal can generalize the strategy to novel stimuli. Successful performance requires that the animal learn that the alternative chosen on the last trial, and the outcome of that choice, predict which alternative will be rewarded on this trial, whatever the nature of the alternatives. Some evidence suggests that primates can generalize rules of this sort more readily than many other animals can. Monkeys trained on a series of reversals of a single discrimination will learn the reversal of any new discrimination with equal facility. By contrast, cats trained on comparable problems show little evidence of such transfer.

Matching-to-sample discriminations

A discriminative problem widely used in the study of transfer is the "matching-to-sample" discrimination. A pigeon, for example, is required to choose between two disks, one illuminated with red light and the other with green light. The correct alternative on any one trial depends on the value of a sample stimulus, which is also part of each trial. If this third light is red, then the red disk is correct; if green, then green is correct. The correct alternative is the one that matches the sample. Although naturally more difficult than the simple red–green discrimination, matching-to-sample discriminations are learned readily enough by a wide variety of animals; however, there appear to be differences among animals in their capabilities to transfer this learning to a new set of stimuli. Primates and dolphins have shown good evidence of such transfer, but pigeons have shown at best only limited transfer. If pigeons are trained with two or three colours to the point where they are responding with essentially no errors, a substitution of a new colour for one of the trained colours may result in a complete breakdown in the discrimination; there is even some question as to whether they can learn a new matching-to-sample discrimination with new stimuli any faster than pigeons with no prior experience of matching problems.

The abilities to respond in terms of certain relationships between stimuli, to abstract those relationships and invariant features from a complex and changing array of stimuli, and, above all perhaps, to transfer such learning to a completely novel set of physical stimuli seem to be some of the more important processes underlying the solution of complex discriminative problems. The fact that certain evidence suggests that animals may differ in some of these abilities has implications for studies of other forms of problem solving.

Köhler's examples of insight learning

Insight and reasoning. Köhler's best known contribution to animal psychology arose from his studies of problem solving in a group of captive chimpanzees. Like other Gestalt psychologists, Köhler was strongly opposed to associationist interpretations of psychological phenomena, and he argued that Thorndike's analysis of problem solving in terms of associations between stimuli and responses was wholly inadequate. The task he set his chimpanzees was usually one of obtaining a banana that was hanging from the ceiling of their cage or lying out of reach outside the cage. After much fruitless endeavour, the chimpanzees would apparently give up and sit quietly in a corner, but some minutes later they might jump up and solve the problem in an apparently novel manner—for example, by using a bamboo pole to rake in the banana from outside or, if one pole was not long enough, by fitting one pole into another to form a longer rake. Other chimpanzees reached the banana hanging from the ceiling by using a wooden box, or a series of boxes stacked precariously on top of one another, as a makeshift ladder.

Criticisms of Köhler's interpretation

Köhler believed that his chimpanzees had shown insight into the nature of the problem and the means necessary to solve it. According to Köhler's interpretation, the solution depended on a perceptual reorganization of the chimpanzee's world—seeing a pole as a rake, or a series of boxes as a ladder—rather than on forming any new associations. But subsequent experimental analysis has cast some doubts on Köhler's claims. The critical observation is that the sorts of solutions that Köhler took as evidence of insight quite clearly depend on relevant prior experience. Chimpanzees will not fit two poles together to form a rake or stack boxes up to form a ladder unless they have had a great deal of prior experience with those objects. This experience may well occur during play, when the young chimpanzee discovers that using a stick can extend the reach of an arm, or that standing on a box can put one within reach of high objects. Thus, what Köhler was studying, without knowing it, was probably the transfer of earlier instrumental conditioning to new situations. As we have already seen, the ability to transfer an old solution to a new stimulus situation is an important one, relevant to a wide range of problem-solving activities. This ability is not at all well understood, but it will not necessarily be greatly illuminated by describing it as insight. Certainly it is not a process unique to the great apes: if the component tasks are sufficiently well-structured, even pigeons can put together two independently learned patterns of behaviour to solve a novel problem.

Combining information from separate sources to reach a new conclusion is one form of reasoning. The paradigm case of reasoning is the solution of syllogisms; for example, when we conclude that Socrates is mortal given the two separate premises that Socrates is a man and that all men are mortal. Employing transitive inference, we can use the premises that Adam is taller than Bertram and that Bertram is taller than Charles to conclude that Adam must be taller than Charles. Reasoning has often been regarded as a uniquely human faculty, one of the few factors, along with the possession of language, that distinguishes us from the rest of the animal kingdom.

Examples of complex reasoning in nonhuman primates

But are humans the only animals that can reason? The unsatisfying answer must be that it depends on what is meant by reasoning. In a very general sense, most animals appear perfectly able to arrive at a conclusion based on combining information obtained on two separate occasions. A formal demonstration is provided by an experiment on instrumental conditioning discussed earlier. If rats learn that pressing a lever provides sucrose pellets and later learn that eating sucrose pellets makes them ill, they will subsequently put these two pieces of information together and refrain from pressing the lever. Monkeys and chimpanzees, however, have been trained to solve problems that appear more similar to transitive inference. They are first given discriminative training between pairs of coloured boxes, called, for example, A, B, C, D, E. Confronted with the choice between A and B, they learn that choice of A is rewarded and B is not. When B and C are the alternatives, they learn that B is correct; when C and D are the alternatives, C is correct; and so on. Although choice of A is always rewarded, and that of E never is, the remaining three boxes each are associated equally often with reward and with nonreward. Nonetheless, given a choice between B and D on a test trial, the animals choose B.

Syllogistic and transitive inference are not the only forms of reasoning: humans also reason inductively or by analogy. Indeed, analogical reasoning problems (black is to white as night is to __?) form a staple ingredient of some IQ tests. One chimpanzee, a mature female called Sarah, was tested by David Premack and his colleagues on a series of analogical reasoning tasks. Sarah previously had been extensively trained in solving matching-to-sample discriminations, to the point where she could use two plastic tokens, one meaning *same,* which she would place between any two objects that were the same, and another meaning *different*, which she would place between two different objects. For her analogical reasoning tasks, Sarah was shown four objects grouped into two pairs, with each pair symmetrically placed on either side of an empty space. If the relationship between the paired objects on the left was the same as the relationship between those on the right, her task was to place the *same* token in the space between the two pairs. Thus in one series of geometrical analogies, a simple problem would display a blue circle and a red circle on the left and a blue triangle and a red triangle on the right; the correct answer, of course, was *same.* But Sarah was equally correct on more complex problems, even when the relationships in question were functional rather than simply perceptual. For example, she correctly answered *same* when the two objects on the left were a tin can and a can opener and the two on the right a padlock and a key.

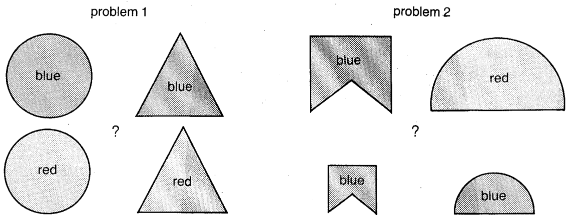


Figure 5: Example of problems in analogical reasoning solved by the chimpanzee Sarah. In each problem her task was to place the token for *same* or *different* between the two pairs of diagrams, depending on whether the relationship between those on the left was the same as, or different from, that between the pair on the right. The correct answer to the first problem is *same;* to the second, *different.*

Solution of analogies requires one to see that the relationship between one pair of items (whether they are words, diagrams, pictures, or objects) is the same as the relationship between a different pair of items. If simple matching-to-sample requires animals to see that one comparison stimulus is the same as the sample and another is different, solving analogies requires them to match relationships between stimuli. The difficulties encountered in training pigeons to generalize simple matching-to-sample discriminations does not encourage one to believe that they would find analogies very easy.

Language learning. The ability to speak was regarded by Descartes as the single most important distinction between humans and other animals, and many modern linguists, most notably Noam Chomsky, have agreed that language is a uniquely human characteristic. Once again, of course, there are problems of definition. Animals of many species undoubtedly communicate with one another. Honeybees communicate the direction and distance of a new source of nectar; a male songbird informs rival males of the location of his territory's boundaries and lets females know of the presence of a territory-owning potential mate; vervet monkeys give different calls to signal to other members of the troop the presence of a snake, a leopard, or a bird of prey. None of these naturally occurring examples of communication, however, contains all of the most salient features of human language. In human language, the relationship between a word and its referent is a purely arbitrary and conventional one, which must be learned by anyone wishing to speak that language; many words, of course, have no obvious referent at all. Moreover, language can be used flexibly and innovatively to talk about situations that have never yet arisen in the speaker's experience—or indeed, about situations that never could arise. Finally, the same words in a different order may mean something quite different, and the rules of syntax that dictate this change of meaning are general ones applying to an indefinite number of other sequences of words in the language.

Defining characteristics of human language

During the first half of the 20th century, several psychologists bravely attempted to teach human language to chimpanzees. They were uniformly unsuccessful, and it is now known that the structure of the ape's vocal tract differs in critical ways from that of a human, thus dooming these attempts to failure. Since then, however, several groups of investigators have employed the idea of teaching a nonvocal language to apes. Some have used a gestural sign language widely used by the deaf to communicate with one another; others have used plastic tokens that stand for words; still others have taught chimpanzees to press symbols on a keyboard. All have had significant success, and several apes have acquired what appears to be a vocabulary of several dozen, and in some cases 100 or 200, "words."

Language learning experiments with apes

Washoe, a female chimpanzee trained by Beatrice and Allan Gardner, learned to use well over 150 signs. Some apparently were used as nouns, standing for people and objects in her daily life, such as the names of her trainers, various kinds of food and drink, clothes, dolls, etc. Others she used as requests, such as *please, hurry,* and *more;* and yet others as verbs, such as *come, go, tickle,* and so on. Sarah, the chimpanzee trained by Premack to use plastic tokens as words, also apparently learned to use tokens for nouns, verbs (*give, take, put*), adjectives (*red, round, large*), and prepositions (*in, under*). But do these signs or tokens really function as words? Does the ape using them, or obeying instructions from a trainer who uses them, really understand their meaning? Or is the ape simply performing various arbitrary instrumental responses in the presence of particular stimuli because she had previously been rewarded for doing so?

There can be little doubt that chimpanzees do have some understanding of what their "words" refer to. Sarah responded appropriately with her token for *red* if asked the question "What colour of apple?" both when an actual red apple was shown as part of the question and when only the token for an apple (which happened to be a blue triangle) was presented. To Sarah, the blue triangle surely stood for, or was associated with, the red apple. In another study, after two chimpanzees had been taught the meaning of a number of symbols for different kinds of food and different tools, they were able not only to fetch the appropriate but absent object when requested to do so, but they could also sort the symbols into two groups, one for foods and one for tools. In another series of studies, a pygmy chimpanzee named Kanzi demonstrated remarkable linguistic abilities. Unlike other apes, he learned to communicate using keyboard symbols without undergoing long training sessions involving food rewards. Even more impressive, he demonstrated an understanding of spoken English words under rigorous testing conditions in which gestural clues from his trainers were eliminated.

As noted above, human language is more than a large number of unrelated words: in accordance with certain implicitly understood syntactic rules, humans combine words to form sentences that communicate a more or less complex meaning to a listener. Can apes understand or use sentences? Undoubtedly they can put together several gestures or tokens in a row. A chimpanzee named Lana, who was trained to press symbols on a keyboard, could type out "Please machine give Lana drink"; Washoe and other chimpanzees trained in gestural sign language frequently produced strings of gestures such as "You me go out," "Roger tickle Washoe," and so on. Skeptical critics, however, have raised doubts about the significance of these strings of signs and symbols. They have pointed out, for example, that when Lana pressed a series of coloured symbols on her keyboard, it was humans who interpreted her actions as the production of a sentence meaning "Please machine give Lana drink." Might it not be equally reasonable to say that she learned to perform an arbitrary sequence of responses in order to obtain a drink? Pigeons can be trained to press four coloured keys—red, white, yellow, and green—in a particular order to obtain food. Psychologists do not feel any temptation to interpret this behaviour as the production of a sentence. What is it about Lana's behaviour that requires this richer interpretation?

Critical interpretations of language experiments

In the case of apes trained to use sign language, two other doubts have been raised. First, there is some reason to believe that a disappointingly high proportion of the apes' gestures may be direct imitations of gestures recently executed by their trainers. Second, a sequence of gestures interpreted as a single sentence is often just as readily interpreted as a number of independent gestures, each prompted, in turn, by a gesture from the trainer. Both these conclusions are based on careful examinations of video recordings of interactions between trainers and apes. Whether they will turn out to be generally true remains an open, and heatedly debated, question.

Without any explicit training, apes have nevertheless learned to produce strings of two or three signs in certain preferred orders: "more drink" or "give me," for example, rather than "drink more" or "me give." Do the animals understand that a string of signs in one order means something different from the same signs in a different order? The following anecdote is suggestive. A chimpanzee called Lucy was accustomed to instructing her trainer, Roger Fouts, by gesturing "Roger tickle Lucy." One day, instead of complying with this request, Fouts signed back "No, Lucy tickle Roger." Although at first nonplussed, after several similar exchanges Lucy eventually did as asked. A simple instance of this sort proves little or nothing, but it may suggest what is needed—namely, that Lucy should understand that changing the order of a set of signs alters their meaning in certain predictable ways. She must generalize the rule that the relationship between the meanings of the signs A-B-C and C-B-A (the same signs in reverse order) is similar to the relationship between the meanings of certain other triplets of signs in her vocabulary when their order is reversed.

Conflicts in approach to the study of animal learning

The research on language in apes forcefully illustrates a conflict, or tension, that is common to many other areas of research on learning in animals. If the investigators are interested in language and communication, they can attempt to communicate as naturally and informally as possible with their apes. This approach involves treating an ape as a fellow social being, with whom one plays and interacts as far as possible as one would with a human child; it also, almost inevitably, results in a style of research where it is exceptionally difficult to control precisely the cues that the ape may be using and even hard to avoid an overly rich, anthropomorphic interpretation of the ape's behaviour. If, on the other hand, the researchers are interested in rigorous experimental control and economical interpretation of the processes underlying the ape's performances, they are likely to set the ape formal problems to solve, with rewards for correct responses and no rewards for errors. But such an approach, however scientific it may seem, must run the risk of missing the point. This is not language; the investigators are not communicating with the ape in the way they would communicate with a child. The very nature of the experimental problems ensures that the ape will not use its language in the way that a child does: to communicate shared interests, to attract a parent's attention to what the child has seen or is doing, to comment on a matter of concern to both.

There is no resolution to this conflict, for both approaches have their virtues as well as their dangers, and both are therefore necessary. In just the same way, the study of a rat pressing a lever in a Skinner box or of a dog salivating to the ticking of a metronome seems to many critics a sterile and narrow approach to animal learning—one that simply misses the point that, if the ability to learn or profit from experience has evolved by natural selection, it must have done so in particular settings or environments because it paid the learner to learn something. It would be foolish to deny this obvious truism: of course it pays animals to learn. Indeed, it may pay them to learn quite particular things in specific situations, and different groups of animals may be particularly adapted to learning rather different things in similar situations. None of this should be forgotten, and the study of such questions requires the scientist to forsake the laboratory for the real world, where animals live and struggle to survive. But few sciences can afford to miss the opportunity to manipulate and experiment under laboratory conditions where this is possible, and none can afford to forget the benefits of precise observation under controlled conditions.

BIBLIOGRAPHY

General works: Historical background is provided by ROBERT BOAKES, *From Darwin to Behaviourism: Psychology and the Minds of Animals* (1984). See also T.R. HALLIDAY and P.J.B. SLATER (eds.), *Animal Behaviour,* vol. 3: *Genes, Development, and Learning* (1983); ROBERT A. HINDE, *Animal Behaviour: A Synthesis of Ethology and Comparative Psychology,* 2nd ed. (1970); J.E.R. STADDON, *Adaptive Behavior and Learning* (1983); and DAVID MCFARLAND, *Animal Behavior: Psychology, Ethology, and Evolution* (1985).

Simple nonassociative learning: GABRIEL HORN and ROBERT A. HINDE (eds.), *Short-Term Changes in Neural Activity and Behaviour* (1970); and HARMAN V.S. PEEKE and MICHAEL J. HERZ (eds.), *Habituation,* 2 vol. (1973).

Associative learning and conditioning: ANTHONY DICKINSON, *Contemporary Animal Learning Theory* (1980); MICHAEL DOMJAN and BARBARA BURKHARD, *The Principles of Learning and Behavior,* 2nd ed. (1986); N.J. MACKINTOSH, *The Psychology of Animal Learning* (1974), and *Conditioning and Associative Learning* (1983); and BARRY SCHWARTZ, *Psychology of Learning and Behavior,* 2nd ed. (1984).

Biological functions of and constraints on learning: ROBERT A. HINDE and J. STEVENSON-HINDE (eds.), *Constraints on Learning: Limitations and Predispositions* (1973); TIMOTHY D. JOHNSTON, "Contrasting Approaches to a Theory of Learning," *Behavioral and Brain Sciences,* 4(1):125–173 (March 1981); and J.R. KREBS and N.B. DAVIES, *An Introduction to Behavioural Ecology* (1981).

Physiological basis of learning: DANIEL L. ALKON and JOSEPH FARLEY (eds.), *Primary Neural Substrates of Learning and Behavioral Change* (1984); and ERIC R. KANDEL, *Cellular Basis of Behavior: An Introduction to Behavioral Neurobiology* (1976).

Spatial learning and navigation: P. ROBIN BAKER, *Bird Navigation: The Solution of a Mystery?* (1984); JOHN O'KEEFE and LYNN NADEL, *The Hippocampus as a Cognitive Map* (1978); and K. SCHMIDT-KOENIG and W.T. KEETON (eds.), *Animal Migration, Navigation, and Homing* (1978).

Song learning and imprinting: P.P.G. BATESON, "The Imprinting of Birds," in S.A. BARNETT (ed.), *Ethology and Development* (1973); HOWARD S. HOFFMAN, "Experimental Analysis of Imprinting and Its Behavioral Effects," *The Psychology of Learning and Motivation,* 12:1–39 (1978); DONALD E. KROODSMA, "Aspects of Learning in the Development of Bird Song," in GORDON M. BURGHARDT and MARC BEKOFF (eds.), *The Development of Behavior: Comparative and Evolutionary Aspects* (1978); and P.J.B. SLATER, "Bird Song Learning: Theme and Variations," in ALAN H. BRUSH and GEORGE A. CLARK, JR. (eds.), *Perspectives in Ornithology* (1983).

Comparative psychology and complex training: E.M. MACPHAIL, *Brain and Intelligence in Vertebrates* (1982); R.E. PASSINGHAM, *The Human Primate* (1982); and H.L. ROITBLAT, T.G. BEVER, and H.S. TERRACE (eds.), *Animal Cognition* (1985).

(N.J.M.)

Human Learning and Cognition

A common goal in defining any psychological concept is a statement that corresponds to common usage. Acceptance of that aim, however, entails some peril. It implicitly assumes that common language categorizes in scientifically meaningful ways; that the word learning, for example, corresponds to a definite psychological process. However, there appears to be good reason to doubt the validity of this assumption. The phenomena of learning are so varied and diverse that their inclusion in a single category may not be warranted.

Recognizing this danger (and the corollary that no definition of learning is likely to be totally satisfactory) a definition proposed in 1961 by G.A. Kimble may be considered representative: Learning is a relatively permanent change in a behavioral potentiality that occurs as a result of reinforced practice. Although the definition is useful, it still leaves problems.

The definition may be helpful by indicating that the change need not be an improvement; addictions and prejudices are learned as well as high-level skills and useful knowledge.

The phrase relatively permanent serves to exclude temporary behavioral changes that may depend on such factors as fatigue, the effects of drugs, or alterations in motives.

The word potentiality covers effects that do not appear at once; one might learn about tourniquets by reading a first-aid manual and put the information to use later.

To say that learning occurs as a result of practice excludes the effects of physiological development, aging, and brain damage.

The stipulation that practice must be reinforced serves to distinguish learning from the opposed loss of unreinforced habits. Reinforcement objectively refers to any condition—often reward or punishment—that may promote learning.

However, the definition raises difficulties. How permanent is relatively permanent? Suppose one looks up an address, writes it on an envelope, but five minutes later has to look it up again to be sure it is correct. Does this qualify as relatively permanent? While commonly accepted as learning, it seems to violate the definition.

What exactly is the result that occurs with practice? Is it a change in the nervous system? Is it a matter of providing stimuli that can evoke responses they previously would not? Does it mean developing associations, gaining insights, or gaining new perspective?

Such questions serve to distinguish Kimble's descriptive definition from theoretical attempts to define learning by identifying the nature of its underlying process. These may be neurophysiological, perceptual, or associationistic; they begin to delineate theoretical issues and to identify the bases for and manifestations of learning. (The processes of perceptual learning are treated in the article PERCEPTION.)

This article is divided into the following sections:

Theories of learning

THE RANGE OF PHENOMENA CALLED LEARNING

Even the simplest animals display such primitive forms of adaptive activity as habituation, the elimination of practiced responses. For example, a paramecium can learn to escape from a narrow glass tube to get to food. Learning in this case consists of the elimination (habituation) of unnecessary movements. Habituation also has been demonstrated for mammals in which control normally exercised by higher (brain) centres has been impaired by severing the spinal cord. For example, repeated application of electric shock to the paw of a cat so treated leads to habituation of the reflex withdrawal reaction. Whether single-celled animals or cats that function only through the spinal cord are capable of higher forms of learning is a matter of controversy. Sporadic reports that conditioned responses may be possible among such animals have been sharply debated.

At higher evolutionary levels the range of phenomena called learning is more extensive. Many mammalian species display the following varieties of learning.

Classical conditioning. This is the form of learning studied by Ivan Petrovich Pavlov (1849–1936). Some neutral stimulus, such as a bell, is presented just before delivery of some effective stimulus (say, food or acid placed in the mouth of a dog). A response such as salivation, originally evoked only by the effective stimulus, eventually appears when the initially neutral stimulus is presented. The response is said to have become conditioned. Classical conditioning seems easiest to establish for involuntary reactions mediated by the autonomic nervous system.

Conditioning of involuntary reactions

Instrumental conditioning. This indicates learning to obtain reward or to avoid punishment. Laboratory examples of such conditioning among small mammals or birds are common. Rats or pigeons may be taught to press levers for food; they also learn to avoid or terminate electric shock.

Chaining. In the form of learning called chaining the subject is required to make a series of responses in a definite order. For example, a sequence of correct turns in a maze is to be mastered, or a list of words is to be learned in specific sequence.

Acquisition of skill. Within limits, laboratory animals can be taught to regulate the force with which they press a lever or to control the speed at which they run down an alley. Such skills are learned when a reward is made contingent on quantitatively constrained performance. Among human learners complex, precise skills (*e.g.,* tying shoelaces) are routine.

Discrimination learning. In discrimination learning the subject is reinforced to respond only to selected sensory characteristics of stimuli. Discriminations that can be established in this way may be quite subtle. Pigeons, for example, can learn to discriminate differences in colours that are indistinguishable to human beings without the use of special devices.

Concept formation. An organism is said to have learned a concept when it responds uniquely to all objects or events in a given logical class as distinct from other classes. Even geese can master such concepts as roundness and triangularity; after training, they can respond appropriately to round or triangular figures they have never seen before.

Principle learning. A subject may be shown sets of three figures (say, two round and one triangular; next, two square and one round, and so on). With proper rewards, the subject may learn to distinguish any "odd" member of any set from those that are similar. Animals as low in the evolutionary scale as the pigeon can master the principle of this so-called oddity problem.

The oddity problem

Problem solving. Examples of human problem solving are familiar: finding the roots of a quadratic equation, solving a mechanical puzzle, and navigating by the stars. Among other animals, chimpanzees have been observed to solve problems requiring toolmaking.

This list only samples from the remarkable array of animal activities categorized as learning. Beginning with habituation, they range from the simple adjustments of single-celled animals up to the highest intellectual accomplishments of mankind. It would be wonderful indeed if a single theory of learning were enough to account for all this diversity. So far, however, no theory of learning adequately covers more than a small fraction of these phenomena.

THE STATE OF LEARNING THEORIES

Yet, at the start of the 20th century, vast psychological systems, such as behaviourism and Gestalt psychology, indeed were offered as explanations of learning (and of much wider ranges of behaviour as well). And as late as the 1940s, comprehensive theories of learning were still believed to be reasonably near at hand. But during the next three decades it grew clear that such theories are tenable only for very limited sets of data. By the late 20th century learning theory seemed to consist of a set of hypotheses of limited applicability.

Important earlier theorists. Beginning in the 1930s a number of general theories were advanced in attempts to organize most or all of the psychology of learning. The most influential of the contributing theorists are noted below.

E.R. Guthrie (1886–1959) wrote that learning requires only that a response be made in a changing situation. Any response was held to be linked specifically to the situation in which it was learned. Guthrie argued that learning is complete in one trial, that the most recent response in a situation is the one that is learned, and that responses (rather than perceptions or psychological states) provide the raw materials for the learning process.

For E.C. Tolman (1886–1959) the essence of learning was the acquisition by the organism of a set of what he called Sign-Gestalt-Expectations. These referred to propositions said to be made by the learner that his own specific response to given signs (or stimuli) would result in such and such circumstances later on. Tolman seemed to be saying that what the learner acquires is a specific knowledge of "what leads to what." In brief, his theory was that the learner develops expectations based on experience and that learning depends entirely on successions of events. Although less vocal on the point than others, Tolman implied that learning was a gradual process.

The theory offered by Clark L. Hull (1884–1952), over the period between 1929 and his death, was the most detailed and complex of the great theories of learning. The basic concept for Hull was "habit strength," which was said to develop as a function of practice. Habits were depicted as stimulus-response connections based on reward. According to Hull, responses (rather than perceptions or expectancies) participate in habit formation, the process is gradual, and reward is an essential condition.

Comparison of these theories yields major questions for empirical investigation. Is learning continuous or discontinuous; is it a gradual or sudden (one-trial) process? Is learning a matter of establishing stimulus–response (S–R) connections or does it depend on the learner's understanding of perceptual relationships? Is reward necessary for learning?

Are theories of learning necessary? Such major investigators of learning as B.F. Skinner and J.A. McGeoch maintained in the 1930s and 1940s that preoccupation with theory was misguided. For them the approach simply was to discover the conditions that produce and control learned behaviour. Beyond this, their interests diverged. Skinner studied instrumental conditioning (operant conditioning, as he called it) among rats; McGeoch specialized in human rote memory. Although study of rote verbal learning had become heavily theoretical by the 1970s, Skinner and his associates stuck to their empirical guns, guiding a variety of programs for the practical control of behaviour. Teaching machines and computer-aided instruction, behaviour modification (*e.g.,* the use of tokens to reward desired behaviour among psychiatric patients), and planned utopian societies (Walden II) all found scientific origins in Skinner's rejection of theory in favour of direct efforts to produce results.

Intervening variables and hypothetical constructs. Learning is a concept and not a thing, and the activity called learning is inferred only through behavioral symptoms. The distinction implicit here between behaviour and inferred process is one of Tolman's major contributions and serves to reconcile influential views that might seem completely at odds. Classical behaviourism, as developed by John B. Watson (1878–1958), rejected every mentalistic account and sought to limit analysis to such physiological mechanisms as reflexes. Watson argued that these are objective in a way that so-called thoughts, hopes, expectancies, and images cannot be. The opposing view holds that experiential (introspective) activity (exactly what Watson sought to dismiss) does require discussion.

Watson's reflex theory

Tolman called himself a behaviourist and ostensibly was bound by Watson's insistence on objectivity. But he also was interested in thinking, expectancy, and consciousness. Tolman found his solution to this problem of incompatible theories after his association with the Vienna Circle of Logical Positivists, whose deterministic teachings he brought to the attention of U.S. psychologists about 1920. He maintained that learning is inexorably produced (determined) by such independent (directly manipulable) variables as the organism's previous training and physiological condition and by the response the environment requires. According to Tolman, the development of learning is revealed through the changing probability that given behaviour (the dependent variable) will result. He held that learning itself is not directly observable; it is an intervening variable, one that is *inferred* as a connecting process between antecedent (independent) variables and consequent (dependent) behaviour.

An attractive possibility is that intervening variables may have discoverable physiological bases. Psychologists Paul E. Meehl and Kenneth MacCorquodale proposed a distinction between the abstractions advocated by some and the physiological mechanisms sought by others. Meehl and MacCorquodale recommended using the term *intervening variable* for the abstraction and *hypothetical construct* for the physiological foundation. To illustrate: Hull treated habit strength as an intervening variable, defining it as an abstract mathematical function of the number of times a given response is rewarded. By contrast, Edward L. Thorndike (1874–1949) handled learning as a hypothetical construct, positing a physiological mechanism: improved conduction of nerve impulses.

Intervening variables and hypothetical constructs need not be incompatible; Thorndike's hypothetical neural process could empirically be found to be the mechanism through which Hull's abstraction operates.

Miniature theories. With growing realization of the complexity of learning, the grand theories of Guthrie, Hull, and Tolman generally have been abandoned except as historic landmarks. Hope for any impending, comprehensive theory was almost dead in the 1970s. More modest miniature theories remain, many likely to be of temporary value. An account of their major themes and issues, however, should have more enduring interest.

MAJOR THEMES AND ISSUES

Association. A dominant ancient theme in theories of learning has been that of association. Although the concept was accepted by Aristotle, it was brought into the developing psychology of learning by British empiricist philosophers (Locke, Berkeley, Hume, the Mills, and Hartley) during the 17th, 18th, and 19th centuries. Popular acceptability of the notion of association was related to progress in the physical sciences. The physical universe had been shown to consist of a limited number of chemical elements that can combine in innumerable ways. By analogy, a science of "mental chemistry" seemed appealing. The theorized elements in this new "science" were called ideas, said to be based on what were named sensations. The synthesizing principles by which these posited ideas combined in conscious experience were expressed as so-called laws of association. It was suggested that such conditions as temporal and spatial contiguity, repetition, similarity, and vividness favoured the formation of associations, and each was called a law of association. Thus, there were "laws" of repetition, of similarity, and so on.

Rote learning

At the end of the 19th century the notion of association was widely accepted among psychologists. German psychologist Wilhelm Wundt (1832–1920) took a position nearly identical with that of the British empiricist philosophers. Also in Germany, Hermann Ebbinghaus (1850–1909) began to study rote learning of lists of nonsense verbal items (*e.g., XOQ, ZUN, ZIB*). He maintained that the association of each word with every succeeding word was the primary mechanism in learning these lists. Pavlov in Russia offered temporary associative connections in the nervous system as a hypothetical basis for conditioned reflexes.

These European influences coalesced in North America. Wundt's notions were introduced there when a student of his from England, Edward Bradford Titchener (1867–1927), came to teach at Cornell University in Ithaca, New York. Ebbinghaus' method and theory became standard in Canadian and U.S. studies of verbal learning; Watson and other behaviourists applied Pavlov's conceptions to their learning experiments. Experimental psychology in the Western Hemisphere came to be dominated by what seemed to be a search for laws of association.

What is associated? Investigators asked whether associations are formed between observable stimuli and responses (S–R) or between subjective sensory impressions (S–S). One group that included Hull, Guthrie, and Thorndike took the relatively objective S–R position, while Tolman and others favoured the more introspective, perceptual S–S approach. For a time S–R theorists held popularity; behavioral responses are readily observable evidence of learning, and many included them in the associative process itself.

But the reduction of learning to mere external stimuli and overt responses raised discordant theoretical objections that the inner activities of the organism were being ignored. S–R theories failed to account for a host of learned phenomena. For example, people could be trained to say they heard sounds even when such auditory stimuli were absent. They said they dreamed about what they had learned, too; yet there need be no immediate external stimulus, nor does the dreamer always make the responses he dreams about.

Physiological psychologists and biologists found ways of delivering electrical stimulation directly to the brain; this eliminated the sensory stimuli and vocal or motor responses on which S–R theories hinge. Direct neural stimulation was found to be an adequate signal and the electrical response of the brain itself proved susceptible to conditioning. At this level of the nervous system, distinctions between stimulus and response mean less than at the periphery, and the S–S versus S–R controversy is no longer such a burning issue.

Direction of association. Classical conditioning dependably has been shown to proceed only forward in time. Bell must precede food if a conditioned reaction is to be established. If it had any effect, the reverse procedure (food before bell) would be called backward conditioning; but at most it only inhibits other reactions. There seems to be a relatively brief optimal interval in classical conditioning at which associations are most easily made. For quick reflexes such as the eyeblink, this interval is about one-half second; longer or shorter intervals are less effective. For slower reactions such as salivation the interval is longer, perhaps two seconds or so.

Forward and backward association; context association

In learning verbal associations the situation appears to be quite different. When one learns the Russian–English forward association *da*–"yes," he also learns the English–Russian backward association "yes"–*da.* Moreover, timing is much less critical than in classical conditioning. Verbal pairs are learned with almost equal ease whether presented simultaneously or separated by several seconds.

In what is called context association, the general environment may begin to elicit a response that is being conditioned to a specific stimulus. Thus, a dog may salivate simply on being brought into the experimental room—before any bell rings. Verbal associations also can be weakened by changes in the general situation.

Repetition. A major theoretical issue concerns whether associations grow in strength with exercise or whether they are fully established all at once. Evidence is that learning usually proceeds gradually; even when a problem is solved insightfully, practice with similar tasks tends to improve performance. Some (perhaps most) learning theorists have concluded that repetition gradually enhances some underlying process in learning.

Stimulus sampling, a statistical learning theory

The view that associations develop at full strength in a single trial leads to a typical question. How can the gradual nature of most learning be explained if all-or-nothing is the rule? One possible answer suggested by Guthrie has led to so-called stimulus-sampling theory. The theory assumes that associations indeed are made in just one trial. However, learning *seems* slow, it is said, because the environment (context) in which it occurs is complex and constantly changing. Given a changing environment, the sample of stimuli will differ from trial to trial. Thus, it is reasoned, it should take many trials before a response is associated with a relatively complete set of all possible stimuli.

In this light, the strength (or probability) of a response should increase with practice even if the elementary associative process occurs in a single trial.

These stimulus-sampling notions translate easily into mathematical form; they are an example of statistical learning theory, a more general development in the quantitative treatment of learning.

Reinforcement. Repetition alone does not ensure learning; eventually it produces fatigue and suppresses responses. An additional process called reinforcement has been invoked to account for learning, and heated disputes have centred on its theoretical mechanism.

Objectively reinforcement refers to the use of stimuli that have been found to facilitate learning. Under appropriate conditions, these include praise, food, water, opportunity to explore, sexual stimuli, money, electric shock, and direct brain stimulation.

More theoretically, the term reinforcement expresses various theoretical hunches about some specialized subjective quality all such stimuli might share. Food for a hungry animal is a well-established reinforcer, conceivably through its distinctive appearance and odour. It tends to elicit a set of responses: approaching, chewing, tasting, swallowing; these may produce additional perceptual activities that reduce the drive or desire for food (*e.g.,* by halting stomach contractions that are experienced as hunger pangs). But no single subjective quality imagined by theorists seems invariably effective in reinforcement studies. Perhaps some combination of introspective influences is critical, or it may be that perceptual processes apply differently from one learning situation to another.

Anti-associationistic positions. Not all psychologists have accepted the general validity of association theories; many have suggested that considerations other than association are crucial to learning.

Organization. Major critics of association theory included such Gestalt psychologists as Wolfgang Köhler (1887–1967), who held that learning often entails a perceptual restructuring of environmental relationships. Köhler cited his own studies of insightful learning by a chimpanzee. The animal learned to join two sticks (akin to a jointed fishing pole) as a tool to pull in a banana that was out of arm's reach and of either short stick alone. The ape was described as sitting quietly (as if in thought), and then suddenly fitting the sticks together to rake in the fruit. It was argued that the ability to perceive new ways of relating the sticks to the banana was essential in solving the problem.

Learning as perceptual organization

Similar organizational processes in perceiving can be demonstrated in serial verbal learning. Memorizing the list *thick, wall, it, tea, of, myrrh, seize, knots, trained* should demand some rehearsal. Yet, notice the phonetic resemblance to Shakespeare's famous line from *The Merchant of Venice*: "The quality of mercy is not strained. . . ." With that kind of perceptual organization, learning can become quick and easy.

A powerful argument also was made by psycholinguists who criticized what they took to be the associationistic account of language learning. Even assuming one-trial acquisition, it was held that such individually learned associations could not account for all combinations of words people use; there are simply too many. They suggested that learning a language requires some general organizing structure on which words are hung. Some proponents of this position hold that this structure does not depend on learning, being transmitted genetically from parent to child.

Inhibition. Gestalt interpretations often reject the associationistic hypothesis wholesale. Other theorists endorse the notion of association, but hold it to be less important than is a process of inhibition through which errors in learning are eliminated. Such theorists find support in evidence for the development of learning sets (what is called learning to learn).

For example, a monkey may learn a long series of discriminations; *e.g.,* red versus green, black versus white, round versus square, large versus small, triangle versus ellipse. After solving several hundred such problems, some monkeys learn to master each new one in a single trial, as if insightfully. The animal is said to have learned to learn such discriminations.

Learning to learn: error-factor theory

Evidence clearly shows that the monkey gradually abandons erroneous tendencies as learning proceeds. At first it might be prone to choose stimuli that are red, black, round, large, or triangular. Correct choices do not always correspond to the animal's initial biases, and their suppression (inhibition, extinction) eventually permits single-trial learning. Theoretically, organisms learn to learn by inhibiting erroneous behaviour; thus, Harry F. Harlow, a proponent of this view, called it an error-factor theory.

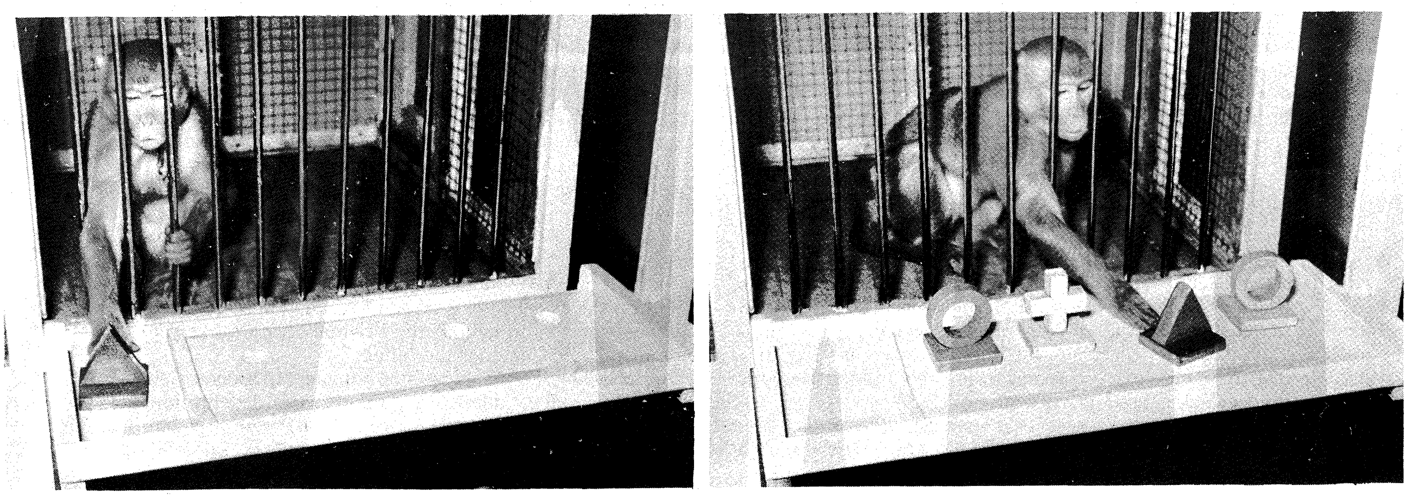
Motivation in learning. Motivation popularly is thought to be essential to learning. Yet many theorists suggest that motives make little or no direct contribution—that they simply tend to promote practice.

Motivation and performance. Learning was defined above as a change in a behavioral potentiality. Realization of such potential seems to be related to the learner's level of motivation. A pupil who has learned the names of all members of the British Commonwealth of Nations would be expected to recite them with particular energy under some sort of incentive (reward or punishment). The incentive is said to raise his level of motivation.

Incentives do seem to invigorate performance up to a point; however, when motivation seems particularly intense, some studies show performance to deteriorate. From such data some theorists conclude that the effect of drive intensity on performance follows a U-shaped course, first helping and later hindering.

Greatly increased motivation also may change performance qualitatively by introducing new inefficient modes of behaviour. A student may be so tautly driven to do well on an examination that his tension, fear of failure, and his visceral and muscular discomfort interfere with performance.

Motivation and learning. To show that motivation af-



Learning-to-learn experiment by Harry F. Harlow.
The monkey has been trained to expect food after lifting an object. In the first picture, the monkey lifts the triangle after having been shown a sample triangular form. In the second picture, taken several minutes later, the monkey selects the triangle from among other forms. The process is repeated with the circle and the cross.

fects performance of what has been learned is not the same as demonstrating its effect on the process of learning itself. This would require that individuals learn under various levels of motivation and be tested under the same incentive levels. (This is to control for the effects of motivation on performance alone.) And, indeed, the best-controlled experiments of this design indicate learning effects to be the same under different levels of motivation.

Varieties of learning. It is debated whether all forms of learning represent the same process. This question applies even to relatively primitive phenomena such as classical and instrumental conditioning.

Classical and instrumental conditioning

In instrumental conditioning reinforcement is contingent on the learner's response; a rat receives food only if it presses the lever. In classical conditioning there is no such contingency; a dog is fed whether or not it salivates. But this is a distinction in experimental procedure. Whether the underlying process of learning is the same for both is quite another question.

Classical conditioning usually has been reported for glandular, autonomically mediated, involuntary responses (*e.g.,* salivation, heart rate). By contrast, voluntary movements of skeletal muscles more typically have been found to be conditionable instrumentally. However, to theorize that classical conditioning is exclusively effective for one class of responses while instrumental conditioning is uniquely applicable to others seems to be a mistake.

Evidence that seems to demolish such theorizing comes from a series of experiments directed by Neal E. Miller at the Rockefeller University in New York City. Rats were immobilized with curare; this drug blocks the junction between muscle and nerve to paralyze the skeletal muscles. However, a curarized individual still can show autonomic, involuntary signs of emotional activity such as a rapidly beating heart.

Electrical stimulation of selected parts of the brain seems to be rewarding; animals behave as if they seek such stimulation and will learn to press a switch for it (voluntary muscle function). Using curarized animals, Miller and others made the rewarding stimulation contingent on such typically *involuntary* responses as changes in heart rate, blood pressure, contractions of the bowel, and salivation. Their research has shown such instrumental conditioning to be effective for all these responses. The evidence appears to destroy the once-popular hypothesis that involuntary autonomic reactions are subject only to classical conditioning. In this sense the two primitive forms of learning seem to be the same.

Stages of learning. Should the basic process prove to be the same for all varieties of learning, there would still be reason to believe that it operates differently from one stage of practice to another. For example, in coping with painful stimuli (*e.g.,* electric shocks) laboratory animals seem to learn in two successive, distinguishable phases. Apparently they first learn to fear the situation, *then* to avoid it.

For example, when an animal learns to avoid painful shock (by turning a paddle wheel or by running away), a warning signal can be given; *e.g.,* with a flash of light or a buzzer. The two stages of learning then can be studied separately. The animal first is subjected to pairings of signal and unavoidable shock to establish (by classical conditioning) signs of fear in response to the signal. In the second stage it is allowed to stop the frightening signal by making an appropriate response. Preconditioned members of the many animal species have learned to avoid the signal itself, even though shock never was presented again.

The role of fear

Theoretically, the classically conditioned signs of fright in response to the initially neutral signal have a motivating function. Termination of that stimulus is seen as instrumental—that is, as rewarding the animal by reducing learned experiences of fear.

Classical conditioning. A two-stage process has been suggested even for classical conditioning. One theory is that in the first stage the subject learns that a neutral stimulus (a ringing bell) is to be presented along with another stimulus (food) whether or not it exhibits a reaction (salivation). Conditioning of any reaction is held to constitute the second stage of learning. The skimpy supporting evidence points to the first stage as a prerequisite, suggesting that responses can only be conditioned after the sensory conditions are recognized.

Verbal learning. Theories that interpret verbal learning as a process that develops in stages also have been worked out. In one variety of rote learning the subject is to respond with a specific word whenever another word with which it has been paired is presented. In learning lists that include such paired-associates as *house–girl, table–happy,* and *parcel–chair,* the correct responses would be *girl* (for *house*), *happy* (for *table*), and *chair* (for *parcel*). By convention the first word in each pair is called the stimulus term and the second the response term. Paired-associate learning is theorized to require subprocesses: one to discriminate among stimulus terms, another to select the second terms as the set of responses, and a third to associate or link each response term with its stimulus term. Although these posited phases seem to overlap, there is evidence indicating that the first two (stimulus discrimination and response selection) precede the associative stage.

Remembering and forgetting. Learning, remembering, and forgetting often have been considered separate processes. Yet these distinctions seem to blur in the face of contemporary research and theory.

Transient and enduring memory. Evidence for stages of learning comes from observations of learners over relatively extended series of trials (or comparatively long periods). The empirical data suggest that several alterations in memory function occur even during a single trial. The process that commits information to memory also seems to have several stages.

Immediate, short-term, and long-term memory

Most theorists attribute at least three stages to memory function: immediate, short-term, and long-term. Immediate memory seems to last little more than a second or so. For example, subjects may be asked to remember where specific objects are located within a complex array they have just seen. Their performance shows that considerable information is retained only briefly, rapidly fading unless it is given special attention.

Short-term memory lasts about 15–30 seconds, as after looking up a telephone number. One makes the call, discovers he has forgotten the number (perhaps in the midst of dialing), and has to look it up again. Nevertheless, such short-term retention does make information available long enough to be rehearsed; if the learner repeats it to himself, the number can be transferred to some sort of longer term storage.

Thus, rehearsal seems to facilitate transfer of data from short-term to long-term memory. Once committed to long-term memory, the results of learning tend to endure but can be abruptly abolished when specific parts of the brain are injured or removed; they also are vulnerable to interference from other learning. Nevertheless, conditioned responses may undergo little or no forgetting over periods of months or years. And electrical stimulation of the surgically exposed brain while a person is awake can make him remember experiences long thought forgotten. Recall is reported to be similarly enhanced during hypnosis.

Retrieval. The amount of information one readily can retrieve from what is stored in memory is prodigious. In locating an item in memory, he apparently activates a system that stores a set of related data; then he searches for the item within that system. For example, a person is shown a long, randomly mixed list of words that belong to different categories (*e.g.,* names of animals, plants, professions, tools). When asked to remember as many words as he can, he spontaneously will tend to group them by category; this is called clustering of recall. Thus, names of animals (spread throughout the original list) are likely to be remembered one after the other.

Clustering; tip-of-the-tongue phenomenon

Studies of the familiar tip-of-the-tongue experience yield analogous results. College students who heard definitions (of this sort: a small, open Chinese boat) were asked to supply the right word (in this case it would be sampan). Those who said they might have it somewhere on the tip of the tongue were significantly accurate in guessing the first letter and the number of syllables. Their tendency also to recall words that sounded the same or that had similar meanings is reminiscent of clustering.

Considerable evidence of this kind supports the theory that the process of retrieval first locates stored data in some sort of associative network and then selects an item with specific characteristics.

Forgetting. Whether immediate and short-term data simply decay or are lost through interference is a matter of controversy. However, evidence is clearer that interference affects retention of information in long-term storage. Retention of the word *happy* (learned as a paired associate of *table*) seems to be subject to the interference of a strong tendency to associate *table* with *chair.* Thus, the paired associate *table–happy* becomes more readily forgotten when followed by *parcel–chair* as the very next item in a list; this seems to help *chair* reassert its old tendency to be associated with *table.* In general, it is found that associations tend to interfere with or to inhibit one another. Interference deriving from earlier (and later) associations is called proactive inhibition (and retroactive inhibition). These two forms of inhibition commonly are accepted as major processes in forgetting, proactive inhibition being assigned greater importance.

Contemporary trends in learning theory. In the early 1930s the distinction between learned and inherited behaviour seemed clearer than it does now. The view that any bit of behaviour either was learned or simply developed without learning seemed straightforward. Studies based on these expectations led investigators to conclude that rat-killing behaviour among cats is learned rather than instinctive, that human fears are all acquired, or that intelligence is completely the result of experience. Learning theorists were saying then that most behaviour is learned and that biological factors are of little or no importance.

Forty years later this position seemed grossly untenable. The once-implied sharp distinction between learned and inherited behaviour had become badly blurred. For example, it has been found that the young of many animal species automatically will learn to follow the first large, moving, noisy object presented (as if it were their mother). This special form of learning is called imprinting and seems to occur only during a critical early stage of life. Among mallard ducklings imprinting is most feasible about 15 hours after hatching. During this period a duckling will imprint as easily on an old man or on a rubber ball as it will on a mother duck. Is this instinctive or learned behaviour? Manifestly it is both. The instinctive tendency to be imprinted is part of the duckling's biological heritage; while the object on which it is imprinted is a matter of experience. What is significant for learning theory is that the contribution of biology cannot be ignored.

Imprinting, a special form of learning

Learning theorists once ruled a number of concepts out of court on the ground that they seemed objectively unclean. Image, cognition, awareness, and volition, all are concepts that were denied acceptance on this basis. They sounded mentalistic, subjective, introspective, and unverifiable. Yet, in the late 20th century these were being given more serious scientific consideration.

For example, the concept of image in learning has begun to show real viability. It has long been reported that the more meaningful a list of words is, the easier it will be to learn. Degree of meaningfulness for a word may be defined by the objectively observed probability that people quickly can give another word in response. Using such empirical scales of meaningfulness, a reliable and substantial relationship has been found between meaningfulness and ease of learning. However, meaningful words also may evoke vivid images that subjects can describe when asked. When they do evoke such imagery, they seem to be learned and remembered even more easily. Thus, learning theory seems to be enriched when introspective data are used.

A final fault in much learning theory stems from earlier tendencies to use the laws of physics as a model. Theorists once sought general laws of wide applicability that tended to obscure differences among individuals. For example, so complete was Hull's faith in universal "laws" of animal behaviour, that he based his hypothesis about the optimal interval for classical conditioning in humans, other mammals, and birds on the pattern of nerve conduction in the optic nerve of the horseshoe crab. There was little concern even for species differences. Within the same species, individual differences were viewed as a mere nuisance; it was believed that, by studying many subjects and by computing averages, basic laws of learning could be found. However, so-called laws were developed in this way that failed to represent even one individual whose behaviour contributed to the average. More than any other consideration, this has led learning theorists to take a belated look at the importance of individual differences and species differences in learning.

(G.A.K.)

Psychomotor learning

Human psychomotor skills are organized patterns of muscular activities guided by changing signals from the environment. Driving a car and eye–hand coordination tasks such as drilling a tooth, throwing a ball, typing, operating a lathe, and playing a trombone are behavioral examples. Also called sensorimotor and perceptual-motor skills, they are studied as special topics in the experimental psychology of human learning and performance. In research concerning psychomotor skills, particular attention is given to the learning of coordinated activity of the arms, hands, fingers, and feet; the role of verbal processes (see below *Concept formation)* is not emphasized.

THE RANGE OF SKILLS

The term "skill" denotes a movement that is reasonably complex and the execution of which requires at least a minimal amount of practice. Thus skill excludes reflex acts. One does not become skilled at sneezing or at blinking the eyes when an object approaches. At the same time, it has become increasingly apparent to scientists investigating the performance and acquisition of motor acts that the performance of complex skills is closely linked to sensations arising from the things the performer looks at, sensations from the muscles that are involved in the movement itself, as well as stimuli received by other sensory end organs. Thus the term "sensorimotor skill," denoting the close relationship between movement and sensation in the acquisition of complex acts, is found within the research literature with increasing frequency.

Simple components of bodily skills. Most of life's skills are continuous and complex and contain a multitude of integrated components; however, these complex skills may be analyzed by examination of their component parts.

Reception of stimulus and selection and initiation of response

The performance of a skill may be broken down into several time intervals. Initially the performer must be attentive and alert enough to be receptive to some kind of sensory information, which may in turn lead toward some kind of motor act. For this to occur the performer usually becomes aware of some kind of stimulus that is intense or distinctive enough to be perceived as different from other sensory information.

As this cue occurs the performer then must make a decision to act or not to act; and this is generally dependent upon his past experience within similar situations and with similar stimuli, as well as upon his feelings about his personal capabilities. If he decides to act, the next thing that occurs is the selection of an appropriate motor response from the entire "collection" of motor responses that he has acquired.

In the laboratory, a subject's reaction time is the time between the presentation of some kind of stimulus and the performer's initiation of response. The individual's speed of reaction in such situations is dependent upon a number of variables, including the intensity of the stimuli; for example, a person will initiate a movement more quickly to increasingly louder sounds until a limit is reached. If too loud a sound is presented it will delay the onset of the movement and result in a longer reaction time. Similarly, a longer reaction time will be recorded in such experiments if the subject is aware that he will have to initiate a complex movement or if the subject must choose from among a number of stimuli before initiating a movement (*e.g.,* if he must move only if one of a number of various coloured lights is turned on).

Factors that affect quality of motor response

The quality of the movement then initiated is dependent upon a number of other factors, the precision of the act required, the past experience of the performer in similar skills, the speed of the movement, the force of the motor act, as well as body part or parts to be moved.

The efficient performance of many types of extremely simple motor skills may be limited by inherent response capacities built into the human nervous system. Finger tapping at more than ten times per second for example is not usually possible. A person's ability to keep a body part relatively steady may be disturbed by natural, regular, and rather predictable oscillation rates of the limbs, fingers, and of the total body (evidenced in measures of body sway).

Individuals vary greatly in their ability to exercise force with various body parts. At the same time, there are limits beyond which it is not realistic to expect humans to go when evidencing strength in various tasks.

Careful experimentation reveals that because of the complexity of the human motor system it is unlikely that an individual ever repeats an apparently similar movement in precisely the same way. Thus the acquisition of skill in a given task involves the performance of a reasonably consistent response pattern, which varies, within limits, from trial to trial.

Basic abilities that contribute to motor skills

It is a common observation that there seem to be a number of basic motor abilities that may underlie the performance of a number of life's activities. This subject was investigated rather extensively during the 1950s and 1960s. The intercorrelation of performance scores elicited from thousands of young adult subjects who participated in these investigations revealed that there are in fact a number of basic abilities that contribute to the efficient performance of both fine and gross motor skills. Although a detailed examination of these abilities is not possible here, in general it was revealed that there are five components of what might be broadly referred to as "manual dexterity," including fine finger dexterity, arm-wrist speed and aiming ability. Similar research has explored the manner in which performance scores group themselves in skills in which larger muscles are involved. It was found, for example, that there are several kinds of strengths, including static strength (pressure measured in pounds exerted against an immovable object); this is independent of what was termed "dynamic strength" (moving the limbs with force). Muscular flexibility and balancing ability were similarly dissected into several components. Thus discussion of a single quality in human movement is inaccurate. Rather, one should refer to several specific types of ability.

Other classifications of motor skills

Motor skills may also be classified by reference to the more general characteristics of the tasks themselves rather than by measuring and intercorrelating the scores elicited by human subjects. It is common, for example, to find the dichotomy "fine" and "gross" motor skill, in which the latter label is applied to acts in which the larger muscles are commonly involved, while the former classification denotes actions of the hands and fingers. One researcher has proposed a three-way classification system, separating human movements into "body transport movements," in which the total body moves in space; limb movements; and manual (hand) actions. In general, however, most skills incorporate precise movements of both larger and smaller muscle groups working in harmony. The basketball player must use his larger skeletal muscles to run and jump but at the same time must employ a fine "touch," evidencing accurate finger control, when dribbling the ball. On the other hand, when a person sits at a desk and writes, the large postural muscles that contribute to the writer's stability are invariably active.

Complex, integrated skills. Most of life's skills are not simple ones. They are rhythmic at times and almost always are composed of several integrated parts. Such skills are often controlled by the organization of visual information available to the performer, particularly during the early stages of learning. At the same time, the individual's ability to analyze the mechanics of a motor task, his verbal ability, and other intellectual and perceptual attributes may influence his acquisition of a skill.

Although psychomotor skills are widely distributed—*e.g.*, in military, athletic, musical, and industrial settings—such complex situations typically do not lend themselves to rigorous experimental research. Most scientists have found it more analytically useful to study psychomotor learning under controlled laboratory conditions. Measures of proficiency obtained in the laboratory reflect increasing accuracy and decreasing variability in a learner's performance as training progresses. If there is sufficient genetic aptitude, a person's mastery of a skill depends on his motivation to improve, on his receiving continuous information or sensory feedback about the adequacy of his performance during training, and on such factors as the rewarding effects of corrections made during successive practice periods. Skills are susceptible to inhibitory influences. The full extent of gains in proficiency often is masked by temporary losses and emerges only later, without additional practice sessions.

Factors in performance

Psychomotor habits are mediated primarily by the sensory and motor cortex of the brain and by the neural fibres (commissures) that connect the two cerebral hemispheres. According to the majority of theoreticians, learning proceeds (habit strength develops) as a mathematical function of the amount or duration of rewarded (reinforced) practice. The effects of associative and motivational factors are believed to combine mathematically by multiplying one another, while inhibitory and oscillation (variability) factors are thought to have subtractive effects. Despite theoretical and empirical progress, much remains to be discovered about the learning and performance of psychomotor skills, especially about the interrelationships among training variables, feedback contingencies, and human-factor variables.

(C.E.N./B.J.C.)

LABORATORY RESEARCH IN PSYCHOMOTOR LEARNING

Devices and tasks. Hundreds of electrical and mechanical instruments have been developed for research in psychomotor learning, but those commonly used number less than two dozen. In operating a device called a complex coordinator, the learner is instructed to make prompt, synchronized adjustments of handstick and foot-bar controls to match different combinations of stimulus lights. Another device, a discrimination reaction timer, requires that one of several toggle switches be snapped rapidly in response to designated distinctive spatial patterns of coloured signal lamps. In performing on a manual lever, a blindfolded subject must learn how far to move the handle on the basis of numerical information provided by the experimenter. With a so-called mirror tracer, a six-pointed star pattern is followed with an electrical stylus as accurately and quickly as possible, the learner being guided visually only by a mirror image. The operator of an instrument called a multidimensional pursuitmeter is required to scan four dials and to keep their indicators steady by making corrections with four controls of the type found in an airplane cockpit. On a rotary pursuitmeter the trainee's task is to hold a flexible stylus in continuous electrical contact with a small, circular metal target set into a revolving turntable.

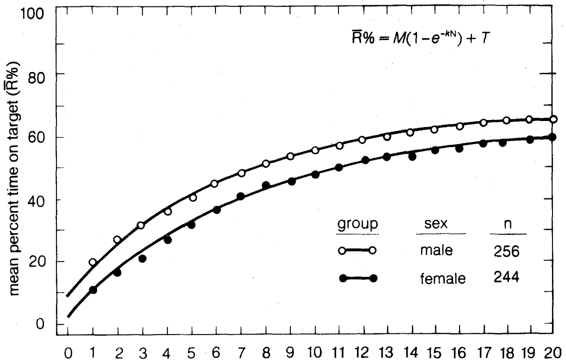
The pursuitmeter and the mathometer

Also employed in such research is a selective mathometer, on which the subject's problem is to discover, with cues provided by a signal lamp, which of 19 pushbuttons should be pressed in response to each of a series of distinctive images projected on a screen. While using a star discrimeter, a person receives information about his errors through earphones; his task is to learn to selectively position one lever among six radial slots in accordance with signals from differently coloured stimulus lights. A trainee on a two-hand coordinator has to manipulate two lathe crank handles synchronously to maintain contact with a target disk as it moves through an irregular course.

Measurements. The tasks required by the above devices produce a substantial range of psychomotor difficulty. The elements of skilled behaviour are expressed as numerical scores; *e.g.*, correct response and error percentages, amplitude and speed of movement, hand or foot pressures exerted, time on target, reaction time, rate of response, and indices of time-sharing activity. Most of the behaviour thus recorded lends itself readily to mathematical treatment. Laboratory devices for studying psychomotor learning characteristically exhibit high reliability (*i.e.*, intra-task consistency) and yield scores of useful validity (extra-task correlation) in predicting such behaviour as performance in factory work and the operation of motor vehicles and aircraft. In other words, it would appear that perceptual-motor devices reliably measure what they are designed to measure, and they also tap a significant proportion of the abilities required in real-life situations (see also PSYCHOLOGICAL TESTS AND MEASUREMENT). When properly maintained and used under standardized conditions, and when the resulting measurements are treated by statistical methods, the above devices are prime choices for many applied and basic research programs.

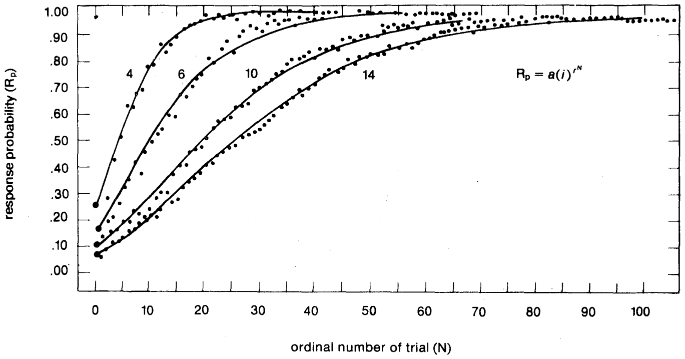
PHENOMENA OF PSYCHOMOTOR LEARNING

Acquisition. Speed and accuracy in the majority of psychomotor tasks studied are typically acquired very rapidly during the early stages of reinforced practice, the average rate of gain tending to drop off as the number of trials or training time increases (Figure 1). Curves based on such measures as reaction time or errors reflect the learner's improvement by a series of decreasing scores, giving an inverted picture of Figure 1. Tracking scores from the two sexes are seen in Figure 1. Other devices have yielded more complicated functions—*e.g.,* S-shaped curves for complex multiple-choice problems on the selective mathometer (Figure 2). Most acquisition curves obey a law of diminishing returns as high levels of skill are approached.



From Clyde E. Noble, "Acquisition of Pursuit Tracking Skill Under Extended Training as a Joint Function of Sex and Initial Ability," *Journal of Experimental Psychology,* vol. 86, pp. 360–373 (1970); Copyright (1970) by the American Psychological Association, and reproduced by permission

Figure 1: *Rotary pursuit acquisition curves.*
Average percent of response time on target ($\bar{R}\%$) while using a rotary pursuitmeter is plotted as a mathematical function of the number of successive 5-trial blocks of practice (N), with sex as the distinguishing factor, for 500 subjects. Both curves are predicted by the exponential equation shown in the figure, where *T, k,* and *M* represent origin, rate, and limit of theoretical response time, respectively.



From Clyde E. Noble, "The Length-Difficulty Relationship in Compound Trial-and-Error Learning," *Journal of Experimental Psychology,* vol. 54 (1957), pp. 246–252 (1957); Copyright (1957) by the American Psychological Association, and reproduced by permission

Figure 2: *Selective mathometer acquisition curves.*
Response probability (R_p), or relative frequency of correct multiple choices, in a selective mathometer learning experiment is plotted as a mathematical function of the number of successive practice trials (N), with task length (4,6,10, or 14 units) as the experimental condition, for 192 male subjects. All four curves are predicted by the double-exponential equation shown in the figure, where *I, r,* and *a* represent origin, rate, and limit of theoretical response probability, respectively.

Data such as those from tracking and multiple-choice tasks can be explained by rational mathematical equations derived from theoretical models (see formulas and captions in Figures 1 and 2). Between them, these two equations describe psychomotor acquisition curves from a wide variety of learning situations and of trainees with less than a 2 percent average error of prediction. Contrary to lay opinion, stepwise plateaus of proficiency are seldom seen, not even in learning Morse code. The "natural plateau" is a phantom.

Similarity and performance

Generalization and transfer. The occurrence of the phenomenon of generalization is seen in the tendency of laboratory subjects (conditioned to respond to a particular stimulus—*e.g.,* a light) to respond as well to similar stimuli beyond the original conditions of training. As differences along a physical continuum (*e.g.,* brightness) between the stimuli used in training and those encountered on test trials increase, the effects of generalization decrease until there may be no transfer from one situation to another. Alternatively, the more the two situations have in common, the greater is the amount of predictable transfer. Generalization (or transfer) may be based on temporal patterns of stimuli (*e.g.,* rhythms), spatial cues (*e.g.,* triangularity), or other physical characteristics (see below *Transfer of training.*).

The measured effects of prior training on the performance of a subsequent task define the transfer of psychomotor learning. Although similar, the latter task usually differs measurably from that originally practiced. A common example is the ability required of many automobile drivers to change easily from, say, a three-speed transmission with a horizontal gear lever on the steering wheel to a four--speed mechanism with a vertical floor-mounted gearshift. In laboratory tasks the amount and direction of transfer effects are accurately predicted. In practical skills, transfer is more likely between tennis and badminton than between swimming and football, between cornet and trumpet than between piano and tuba. Similarity is not the only correlate of transfer, however, and empirical studies must take account of such factors as the amount of practice and the sequence of events in previous training.

Transfer effects may be positive, negative, or zero; *i.e.,* learning one task may facilitate, hinder, or have no observable influence upon performance of the next task. Flight simulators are designed to maximize the amount of positive transfer, often by ensuring high levels of behavioral similarity. Negative transfer effects (*e.g.,* reaching for the floor to shift gears when the lever is on the steering wheel) appear occasionally but tend to be easily overcome. Since transfer necessarily involves retention, the best schedules minimize forgetting by the inclusion of short time intervals between training and transfer.

The degree and amount of transfer are contingent upon such factors as number of common elements or principles, stimulus and response similarity, amount of predifferentiation training, the variety of learning-to-learn experiences, part–whole relationships, differences in intertask complexity, use of mnemonic aids, and the extent of proactive or retroactive interference. Transfer equations usually assume that the basic indices of performance for experimental and control groups will increase with practice, that the possible measures range from negative 100 percent through zero to positive 100 percent, and that the groups have been equated in aptitude or initial ability to learn before the experimental treatments are begun. Retroactive interference designs typically employ a sequence of original learning, interpolated learning, and relearning.

Effects of rest intervals

Retention. Learning is to acquisition as memory is to retention. Psychomotor retention scores indicate the percentage or degree of originally learned skill that is remembered or recalled as a function of elapsed time. Alterations of motor memory are reflected by changes in means, variances, and correlations between test results. In contrast to verbal behaviour, which is notoriously susceptible to forgetting through interference within a matter of seconds, mean scores for tracking and coordination skills recorded over periods ranging from two days to two years diminish scarcely at all. Yet, when intervals of three minutes to six weeks are interpolated between discrete responses on a manual lever device, performance remains stable for about two days and then becomes inconsistent; variabilities increase and correlations decrease as the subjects mis-recall more and more of their original skill. In the light of this evidence, motor memory may be viewed as a phenomenon of persistence, while forgetting is a case of inconsistence.

One hypothesis advanced to account for the greater retentivity of psychomotor behaviour, as compared to that of newly acquired verbal behaviour, is that nonverbal striped-muscle responses are more often overlearned and are less susceptible to proactive interference (*i.e.,* competition arising from things learned in the past). Distinctions between immediate, short-term, and long-term memory are also less prominent in studies of motor learning, possibly because of the devotion of skills specialists to efficient practice and feedback methods that ensure permanent storage of habits in the brain. This is not to say that motor skills are unforgettable; studies of short-term memory suggest that psychomotor forgetting can be swift indeed. Regardless of theoretical differences, however, psychologists generally agree that psychomotor behaviour is best remembered (and least forgotten) when overlearning is high, interference is low, reinforcing feedback is optimal, and interpolated activities are unrelated to the task being learned. Time is less important in the degradation of memory than are the events that fill the time (see also MEMORY).

Reminiscence. The phenomenon of reminiscence is a gain in performance without practice. Thus, when subjects performing trial after trial without rest (massed practice) are given a short break, perhaps midway through training, scores on the very next trial will show a significant improvement when compared with those of a massed group given no break. Reminiscence effects are most prominent in tasks demanding continuous attending and responding; they are least often observed with discrete-responding apparatus. The theoretical importance of this concept derives from its role in testing a hypothesis of reactive inhibition that asserts that a decremental process cumulates in the organism as a positive function of responding to stimulation and a negative function of resting time. That the phenomenon of reminiscence also manifests bilateral transfer of skill (*e.g.,* from the left to the right hand) suggests that the locus of the decrement is in the central nervous system rather than in the peripheral effector organs. Indeed, merely watching another subject practicing on a rotary pursuitmeter has an inhibitory effect on a person's performance; yet the cause is neither boredom nor fatigue.

Warm-up. Athletes and musicians often report that they get "cold" during a layoff (even for a rest period of a mere five minutes); when practice is resumed, the decrement in performance requires a warm-up before it is overcome. Similarly, on a rotary pursuitmeter, it is necessary to regain the optimal posture, grip the stylus correctly, begin the coordinated movements of eyes and hand, and recapture the proper whole-body rhythm. Warm-up produces a further gain in proficiency following the initial reminiscence effect. Mean scores continue to rise for several trials, reach a peak at the level found for distributed practice, and then fall more gradually until they merge with the curve for massed practice. When the duration of rest is extended, the amount of warm-up decrement first increases rapidly and then decreases; similar findings obtain for a succession of work and rest periods. Investigators who have tried to substitute warm-up activities other than actual pursuitmeter practice to offset or reduce the magnitude of the decrement have not been successful. At least for continuous psychomotor tasks of this sort, the need for proper, task-specific warm-up appears to be an intrinsic requirement of efficient performance. Wherever reminiscence goes, warm-up seems to follow; yet the converse does not always hold. The connection between warm-up and forgetting is uncertain.

Warm-up as a practice factor

Refractory period and anticipation. When required to make quick, discrete responses to two stimuli separated in time by one-half second or less, an operator's reaction time (latency) for executing the second response is typically longer than that of his first response. This difference in reaction time is called the psychological refractory period. At one time, it was thought possible that sensory feedback from the first response might stack up in the nerve centres to make the system refractory for a brief time, thereby delaying the processing of the second stimulus. Research findings that erroneous reactions could be corrected within one-tenth second would seem to negate the hypothesis. An alternative suggestion is that corrective movements are facilitated by feedback from the incorrect ones, and controlled observations appear to confirm that error-correcting responses have shorter latencies than those that are either correct or erroneous. Apparently, a false movement can be stopped on the basis of internal cues more promptly than on that of external stimuli.

Expectancy is a collateral factor with which researchers have had to reckon; *i.e.,* a subject may learn to accommodate himself to expect a delay between the first and second stimulus and thus be relatively unprepared should the second arrive earlier than usual. Further, people learn to be more expectant for particular kinds of stimuli than for others. When a person is uncertain about whether regularly occurring stimuli will be auditory or visual, or when their spatial direction is uncertain, performance is significantly degraded. This would suggest the possibility of divided attention; indeed, when pairs of stimuli are made perfectly predictable as to time and type, no impairment of response is observed.

Learning to divide attention

If a subject can acquire suitable expectancies via training and experience, then he can improve the skill of dividing his attention and, within physiological limits, simultaneously handle an increased range of stimuli without loss of proficiency. Results from extended practice on a task requiring successive choice and dual reaction indicate that, with learning, people can reduce the psychological refractory period. The ability to develop anticipatory responses to regularly occurring stimulus cues is well established. A military gunner scanning a distant fixed target for azimuth and elevation, for example, is engaging in a preview of receptor anticipation to maximize his score. An operatic tenor who rehearses covertly the opening notes of his cadenza while the orchestra finishes the introduction is employing perceptual anticipation to optimize his rendition. Anticipatory timing is learned, and reinforcing feedback is necessary.

FACTORS AFFECTING PSYCHOMOTOR SKILL

Amount of practice. It has been noted above (Figure 1) that the practice of sensorimotor tasks usually produces changes in scores that reflect diminishing returns. A major influence in learning generally, repetition is the most powerful experimental variable known in psychomotor-skills research. But practice alone does not make perfect; psychological feedback is also necessary. The consensus among theoreticians is that feedback must be relevant and reinforcing to effect permanent increments of habit strength. Once developed, habit never dies; it does not even fade away.

The effects of feedback and four other important performance variables (*i.e.,* task complexity, work distribution, motive-incentive conditions, and environmental factors) remain to be summarized.

Psychological feedback. Ranking prominently among experimental variables are so-called feedback contingencies (aftereffects, knowledge of results) that may be controlled by the experimenter so as to occur concurrently with or soon after a subject's response. A learner appears to improve by knowing the discrepancy between a response he has made and the response required of him; but, in experimental practice, the investigator manipulates behaviour by transforming functions of error. Since transformations are usually numerical or spatial, sensory returns from one's action may be informative, motivating, or reinforcing. Response-produced stimulation is intrinsic to most skeletal–muscular circuits; the neural consequences of bodily movement are fed back into the central nervous system to serve the organism's regulatory and adaptive functions. When this normal feedback is interrupted or delayed, psychomotor skill is often seriously degraded. Experimentally delayed auditory feedback of a subject's oral reading produces stuttering and other speech problems; delayed visual feedback in simulated automobile steering is a greater hazard under emergency conditions than is the driver's reaction time.

Generalizations about psychomotor learning

Laboratory investigations have supported the following generalizations about psychomotor learning: (1) without some kind of relevant feedback, there is no acquisition of skill; (2) progressive gains in proficiency occur in the presence of relevant feedback; (3) performance is disrupted when relevant feedback is withdrawn; (4) delayed feedback in continuous (but not discrete) tasks is typically decremental; (5) augmented or supplementary feedback usually results in increments; (6) the higher the relative frequency of reinforcing feedback, the greater is the facilitation of skill; and (7) the more specific the feedback (*e.g.,* in designating location, direction, amount), the better is the performance.

Experiments with a manual lever device, for example, suggest that when feedback is introduced and withdrawn at four stages of practice, the effect on error scores is profound. Knowledge of results given early and late has effects similar enough to reject any hypothesis that learning arises merely from repetition. These experiments indicate that practice makes perfect only if reinforced; the result of unreinforced practice is extinction of the correct response and a proliferation of errors. Studies employing a complex mirror-tracking apparatus have clarified the role of reinforcing feedback. Targeting performance was facilitated by presenting distinctive supplementary visual feedback cues previously associated with aversive (electrical shock) and nonaversive consequences. Moreover, the amount of facilitation grew curvilinearly with the number of cue conditioning trials. Work on human incentive learning thus demonstrates that the rate of gain in psychomotor proficiency can be regulated by stimuli that have been accompanied by positive or negative aftereffects. Persistence of the acquired reinforcing effects, considered with their cumulative quantitative properties, enhances the attractiveness of theoretical interpretations that emphasize continuity and reinforcement as contrasted with theories based on discontinuity and contiguity alone. Clark Hull's system (1943) is the classic model.

Human incentive learning

Task complexity. The complexity of discrete psychomotor tasks may be specified either as the number of response sequences a subject can make or as some measure of a subject's uncertainty about choices among stimuli. Still other factors that have been investigated as instances of complexity include variations in the number of possible responses at each choice point, different lengths of series, and regular versus unpredictable stimulus sequences.

Experimental procedures involving an increase of complexity produce more errors, require more trials to reach proficiency, and result in longer latencies per trial. Difficulty in psychomotor learning, therefore, generally increases with the complexity of the task to be mastered. An example of this phenomenon appears in Figure 2. Subjects exhibit continually altered probabilities of response during training sessions, and an average person with enough practice on a discrete sensorimotor task can learn to perceive, select, and react as fast to ten stimuli as he can to two. Apparently, it is not the number of choices among stimuli as much as it is the number of choices among responses that slows up a subject's processing activities and complicates his decision problems. Indeed, by limiting response alternatives (*e.g.,* circumscribing the physical range of a trainee's movements or providing supplementary auditory and visual indicators of error), a training device can facilitate the acquisition or transfer of skill.

Work distribution. Hazardous though generalizations about work and rest in psychomotor learning may be, a few guiding principles are notable: (1) massed practice is usually superior to distributed practice for simple discrete-trial tasks; (2) distributed practice is usually superior for complex continuous-action tasks; (3) short practice sessions are generally superior to long practice sessions; (4) long rest periods are generally superior to short rest periods, although forgetting must be counteracted; (5) for continuous-tracking tasks practiced under constant work sessions and variable rest periods, the final proficiency level grows curvilinearly as the intertrial interval is lengthened; (6) gains in proficiency under distributed practice, or with interpolated rest periods during massed practice, are usually in terms of performance rather than of learning; (7) losses in proficiency under massed practice, or with increased work load, usually pertain to inhibitory rather than motivational decrements; (8) under certain conditions (*e.g.,* "cramming" for examinations) it may be most efficient to mass practice as long as adequate rest can be obtained before criterion performance is demanded; (9) reminiscence increments and warm-up decrements are intimately related to schedules of work and rest; (10) decrement is not the same as fatigue.

Quite apart from the practical question of the optimal management of training programs (*e.g.,* in coaching oarsmen in racing shells), the aversive inhibitory consequences of sustained action that are recognized as subjective fatigue and behavioral decrement are clearly adaptive. By a reflex negative-feedback mechanism, inhibitory impulses may prevent an organism from working itself to exhaustion. With few exceptions, the presumption in favour of spaced practice can safely be taken out of the psychomotor-skills laboratory and applied in the gymnasium, lake, and playing field. Research on the skills involved in, for example, archery, badminton, basketball, golf, javelin throwing, juggling, marksmanship, rowing, and tennis supports the notion of distributing training by means of short workouts and frequent breaks.

Distributed practice applications

Motive-incentive conditions. Motivational processes are states of the organism that serve to activate reaction tendencies. Such states are classified as primary (innate) or secondary (acquired, learned) motivation. Though common physiological needs (*e.g.,* for food, water, avoidance of pain) may evoke psychological drives (*e.g.,* hunger, thirst, pain), the concepts of need and drive are not perfectly correlated. Some needs (*e.g.,* oxygen demand) seem to have no specific behavioral drive, and for some drives, clear-cut biological needs remain to be identified (*e.g.,* curiosity). Despite this apparent discrepancy, there is a theoretical consensus that psychological drive arouses the body to action, energizes its latent responses, and supports its behaviour over time. Most theorists believe that motivation (drive) and learning (habit) interact (in a multiplicative—drive times habit equals action—manner) in generating response. In other words, to produce action both are theoretically indispensable, but neither is sufficient alone. A person is not likely to perform a skill if he does not want to and cannot do so if he does not know what to do. The multiplicative theory implies that the same level of psychomotor proficiency may arise from quite different combinations of learning and motivation. Moreover, the organism's temporary drive state seems clearly to affect the adequacy of reinforcing feedback (*e.g.,* offers of monetary reward do little to arouse one who is already trying his level best). While these theoretical interpretations often apply well to laboratory animals, their application to human acquisition of skill is complicated because incentive learning in man can become very abstract.

Physiological explanations of human behaviour that depend on the concept of primary motives (derived from research with rats and dogs) run into difficulties in view of the fact that primary motivation and reward do not appear to be critical in most studies of human skill acquisition. Thus, instead of giving food pellets (as to a rat), an experimenter delivers praise to a human subject; rather than receiving feedback by electric shock, the human can be guided by a needle moving on a dial or a buzzer signalling an error. At any rate, despite efforts to distinguish such motivational factors as general drives from selective incentives, attempts to demonstrate significant motivational effects in human psychomotor learning have met with only modest success. Among exceptions to the above are a few studies with standard apparatus (*e.g.,* the complex coordinator) and with special devices that have indicated that such incentives as money, verbal threats, electric shock, exhortations, and social competition may be relevant. Significant effects frequently fail to appear in experiments, and findings are often contradictory, so it has been suggested that the intrinsic challenge of the

Contradictory findings about motivation

gadgetry, coupled with the subjects' already high preexperimental motivation, leaves human volunteers unaffected by such weak laboratory manipulations of motive-incentive conditions as the foregoing (see also EMOTION AND MOTIVATION).

Environmental factors. Many practical skills must be executed outside the laboratory under unfavourable conditions of temperature, humidity, illumination, and motion. It is generally found that below the limiting levels of extreme stress, such conditions affect psychomotor performance to a greater extent than they affect psychomotor learning. Representative findings have included the following: (1) isolation and sensory deprivation cause dramatic reductions in vigilance and monitoring skills within an hour; (2) environmental temperatures above or below 70° ± 5° F tend to lower scores on tracking apparatus but do not impair learning; (3) lack of oxygen slows reaction time, especially when the atmosphere corresponds to altitudes of 20,000 feet or higher; (4) accelerations of the body in a centrifuge or rotating platform disrupt postural coordination and produce systematic shifts in the perception of the vertical; (5) although such people as acrobats, dancers, pilots, and skaters can adapt well to high accelerations, even they lose equilibrium if deprived of the customary visual frame of reference; (6) rather mild centrifugal effects of slow, constant rotation may induce acute motion sickness and associated degradation of psychomotor proficiency in normal persons; (7) while some controlled work-rest schedules of crews during confinement in a small cabin upset daily sleep rhythms and lead to decrements in watchkeeping, memory, and procedural skills, a schedule of four-hours-on versus four-hours-off duty can be maintained for several months without significant impairment; (8) faulty identifications of visual displays on an eye–hand matching task have been produced in volunteer subjects exposed to controlled infectious diseases (*e.g.*, respiratory tularemia, phlebotomus fever, viral encephalitis).

Other environmental stress variables found to exert negative influences are vibration, low illumination, high atmospheric pressure, noise, glare, toxic gases, ionization, and subgravity. Certain drugs have positive effects on psychomotor performance (*e.g.*, amphetamines, magnesium pemoline, methyl caffeine, pipradrol); some have deleterious effects (*e.g.*, alcohol, barbiturates, diphenhydramine hydrochloride, lysergic acid, meprobamate, phenothiazines, scopolamine, tetrahydrocannabinol, tripelennamine); and others are either neutral or have inconsistent effects (*e.g.*, caffeine, nicotine).

INDIVIDUAL AND GROUP DIFFERENCES

Statistical indices of psychomotor ability (*e.g.*, means, variances, correlations) not only differ among individuals but may also serve to distinguish from each other groups of persons classified by such traits as age, sex, race, personality, and intelligence. Comparative psychological studies of monozygotic (identical, one-egg) and dizygotic (fraternal, two-egg) twins have indicated that high coefficients of heritability—measured as a ratio of genotypic to phenotypic variance—exist for perceptual, spatial, and motor abilities.

Types of changes with age

Age. The most pervasive differences in human performance on psychomotor apparatus are associated with chronological age, and scores obtained from nearly all the devices mentioned above are sensitive to age differences. Researchers generally report a rapid increase in psychomotor proficiency from about the age of five years to the end of the second decade, followed by a few years of relative stability and then by a slow, almost linear decrease as the ninth decade is approached. For simple hand or foot reactions, complex discrimination-reaction time, and coordinated automobile steering, the peak of skill is attained between ages 15 and 20 on the average, and then performance at age 70 declines to about the level of age 10. This is a two-stage process: first, a developmental phase (*e.g.*, through maturation), followed by the more gradual deterioration of aging. Common athletic skills (*e.g.*, balancing, catching, gripping, jumping, reaching, running, and throwing) also improve through childhood, and it is well known that most athletes reach their prime before the end of the third decade. Olympic events requiring great muscular strength or stamina (*e.g.*, swimming) are dominated by athletes in their teens and 20s, whereas practitioners of more refined technical skills (*e.g.*, gymnastics) tend to be older. Self-paced, leisurely sports (*e.g.*, golf) are favoured over opponent-paced, combative activities (*e.g.*, tennis) as the aging process continues. Hereditary potentialities require several years to become established. It is probable that the genetic factors that underlie growth rates, and the age sequence in which different kinds of behaviour first appear, affect learning as well as performance.

Sex. Although the assessment of sexual differences in perceptual and reactive abilities is complicated by a number of factors (*e.g.*, age, race, and personality), girls and women tend to be more proficient than boys and men in such psychomotor skills as finger dexterity and inverted-alphabet printing. On the other hand, males generally do better than females at pursuit tracking, repetitive tapping, maze learning, and reaction-time tasks. On rotary pursuitmeter tests, women are not only less accurate but more variable than men of the same age and race (Figure 1). Although males appear to be superior to females in aptitude and capacity, these advantages disappear when subgroups are carefully matched for initial ability. In contrast, speed scores on discrimination-reaction tests reveal clearly diverging trends for college men and women trained intensively for several days (960 trials). This seems to be a genuine sex difference rather than an element of measurement or selection. Though both groups were equated for intelligence and had similar error scores, females began to suffer cumulative impairment on the fourth day of practice, whereas males kept improving. Sizable average differences in reaction latency as well as in movement time are characteristic of the sexes on other tasks.

Athletic skills

Whereas girls tend to attain their maximum proficiency in speeded tasks earlier in life than boys do, males continue to gain over a longer period and maintain their superiority over females for about half a century of the lifespan. After puberty, boys excel at most athletic skills demanding stamina and strength (*e.g.*, jumping, running, throwing). Thus, female Olympic swimming and track-and-field records are inferior to those of males and are achieved by girls who are noticeably younger than male champions in the same events. Sex has also been implicated in experiments employing complex coordinators, mirror tracers, and selective mathometers, with boys and men typically surpassing girls and women. Not all psychomotor differences associated with sex are intrinsically biological; unequal opportunities, distinctive social learning, role playing, and other culturally conditioned influences undoubtedly modulate the learning and execution of skills by males and females.

Race. All mankind is of one species. Zoologically, human races are all mutually interfertile subspecies—*i.e.*, breeding populations that differ in the relative frequencies of one or more genes. Although the variety of possible traits is practically limitless, random mating is not the case. Because of historical inbreeding tendencies, it is statistically improbable that any two human races have the same means and variances for all psychological traits. Not surprisingly, therefore, significant differences in psychomotor behaviour are found among ethnic groups throughout the world.

In one classic set of data (1904) on form-board skill (fitting nine geometric forms into correct holes), the average time in seconds for completing the task varied for different races—*e.g.*, seven African Pygmies (82.20), 12 Philippine Negritos (63.30), 55 American Indians and Eskimos (34.24), and 74 U.S. whites (27.80). Average error scores fell in the same order and, consistent with a genetic hypothesis, hybrid groups were appropriately ranked in between. These small samples were, however, not necessarily representative; *i.e.*, no effort was made to equate for differences in cultural values and psychomotor experiences found in different societies. Furthermore, since these were average differences, no conclusions about individual persons were warranted. It is interesting that the rankings were found to change for other psychomotor tasks; *e.g.*, on a test of tapping and aiming skill, Eskimos surpassed all others, followed by Filipinos, who were in turn trailed by

Caucasians. A more recent study (1967) discovered that a sample of Mongoloid Chamorro people from Saipan and another of Indians from the U.S. exceeded white norms on a test of maze-tracing ability. In both studies, people of mixed races tended to make intermediate scores, a fact consistent with contemporary research in behavioral genetics and physical anthropology.

Absence of a superior race

No particular race is found to be uniformly superior in all psychomotor aptitudes and capacities, but environmental causes seem to be inadequate to explain differences in rate of acquisition and final level of performance on standard apparatus. For some tasks (*e.g.,* on the rotary pursuitmeter) psychologists report that the degree of initial hereditary determination seems to reach 90 percent, thus leaving little room for sociological variables to operate. Teams conducting research with infant tests have noted that Congoid babies in both Africa and the United States are more precocious in sensorimotor maturation than are Caucasoid babies in Africa, the United States, or Europe, and that this precocity lasts about three years, after which whites outperform the blacks. By adolescence, Negro subjects score significantly less well than whites on complex coordinator, rotary pursuitmeter, discrimination reaction, selective mathometer, and two-hand coordinator tasks. On the other hand, Chinese- and Japanese-American infants lag behind Caucasians in motor development but perform better than both whites and blacks on certain eye–hand coordination and dexterity tasks at later ages.

Many research workers believe that social, economic, educational, and attitudinal variables that might unequally influence minority groups are of little consequence in the psychomotor field. They point out, for instance, that Chinese, Jewish, Negro, and Puerto Rican children, when tested by members of their own cultures, show distinctive patterns of basic perceptual and motor abilities, and their particular skill profiles are unaffected by differences in socio-economic level. Malnutrition (*e.g.,* protein deficiency) is not believed to be a plausible explanation unless there has been severe deprivation during the perinatal period. According to much of the literature, blacks generally do better than whites on chemical taste tests, rhythmic discrimination, visual acuity, colour perception, and resistance to special optical illusions; Mongoloids show better taste sensitivity and less colour blindness than Caucasoids; and on certain physical-fitness tests male Afro-American athletes are not only superior to their white countrymen but their relative proficiency is inversely correlated with the degree of Caucasoid admixture.

Genetic behavioral differences among human populations—just as those of morphology—appear to be the rule rather than the exception. Tarahumara Indians far surpass other races in endurance at long-distance running contests; Andeans and Tibetans are superbly adapted to working at high altitudes; Eskimos excel on psychomotor tasks performed under low-temperature stress. It is plausible that the inherited factors underlying behavioral aptitudes and capacities have evolved from different selective pressures in different ecological niches. As is true for age and sex, however, hereditary and environmental variables are complexly intertwined in racial studies. Nevertheless, genetic determinants seem to be far more powerful in the etiology of original psychomotor aptitudes. It does not follow that learnability is weak. Quantitative experiments demonstate that heritabilities can be systematically altered by controlled practice; this is a theoretical discovery of broad implications for practical training programs. At the same time, it would appear that the hereditary control of several psychomotor abilities tends to be less pronounced at the end of training than at the beginning.

Other factors. A number of other personal characteristics have been found to be of significance in psychomotor behaviour. For instance: (1) speed scores in reaction-time tasks are positively correlated with body temperature in adults, one of the many indices of variation within the individual; (2) psychotics show longer reaction times and poorer tracking scores than do people of normal personality; (3) right-handed operators are favoured on the rotary pursuitmeter, while left-handed persons tend to do better on the complex coordinator; (4) left-handed people are more variable in finger-dexterity and paper-cutting skills and also are more prone to show signs of ambidexterity; (5) intelligence quotients (IQ) are weakly related to physical strength or endurance yet are strongly associated with performance in such psychomotor activities as running the 35-yard dash, balancing on one foot, discrimination reaction, rotary pursuit, and selective mathometry—these correlations are especially high when based on groups that comprise a full range of IQs (from retardates to college students); (6) typically one's body build (somatotype) is associated with his athletic skills—the best fencers, oarsmen, and basketball players, for example, tend to be tall and lean (ectomorphic); top swimmers, divers, and pole-vaulters are likely to be broad-shouldered and slim-hipped (mesomorphic); champion wrestlers, shot putters, and weight lifters are apt to be thick-trunked and short-limbed (endomorphic). While these genetically determined somatotypes do not guarantee athletic prowess, they definitely do favour success in certain sports rather than others. Similar considerations apply to vocal and instrumental musical aptitudes wherein unique combinations of such anatomical structures as lips, teeth, larynx, tongue, eyes, ears, hands, and arms can facilitate the attainment of virtuoso skill.

IQ correlations

In short, psychomotor abilities and learning underlie some of the most fundamental human activities, contributing to the full spectrum of work, play, creativity, love, and the very survival of individual and species. (C.E.N.)

Concept formation

Concept formation refers to the process by which one learns to sort his specific experiences into general rules or classes. One is observed to meet a given person, to lift a particular stone, and to drive a specific car. When he seems to *think* about things, however, he often appears to deal with classes; apparently he knows that stones (in general) sink, that automobiles (as a class) are powered by engines. He behaves as if he thinks of them in a general sense beyond any particular stone or automobile. Awareness of such classes can help guide behaviour in new situations. Thus two people in a bakery may never have met before; yet, if one can be classified as customer and the other as clerk, they tend to behave appropriately. Similarly, many people seem able to drive almost any automobile by knowing about automobiles in general.

Concepts as rules

Concept formation is a term used to describe how one learns to form classes; conceptual thinking refers to one's subjective manipulations of those abstract classes. A concept is a rule that may be applied to decide if a particular object falls into a certain class. The concept "citizen of the United States" refers to such a decision rule, meaning any person who was born in U.S. territory or who is a child of a U.S. citizen or who has been legally naturalized. The rule suggests questions to ask in checking the citizenship of any particular individual. As most concepts do, it rests on other concepts; "U.S. citizen" is defined in terms of the concepts "child" and "territory." Many scientific or mathematical concepts cannot be understood until the terms in which they are defined have been grasped. Concept formation builds on itself.

Conceptual classification may be contrasted with another type of classification behaviour called discrimination learning. In discrimination learning, objects are classified on the basis of directly perceived properties such as physical size or shape. The usual explanation for discrimination learning is that the sensory features of any stimulus are matched to what is already remembered of these features, and that the learner's response becomes associated with them. The response thus classifies the stimulus. In discrimination learning subjective representations of immediate and past stimuli seem directly to indicate concrete, physical features (in contrast to the more abstact nature of concept formation). When a stimulus is perceived to match several different past experiences, the response may be a compromise; an object need not bear an all-or-none relation to a set of others in discrimination learning—for example, there is no absolute distinction between tall and short people.

While human beings popularly are called abstract thinkers, many of the classifications people make clearly seem to be concrete discriminations. Indeed, people may use the same term either in a discriminative or conceptual way. A child uses the term policeman in discriminating a man in distinctive uniform, while a lawyer may have a concept of a civil servant charged with enforcing criminal codes.

In practice, people seem to think in many ways that combine abstractness and concreteness. They also may blend class membership with assignment along a scale; *e.g.,* such concepts as leadership, an abstract quality that people are said to exhibit in varying degrees. The same applies to vivacity, avarice, and other personality classifications.

People seem to develop more complex sets of classes than do other animals, but this need not mean that human modes of learning are qualitatively unique. It may be that all animals have the same basic biochemical machinery for learning, but that human animals exhibit it in greater variety. Yet, it seems no more appropriate to account for human concept formation in terms of discrimination learning alone than it does to reduce the functions of a piston engine to chemical reactions.

EXPERIMENTAL STUDIES

Since careful observation of informal, everyday behaviour is difficult, most evidence about human concept formation comes from laboratory subjects. For example, each subject is asked to learn a rule for classifying geometric figures (see the Table).

Geometric Patterns of the Type Used in Studying Concept Formation

object number	size	colour	shape
1	big	green	triangle
2	big	green	circle
3	big	red	triangle
4	big	red	circle
5	small	green	triangle
6	small	green	circle
7	small	red	triangle
8	small	red	circle

The experimenter may concoct the rule that all green objects are called GEK. The subject is shown some of the figures, told which are named GEK, and asked to infer the rule or to apply it to other figures. This is roughly akin to teaching a young child to identify a class of barking animals with the name DOG. In both cases a general rule is derived from specific examples.

The problem of discovering that GEK = GREEN is almost trivial when four GEK and four NOT GEK figures are presented at once; it becomes surprisingly difficult if they are presented one at a time and need to be remembered. When two concepts are to be learned together (*e.g.,* JIG = TRIANGLE and GEK = GREEN) memory for each concept tends to be mixed, and it becomes a formidable task to solve either problem. This is evidence that short-term memory functions in concept learning, and that it often is a limiting factor in performance. Efficiency in more complex concept learning often depends on providing enough time for examples of a rule to be fixed in memory for longer periods.

Concept identification

Most such experiments involve very simple rules. They properly concern concept identification (rather than formation) when the learner is asked to recognize rules he already knows. Adult subjects tend to focus on one stimulus attribute after another (*e.g.,* shape or colour) until the answer is found. (This is problem solving with a minimum of thinking; they simply keep guessing until they are right.) People tend to avoid repeating errors but seem to make surprisingly little use of very recent, short-term experience.

Most people are orderly in trying out attributes, first considering such striking features as size, shape, and colour, only later turning to the more abstract (*e.g.,* number of similar figures, or equilateral versus isosceles triangles). This characteristic progression is reminiscent of the quantitative distinction between discrimination learning (relatively concrete) and concept formation (more abstract); there seems to be no sharp, qualitative division. If pairs of arrays are shown in which the same geometric figure is repeated (the rule being that all GEK arrays have exactly 10 figures), people first are apt to react to directly perceivable characteristics (*e.g.,* the extent to which the figures fill the space). They are likely to discriminate most grossly different NOT GEK patterns in this way quickly but to be troubled by arrays of 9 or eleven. Eventually most adults should discover a solution by counting figures. Ordinarily more difficult, higher-order abstractions (*e.g.,* GEK arrays have an even number of figures) become easy to learn if the distinction is directly perceivable; for example, if even-numbered arrays are drawn symmetrically and all others are not.

Concept learning; conjunctive and disjunctive rules

Study can shift from concept identification to concept learning by requiring combination of previously learned rules. A conjunctive concept (in which the rule is based on the joint presence of 2 or more features; *e.g.,* GEK patterns now are LARGE and GREEN) is fairly easy to learn when the common characteristics stand out. But learning a disjunctive rule (*e.g.,* GEK objects now are either LARGE or GREEN but not both) is quite difficult; there is no invariant, relatively concrete feature on which to rely.

Concept learning in adults may be understood as a two-step process: first the discovery of *which* attributes are relevant, then the discovery of *how* they are relevant. In the conjunctive illustration used here, the learner first is likely to notice that size and colour have something to do with the answer and then to determine what it is. This two-step interpretation presupposes that he already has learned rules for colour, size, shape, or similar dimensions.

In an example of what is called intradimensional shift, initially the subject learns that GEK = GREEN; then, without warning, the experimenter changes the rule to GEK = RED. The same attribute or dimension (colour) is still relevant, but the way in which it is used has been changed. In extradimensional shift the relevant dimension is changed (*e.g.,* from GEK = GREEN to GEK = TRIANGLE) but the classification of some objects does not change (GREEN TRIANGLE is a GEK under both rules). The relative ease with which a subject handles such problems suggests something about how he learns. If he tends to learn simply by associating GEK with specific figures without considering the selected attribute, then he should find extradimensional-shift problems easier, since only some of his associations need be relearned. But if he has learned stepwise in terms of relevant attributes (*e.g.,* to say "What is the colour? . . . Ah, *that* colour means it is GEK"), intradimensional shift should be easier, since only the *how* phase of the two-step process need be relearned.

College students tend to find intradimensional-shift problems easier, indicating that they are prone to use the two-step process. On the other hand, suppose a rat initially is rewarded when it runs into the right-hand side of a maze for food, then a change is made by rewarding entries to the left (intradimensional shift) or by rewarding entries to any brightly lighted alley regardless of location (extradimensional shift). The rat will perform best on the extradimensional-shift problem. Among children, performance depends substantially on age; preschool children are likely to do best with extradimensional shifts (as rats do) but beyond kindergarten age they tend to find the intradimensional shift easiest. Perhaps these differences are related to how children learn to apply language in problem solving.

Concepts as models of change

Concepts need not be defined as limited to simple classification but also can be interpreted as models or rules that reflect crucial possibilities for change. To take a simple case, an adult is not apt to think that the volume of water changes when it is poured into a container of different shape. (Young children may claim that it does.) In the adult's concept, volume is not synonymous with the shape of a container but is based on a model of how fluids behave. The concept of "heat" does not serve simply to sort objects as hot or cold; it implies a rule or model of energy transfer that can be used widely (*e.g.,* to explain how water boils at a lower temperature when pressure drops). Concepts can be understood as models on which to decide if particular changes will have significant effects. This also implies classification (of sets of equivalent situations), but the way people learn to make this sort of classification may be quite different from the processes described so far.

AGE AND CONCEPTUAL BEHAVIOUR

Piaget's observations. The provocative clinical observations of Swiss psychologist Jean Piaget (1896–1980) have initiated considerable study of how young children learn concepts for coping with their physical surroundings. As models for defining feasible change, concepts are at least as important in such contexts as they are for classification. Piaget stressed that an infant normally first must learn that he is a thing apart from his external environment; next that he must form enough concepts of physical invariances (*e.g.,* that objects fall) to let him explore his world. Later in the preschool period the child typically grasps the concept of spatial localization (of objects separated in space). Piaget characterized the child during this period as classifying objects only on the basis of perceptually attractive, concrete physical features (in agreement with laboratory studies of intradimensional and extradimensional shift).

He and others who used his methods reported that preschool children are apt to explain external change in terms of their own needs; *e.g.,* a four-year-old is likely to say that a cloud moves "because the sun is in my eyes." Other distinctions between cause and effect emerge among children in early primary grades, who may say a moving cloud "wants to hide the sun." In later primary grades, volitional and passive movement usually become conceptually distinct.

Ability to deal analytically with objects apart from their immediate perceptual characteristics is reported typically to become effective only just before adolescence. At that time the concept of hierarchies of subclasses within more general classes commonly develops. A normal child of eleven applies the properties of all living things to the class called birds.

The role of learning

Progressive use of abstract concepts seems to reflect both maturation and learning. Given proper information, by the age of six many children display impressive concept-forming abilities. They ordinarily have considerable linguistic competence, using (though often not being able to explain) such abstract transformations as present and past tense. Rules of formal logic can be taught in the elementary grades; the so-called new mathematics reflects efforts to improve the order in which such concepts are introduced.

The role of instruction in concept formation remains poorly understood, yet practically all cultural heritage is explicitly taught. How many people would develop the concept of number (let alone that of odd and even) if left to themselves? Human societies have existed for thousands of years without these concepts. Better knowledge of how to instruct and of the role of imitation in transmitting cultural concepts is needed.

Aging. It is generally reported that potential for learning new abstractions decreases in old age. In such extreme cases as senility, severe alcoholism, or brain injury, the deficit is dramatic. Much less is known, however, about changes in conceptual ability during the active period of adult life, and what evidence there is is conflicting. Perhaps such adults are too busy to serve as research subjects.

People deemed gifted as children tend to retain superior ability into their later years in grasping new abstractions. However, among more typical people, little correlation is found between conceptual ability evaluated in the early teens and the same ability measured ten or more years later. Very gifted youngsters are likely to be given special scholastic challenges and opportunities that may enhance their skill in abstract thinking.

In such abstract pursuits as pure mathematics or theoretical physics there is a tendency for creative scientists and writers to be most productive in their late 20s and early 30s, but there are many exceptions. Among people in general, there probably is a slight decrease in ability to form new concepts starting from the late 20s. At the same time, as people get older they have more learned concepts that they can apply to a problem, so the net change in ability is hard to predict. Deterioration in learning new concepts is likely to be more rapid past 60, its severity varying markedly from person to person. Deterioration may be associated with illness or injury rather than with mere age in years.

LANGUAGE

Language, as a system of symbols abstracted from experience, provides an important vehicle for thinking. Some theorists treat linguistic concept formation as being a complex type of discrimination learning. The U.S. psychologist B.F. Skinner, for example, held that linguistic concept formation is based on the same principles that describe how a rat learns to push a bar in response to a specific signal. This seems to account for name learning (*e.g.,* that some objects are called horse and others dog). Other aspects of language learning, however, do not seem to fit Skinner's discrimination-learning model; its adequacy in explaining how one learns the concept of grammatically equivalent sentences has been challenged. Considering concepts as specifications of feasible transformations, sentences are equivalent if one can be derived from another by allowable change (*e.g.,* from active to passive tense). It is hard to see how learning to handle transformations could be based on the learning of primitive discriminations.

Genetic theories

Another explanation of language acquisition favoured by some linguists and biologists lays less stress on learning. It could be that humans are genetically prepared to acquire some language at an early age, much as some birds show readiness to learn any song pattern to which they are exposed when they reach a certain age. In humans, this period seems to stretch from about age one to six. If this explanation holds, all human languages should obey constraints established by the linguistic limits of genetic endowment. Should language prove to be a relatively independent biological function, the high linguistic competence of many young children with poor ability for abstract reasoning would appear less paradoxical.

Perhaps some rudimentary bases for language among other animals can be learned by methods appropriate for discrimination learning; even very young children are among the best discrimination learners in the animal kingdom. Once basic linguistic discriminations have been grasped, they can be used as tools with which the remainder of any language is learned. This suggests a theoretical position that falls between strictly cognitive accounts of language learning and Skinner's ideas. Nevertheless, biological bases for language learning remain to be identified and incorporated in such theories.

At any rate, such an array of theories is a sign of how little is known of the way people learn the concepts of a language.

CONCEPT FORMATION IN ANIMALS

Rats learn to enter lighted or unlighted alleys to get food, and goldfish can be taught to swim toward or away from an object. In such discrimination learning the animal is said to associate a physical property of the stimulus with its response, and with some contingency of reward or punishment. Thus, while a dog can be trained to come when called, it need not mean that he knows his name in the same sense that a man apparently does. But how can one prove that the dog does not know his name, or even that another person has a deeper concept of his name?

Most animals show classification behaviour that clearly seems to be discrimination learning. A crow will respond to the danger call of a bird of another species, but, only if that call physically resembles the crow's. Chimpanzees, however, have been observed using sticks as primitive tools; they behave as if they have a concept of things that extend reach. On considerable evidence of this sort, many are reluctant to say flatly that the animals are incapable of abstract thinking.

The oddity problem

Most studies aimed at evidence of concept formation among laboratory animals have involved primates, although there are reports of abstract behaviour among such animals as dogs, dolphins, and pigs. Monkeys have been taught to solve the oddity problem: presented with two objects of one kind and one of another, they can be trained to select the discrepant one. This behaviour persists even for sets of objects that have never been presented to them before. The animals behave as if they grasp the general concept of similarity, an abstraction rather than a simple discrimination. With great effort chimpanzees have been taught to speak and to use correctly a very few words. A much more successful attempt has been made to teach a chimpanzee the sign language used by deaf people, gestures apparently being more appropriate to the anatomic structure of chimpanzees. (Human beings ordinarily seem more prepared to learn spoken language.) The chimpanzee learned to use the signs for hat, dog, food, yes, me (self), sorry, funny, go, come, and many others. This work was reported in 1967 by R.A. Gardner and B.T. Gardner.

CONCEPT FORMATION BY MACHINE

Computers can be programmed to process information and to develop classification rules (*e.g.,* they can play chess and make decisions about business or military problems). Essentially such devices are programmed to mimic the process of problem solving required of subjects in laboratory experiments on concept learning. In this sense, machines have formed concepts; but their functions remain relatively impoverished. Efficient linguistic behaviour has proven particularly difficult to produce in a machine, despite numerous attempts. Yet there is no evidence that human concept formation is based on any mode of handling information that in principle could not be built into a machine. It is almost an article of faith among many investigators that human thinking can be explained mechanistically in physiological terms, but the scientists themselves do not yet seem to have developed concepts adequate for producing machines that can approach the full range of human talent. (E.B.H.)

Transfer of training

Will one's knowledge of English help him learn German? Are skillful table-tennis (Ping-Pong) players generally good court-tennis players? Can a child learn to multiply if he does not known how to add? These questions represent the problems of transfer of training: the influence the learning of one skill has on the learning or performance of another.

KINDS OF TRANSFER

Basically three kinds of transfer can occur: positive, negative, and zero. The following examples from hypothetical experiments, purposely uncomplicated by distracting detail, illustrate each. Suppose a group of students learn a task, B, in 10 practice sessions. Another group of equivalent students, who previously had learned another task, A, is found to reach the same level of performance on task B in only five practice sessions. Since the average number of practice sessions required to learn B was reduced from 10 to five, transfer of training from task A to task B is said to be positive (10 − 5 = +5). Many successful training aids, such as those that simulate the cockpit of an airplane and that are applied to teach people how to use instruments for flying blind without leaving the ground, produce positive transfer; when students who have preliminary training in such trainers are compared to those who do not, those with preliminary training almost invariably require less practice in achieving the desired level of skill.

Positive, negative, and zero transfer

Sometimes the effect of transfer of training is to hamper effectiveness in subsequent activity. If after learning task A a group of people need 15 practice sessions to learn task B whereas only 10 sessions are required for those without any previous training in task A, then task A is said to lead to negative transfer of training on task B (10 − 15 = −5). Having learned to drive on the right side of the road often is observed to produce negative transfer for the tourist from Japan or continental Europe or North America when he is travelling in Great Britain, where cars are to be driven on the left-hand side of the road.

The degree to which transfer of training occurs between two different tasks is often minimal and may be so small that it is called zero transfer. If learning task B with or without previous training in task A requires 10 practice sessions, then the amount of transfer from one task to the other is said to be zero (10 − 10 = 0). Learning to knit Argyle socks is apt to produce zero transfer of training in learning to sing an operatic aria in French.

Although in contemporary psychology transfer of training is a distinct topic of investigation with its own experimental designs and procedures for measurement, its implications pervade practically all of psychology, from conditioning to personality development. Ivan P. Pavlov discovered that when a dog is conditioned to salivate in response to a sound wave of 1,000 cycles per second, it will also salivate if it is next exposed to a tone of 900 cycles per second, although typically the volume of saliva will be slightly reduced. In this case, transfer of training occurs between two similar auditory stimuli; in general, phenomena of this sort are called stimulus generalization. At the very root of modern theories of personality development is the assumption that what a person learns during his childhood will show a pervasive degree of transfer to his adult behaviour. In some cases stimulus generalization mediates this transfer. Some cases of excessive fears may have their origins in unpleasant experiences during early life.

EDUCATION AND TRANSFER

The experimental study of transfer of training has historical roots in problems of educational practice. Educators in Western countries at the end of the 19th century widely endorsed the doctrine of formal discipline, contending that psychological abilities, called "mental faculties" by such philosophers as Thomas Aquinas (1225–74), could be strengthened, like muscles, through exercise. By learning geometry, one was expected to improve his ability to reason; studying Latin was held to "strengthen" the so-called faculty of memory, and so on. Although what contemporary educators have demoted from the doctrine to the theory of formal discipline once seemed reasonable to many, experimental tests have refuted it. When the reasoning abilities of groups of mathematics students in secondary schools were compared with those of other equally talented students who had not had the same mathematical training, no differences in general logical effectiveness were observed between the groups.

An alternative theory of identical elements was proposed in which it was postulated that transfer between activities would take place only if they shared common elements or features. Thus it was predicted that one's training in addition would transfer to his ability to learn how to multiply. It was reasoned that both tasks share identical features, multiplication basically requiring a series of successive additions, and that both tasks demand the individual's concentration.

But the identical-elements formulation soon came under attack when experimental results suggested that one's understanding of general principles, rather than the presence of identical task elements, has substantial effects on transfer of training. In one notable experiment, two groups of boys practiced throwing darts at a target placed under about a foot of water. Only one group, however, was instructed about the principle that water bends (refracts) light. According to this principle, the apparent position of the target should vary with the depth of the water. When the target depth was reduced to four inches, the group that had been taught the general principle of refraction adjusted rapidly to the change and exhibited substantial positive transfer; the other boys showed comparative difficulty in learning to hit the target at the shallower level.

Theories of transfer

These formulations (formal discipline, identical elements, and general principles), when considered carefully, might be recognized as points of view rather than as rigorously specified theories that could lead to unequivocal predictions of the results of new experiments in transfer of training. For example, failure to demonstrate positive transfer between mathematical training and general reasoning ability could be attributed to ineffective teaching of mathematics; in such case, the results need not be interpreted as refuting the theory of formal discipline. If the then-traditional manner of teaching mathematics could be changed to emphasize logical thinking (rather than routinized application of formulas), it was argued that perhaps mathematical training could improve reasoning ability in general. Some theorists also suggested that the positive transfer observed to result when boys learned the principle of refraction was consistent with the hypothesis of identical elements; these theorists observed that a general principle may be considered an element common to many tasks. According to this line of reasoning, the group of boys who exhibited positive transfer with the shift to a new target depth shared the principle of refraction as an element in common with the previous task, along with those of aiming and throwing. By contrast, the youngsters who performed without the benefit of knowing about refraction were held to have gained positive transfer from throwing but to have suffered negative transfer as a result of aiming incorrectly.

EXPERIMENTAL ANALYSIS OF TRANSFER OF TRAINING

The indeterminate character of the broad theoretical formulations offered to account for transfer of training and the often unsuccessful ways in which they were applied to the practical problems of classroom teaching led some psychologists to retreat to the laboratory in the hope of identifying more clear-cut, fundamental processes in transfer of training. As a result, a number of different transfer-of-training phenomena were discovered, several of which may be reviewed as follows.

Stimulus and response similarity. The method of paired-associate learning, in which a person is asked to learn to associate one syllable or word with another (*e.g., complete–hot, safe–green, wild–soft*), encouraged the investigation of the influence of stimulus and response similarity on transfer of learning. Typically these pairs of verbal items are presented to the laboratory subject so that the first, or stimulus, member (*e.g., complete*) is exposed alone, followed after a short interval by the second, or response, member (*e.g., hot*). The subject's task is to respond to the stimulus term before the response term appears, as when an English-speaking student in learning French is supposed to respond to *le livre* with *the book.*

When two successive lists of paired associates are learned in which the stimulus elements are the same but the response terms are changed (*e.g., complete–hot* in the first list and *complete–new* in the second), negative transfer typically results. Apparently, in learning the second list the subject tends to respond to the stimulus term (*e.g., complete*) with the previously learned correct response term (*e.g., hot*), the result being interference with new learning to produce negative transfer. If he were learning the second list without having learned the first, the subject would not be so handicapped.

Another question concerns the sort of transfer that results when response terms are different and stimulus elements are similar but not identical; for example, *entire* is similar to *complete.* After one has learned *complete–hot,* the experimental evidence is that his ability to learn *entire–new* becomes definitely more difficult. Both *entire* and *complete* seem to have a tendency to evoke the response *hot* and to be incompatible with subsequently learning the association of *entire* with *new.* The principle that appears to operate in such situations is that the greater the similarity in stimulus elements, the greater the degree of negative transfer.

The influence of response (rather than stimulus) similarity on transfer of training is more complex; in paired-associate learning, the subject needs to learn the response term of each pair (response learning) and then to remember that it is linked with its appropriate stimulus partner (associative learning). When response terms are relatively difficult to learn (as in the case of unfamiliar or foreign words), the subject tends to profit considerably from learning the first list. But when response terms already have been learned (or are easy to learn), little if any positive transfer is likely to occur. The degree of transfer between lists that contain similar response terms depends both on how similar they are and on their level of difficulty; increasing the similarity between response terms is most likely to increase positive transfer when the response terms are relatively difficult to learn.

Although attempts have been made to formulate an all-embracing theory that would account for the effects of similarity among paired associates on transfer of training, a major obstacle that has prevented fully satisfying results is that the degree of positive or negative transfer is typically a product of many interacting influences beyond those of stimulus and response similarity. For example, the amount of training that the subject receives also has significant effects on transfer. When initial training is given on a simulated task (*e.g.,* learning to operate a set of dummy controls in preparation for a second task of acquiring a complicated skill, such as flying an airplane), negative transfer effects frequently appear during the initial stages of learning the second task and then give way with further training to generally positive transfer effects.

Obstacles to a unified theory of similarity in transfer

Another stumbling block in developing theoretical explanations has to do with the meaning of the central concept of similarity. In such experiments as those in which the salivary reflex is conditioned to different auditory stimuli, similarity is measured in terms of physical stimulus properties (*e.g.,* pitch or loudness); in other studies, as in paired-associate learning, similarity typically is expressed in terms of verbal meaning. In neither case has a universally adopted method yet been devised to measure similarity in a reliable and precise way; perhaps none can be, simply since there are so many different aspects of physical and linguistic or semantic similarity. Despite these difficulties, efforts to analyze transfer experimentally in terms of the properties of stimulus and response events have been productive in identifying conditions that can be varied to alter the direction and the degree of transfer of training.

Retroactive and proactive inhibition. Closely related to stimulus and response similarity are phenomena called retroactive inhibition and proactive inhibition; these demonstrate how forgetting seems to result from interfering activities.

In a study of retroactive inhibition, both the experimental and control groups of people learn task A (for example, a list of adjectives) and are tested for their ability to recall A after a specified time interval. The groups differ in what they are asked to do during the interval; the experimental group learns a similar task B (say, another list of words), while the control group is assigned some unrelated activity (for example, naming a series of coloured chips) designed to prevent them from rehearsing task A. The results of numerous studies of retroactive inhibition show that the experimental subjects typically are poorest in recalling information from task A. The interpolated activity, particularly a comparable one such as memorizing a second list of adjectives, apparently interferes with one's ability to recall words from the first list. Habit competition, or what is sometimes called interference, between the items of the original and the interpolated word lists at the time of recall is considered to be one of the major sources of the negative transfer exhibited in retroactive inhibition.

Interference theory

Experimental designs for demonstrating proactive inhibition differ from those used for showing retroactive inhibition in that the experimental group learns task B before, instead of after, task A. Whereas B was a task that was interpolated between the learning and the recall of task A in the retroactive inhibition study, B is a task that precedes the learning of task A in the proactive inhibition study. To evaluate the effects on the experimental subjects of their having learned B prior to A, the control people are instructed to relax during the time the experimental group is learning B. Typically an experimental subject's ability to recall from task A is inferior to that of a control person, the degree of inferiority depending in part on how similar the two tasks are; the greater the similarity, the poorer the recall tends to be. Although proactive inhibition, so called to indicate that it acts forward from the first-learned task to the second, produces appreciably less forgetting than does retroactive inhibition, they both support the theory that interference can produce forgetting (see MEMORY).

Stimulus predifferentiation. Educational films can be considered as everyday examples of stimulus predifferentiation, in which the individual gets preliminary information to be used in subsequent learning. The student who sees a film describing the various parts of a microscope is likely to be better prepared to learn the requisite skills when confronted with the instrument itself. In laboratory studies of stimulus predifferentiation, the subject is given experience with a particular stimulus situation ahead of time; later he is asked to learn new responses in the same situation. In one illustrative study, subjects first practiced labelling four different lights and then later were asked to learn to press selectively one of four switches, each connected to one light. The rate at which they learned the appropriate pressing reactions was related to how well they had learned to label the lights.

The results of a large number of experiments covering a variety of stimulus predifferentiation techniques suggest that when a learner has an opportunity to become generally acquainted with an environment, he retains some information about its different components that prepares him for learning to make new responses to them. Various explanations have been offered to account for this facilitation; some investigators suggest that the process of labelling enhances the distinctiveness of environmental stimuli for the labeller; others hold that perceptual acquaintance can more sharply differentiate an environment into its component parts for the perceiver or that it may encourage appropriate responses of observing or attending. Nevertheless, no single process has been identified as fundamental in stimulus predifferentiation. Perhaps a number of these processes operate in different combinations from one stimulus-predifferentiation transfer experiment to another, each process representing a different method by which a learner can become familiar with the details of his environment.

Transposition. Another phenomenon that has received considerable attention in theories of transfer of training is called transposition. An initial report of transposition came from a study in which chickens were trained by rewards to respond to the darker of two gray squares. After this discrimination task was learned, the chickens were shown the originally rewarded gray square along with one that was still darker. They seemed to prefer the darkest gray to the square that had been previously rewarded. This finding was interpreted to support the hypothesis that the birds had initially learned to respond to a relationship (what a human being would call the concept "darker") and that this response to a relationship had been transposed or transferred to the new discrimination. This relational interpretation later was challenged by theorists who offered a formulation to show, on the basis of principles of stimulus generalization, how a response to a relational stimulus could be explained by assuming that organisms do indeed respond to the absolute properties of the stimuli. Both explanations were found to be too simple for the variety of findings obtained with transposition studies. As a result, the interest of many investigators shifted away from demonstrating the relative merits of absolute versus relational interpretations to identifying conditions that seem to influence transposition behaviour. Within this context, newer, more sophisticated formulations have been proposed that consider both the absolute and relational characteristics of the stimuli in transposition studies.

Transfer among animals

Learning to learn. When people are asked to learn successive lists of words, their performance tends to improve from one task to another so that much less time is commonly required to learn, say, the tenth list than was needed for mastering the first list. This improvement suggests that information beyond the specific content of lists of words is also learned. It would seem as if the subjects are learning how to learn; that is, they seem to be acquiring learning sets, or expectancies, that transfer from list to list to produce continually improving performance.

Some of the most intensive work on learning sets has been carried out with monkeys that were learning how to solve several hundred discrimination problems in succession. In each problem, the monkey learned which one of two objects (for example, a bottle cap and a cookie cutter) consistently contained a piece of food. Although the solution of each successive problem required the animals to discriminate between two previously unfamiliar objects, performance tended to improve on successive tasks; the monkeys made increasing numbers of correct choices on the second trial of each problem as the process continued. Manifestly there was no cue to indicate the correct choice on the first trial of any specific problem. If the animal responded correctly on the first trial, then on the second trial it would only have to choose the same object to be correct thereafter; if the monkey made an error on the first trial, then the other object would inexorably be the one that should be chosen next. During their efforts to solve the first few problems the monkeys were correct approximately half the time on the second attempt to solve each problem. This success increased to an average of 80 percent correct after each animal had solved 100 problems, to 88 percent after 200 correct solutions, and eventually to 95 percent after 300. Thus, after a long series of separate tasks, all of the same type, the monkey's first response to the next problem usually provided sufficient information for the animal to make the correct choice.

Since each of the successive discrimination problems was different, what actually was being transferred from problem to problem? In these discrimination problems, the monkeys seemed to have several items of information to learn in addition to which one of the two objects contained the rewarding bit of food. The animals apparently had to learn to pay attention to that part of their environment where the objects were placed. To make the correct choice, it would seem that a monkey would have to learn to abandon any preference it might exhibit for objects on either the left or the right; indeed, the animals usually did show such preferences. (The correct object was shifted from side to side in a random sequence to control for these preferences.) Ostensibly, the monkeys also had to learn that one object consistently contained food while the other was always empty. Although these learning sets by themselves would not serve to identify the correct object in each new discrimination problem, it seems likely that they could help the animal locate the reward very rapidly by eliminating initially unprofitable responses.

Reversal learning. In reversal learning, the individual first learns to make a discrimination, such as choosing a black object in a black–white discrimination problem, and then is supposed to learn to reverse his choice—*i.e.,* to choose the white object. Such reversals tend to be difficult for most learners since there are negative transfer effects; *e.g.,* the individual tends to persist in responding to the black object that was originally correct. Eventually, however, one's tendency to make the originally learned selection typically becomes weaker, and he makes the competing response (*e.g.,* to white) more frequently until a point is reached where it is almost consistently evoked. Reversal learning can be accomplished very rapidly when a laboratory animal, such as a monkey, is presented with a series of reversal-learning problems in which the same sequence of shifts is repeated (as when black is initially correct, then white, then black, then white, and so on). After extended reversal training, some animals are able to make the next reversal in the sequence in one trial. They behave as if they have mastered the abstract concept of alternation or of regular sequence.

Evolution and transfer

The speed with which representatives of a given species of animal, including human beings, can be taught to make a reversal of this kind seems to be related to the place biologists assign them in a hierarchy of evolutionary development. On first being exposed to a reversal-learning problem, normally competent adult humans who can use language are likely to achieve a solution with great rapidity. Monkeys can learn to perform equally well after a relatively longer series of reversal-learning tasks; but isopods such as pill bugs or sow bugs, small relatives of crabs and shrimp, have such primitive brains that they seem to be unable to improve their performance at all during a series of reversal-learning tasks.

DEVELOPMENTAL PROCESSES AND TRANSFER

The manner in which a problem is learned seems to have an effect on what is transferred. This conclusion is supported by experiments in which comparisons are made of the relative ease with which children of different ages execute reversal and so-called extradimensional shifts (see above *Concept formation*). In performing both kinds of shift, experimental subjects learn two successive discriminations between two pairs of objects that vary simultaneously in two aspects or dimensions—*e.g.,* white triangle versus black square, and black triangle versus white square. In training subjects initially, discrimination of only one dimension (for example, black–white) is made relevant, with the child's selection of one of the cues (for example, white) being rewarded, while the other (black) is incorrect. After they have learned this, the children are shifted to the second discrimination. In the case of a reversal shift, the same stimulus dimension (black–white) remains relevant, but the child is now to learn to reverse his initial choice; black choices are now rewarded, and white selections become incorrect. For an extradimensional shift, the initially irrelevant dimension (square–triangle) is given relevance by rewarding selection of one of its alternatives and by failing to reward choices for the other.

The relative ease with which human beings learn to make extradimensional and reversal shifts is related to how old they are. Reversal shifts are relatively difficult for young children to learn and are relatively easy for adults to master. As people gain maturity, the relative ease with which they execute a reversal shift tends to increase in comparison with their ability to achieve an extradimensional shift.

Age and transfer

Explanations for these developmental changes seem to be found in the manner in which the individual solves a discrimination problem. Very young children and laboratory animals tend to learn simple habits when faced with a discrimination problem for the first time; for example, they are most likely to learn simply to approach black objects and to avoid white. Reversal shift is often extremely difficult for them, and negative transfer effects are substantial. Subjects who primarily learn simple habits are faced with the task of eliminating one habit (*e.g.,* to choose black) that has been rewarded and then of developing another habit (*e.g.,* to choose white) that previously has not been rewarded.

Human adults, on the other hand, generally find a reversal shift relatively easy; they do not behave as if they simply associate their choices to the relevant stimuli (*e.g.,* white and black) but instead appear symbolically (or conceptually) to react to both of them in terms of their common characteristic (brightness). A similar kind of symbolic or logical response is appropriate in solving reversal-shift problems; since the relevant dimension remains the same, this kind of shift tends to be easier to make than is one involving extradimensional shift, which requires the individual to switch to a new symbolic response (*e.g.,* from brightness to size). In short, when they respond concretely, learners favour their potentials for achieving extradimensional transfer; those who tend to respond symbolically enhance the probability for reversal transfer.

Whatever the validity to be found in theoretical explanations of this sort, review of how transfer phenomena may be influenced suggests that no single principle or simple theory thus far put forward accounts for all of the observed data. Instead, the evidence is that several interacting processes underlie transfer of training and that their relative influence depends both on the nature of the tasks between which transfer takes place as well as on the characteristics of the learning organism. If one seeks to control the degree of transfer, as one does in educational settings, it seems useful to analyze transfer behaviour in terms of a number of component processes—*e.g.,* stimulus and response similarity, stimulus predifferentiation and response learning, and the symbolic abilities of the learner.

THE PHYSIOLOGY OF TRANSFER OF TRAINING

Although available evidence for a physiological basis of transfer of training is limited, some impressive data already are recorded. Some central (brain and spinal-cord) mechanisms seem to control transfer of training. A long-established transfer phenomenon is cross education, in which there is positive transfer of a skill learned with one part of the body to another, untrained part. For example, a person who learns to throw a dart with his preferred hand exhibits positive transfer to his non-preferred hand. Since different muscles are involved in the equivalent action of opposite limbs, positive transfer resulting from cross education cannot be attributed simply to common muscular movements; instead it would seem that cross education depends on central processes that control the actions of both limbs.

Cross education

Among highly evolved animals, transfer of training between limbs from opposite sides of the body evidently is mediated through a massive system of neural fibres, known as the corpus callosum, that connects the two hemispheres of the brain. One of the many ways in which the validity of this principle may be demonstrated is first to train blindfolded cats to discriminate with one paw between two different pedals (by feeling raised horizontal lines on one pedal and by detecting raised vertical lines on the other). Since each eye sends some of its nerve impulses to both hemispheres of the cat's brain while each paw only directs impulses to the hemisphere of the brain on the same side of the animal's body, this procedure feeds the sensory information to just one hemisphere. After learning to make the discrimination with one paw (*e.g.,* reward being given only for the pedal with the horizontal pattern), a cat that is confronted with making the same discrimination with the other front paw, which has its connections with the ostensibly "untrained" brain hemisphere, will nevertheless exhibit positive transfer. Indeed, even when the corpus callosum is surgically severed immediately after learning (to "disconnect" the two hemispheres), positive transfer will take place from one front paw to the other; manifestly, transfer of training takes place between connected hemispheres while the animal is learning. If the cat's corpus callosum is severed before it initially learns to discriminate the two pedals, however, no transfer occurs between the animal's limbs; the untrained paw fails to exhibit any benefit from what has been learned with the other paw. In other words, by severing the cat's corpus callosum, the surgeon splits the brain into two independently functioning units. The same kinds of behaviour are observable among other split-brain animals, including chimpanzees and people.

The physiological foundations of transfer of training are not limited merely to the anatomical considerations of the central nervous system. To better understand how physiological processes mediate transfer of training means also to be able to specify more fully the anatomic, electrical, and chemical basis of learning in general, a goal that remains incompletely achieved. Many physiologists and psychologists hold that the search for the neurophysiological foundations of learning can be pursued most profitably by measuring physical and chemical changes that influence the transmission of nerve impulses. It has long been established that chemical changes are part of the process of neural transmission; and it is widely agreed that, in some way, biochemical activities also are responsible for all forms of learning, including transfer of training.

One popular theory in the 1960s was that learning and remembering depend on changes in the molecular structure of such chemicals as ribonucleic acid (RNA) and peptides that are incorporated in the cells of the body, including nerve cells. Some researchers have theorized that memory traces are physically coded within the molecules of cells.

Reports of experiments have been published offering evidence that skills have been transferred from one individual to another by injecting materials taken from the brains (or even other parts of the body) of trained animals into the bodies of untrained organisms (*e.g.,* flatworms, rats, hamsters). These reports have encouraged many to hope that someday one might be able to learn a foreign language, for example, by simply taking a pill instead of through the usual time-consuming practice. Subsequent efforts to repeat such experiments sometimes have given positive results but more often have yielded no evidence of chemical transfer of training from one individual to the next. In view of such inconsistent findings, this question became a matter of considerable controversy. Many investigators seemed inclined to dismiss the notion that organisms can learn by swallowing chemicals or through injection as another of those oversimplified interpretations that continue to be offered in efforts to account for complex psychophysiological phenomena. (H.H.K.)

BIBLIOGRAPHY

General works: J.F. HALL, *The Psychology of Learning* (1966), provides a comprehensive account of the empirical data that theories of learning are designed to integrate. E.R. HILGARD and G.H. BOWER, *Theories of Learning*, 3rd ed. (1966), is the standard reference on this subject. G.A. KIMBLE (ed.), *Hilgard and Marquis' Conditioning and Learning*, 2nd ed. (1961), describes theoretical issues as they apply to simple learning. The separate articles reproduced in KIMBLE, *Foundations of Conditioning and Learning* (1967), develop some of these arguments in more detail and also present synopses of the theoretical positions of Hull, Guthrie, and Tolman. G.A. KIMBLE and N. GARMEZY, *Principles of General Psychology*, 3rd ed. (1968), is a general textbook that covers several theoretical issues and should be particularly useful on the topics of memory and retrieval.

Psychomotor learning: A.L. IRION, "A Brief History of Research on the Acquisition of Skill," in E.A. BILODEAU (ed.), *Acquisition of Skill* (1966), an excellent historical survey; R.S. WOODWORTH and H. SCHLOSBERG, *Experimental Psychology*, rev. ed. (1954), one of the best standard reference works; C.E. NOBLE, "S-O-R and the Psychology of Human Learning," *Psychol. Rep.*, 18:923–943 (1966), a brief introduction to human learning; A.W. MELTON (ed.), *Apparatus Tests* (1947), a classic volume on psychomotor devices used in aviation psychology. Periodic reviews of the psychomotor field include: E.A. and I.MCD. BILODEAU, "Motor-Skills Learning," *A. Rev. Psychol.* 12:243–280 (1961); J.A. ADAMS, "Motor Skills," *ibid.*, 15:181–202 (1964); and C.E. NOBLE, "The Learning of Psychomotor Skills," *ibid.*, 19:203–250 (1968). C.L. HULL, *Principles of Behavior* (1943); and K.W. SPENCE, *Behavior Theory and Conditioning* (1956), are two of the most influential books on general behaviour theory from the reinforcement viewpoint. C.E. NOBLE, "A Theory of Psychomotor Skill: Derivation and Data," *Psychonomic Sci.* 21:344 (1970), makes a specific application to psychomotor learning. Two volumes presenting detailed information-processing analyses of skill are P.M. FITTS and M.I. POSNER, *Human Performance* (1967); and A.T. WELFORD, *Fundamentals of Skill* (1968). E.A. BILODEAU (ed.), *Principles of Skill Acquisition* (1969), provides an eclectic, simplified treatment of current topics by several authors. R.N. SINGER, *Motor Learning and Human Performance* (1968), is oriented mainly toward athletic proficiency and physical education. Articles of specialized interest, as indicated by their titles, are J.A. ADAMS, "Response Feedback and Learning," *Psychol. Bull.*, 70:486–504 (1968); C.E. NOBLE, "Acquisition of Pursuit Tracking Skill Under Extended Training As a Joint Function of Sex and Initial Ability," *J. Exp. Psychol.*, 86:360–373 (1970); and R.B. PAYNE, "Functional Properties of Supplementary Feedback Stimuli," *J. Motor Behav.*, 2:37–43 (1970).

Concept formation: L.E. BOURNE, *Human Conceptual Behavior* (1966), a well-organized review of laboratory studies; J.S. BRUNER, J.J. GOODNOW, and G.A. AUSTIN, *A Study of Thinking* (1956), a classic, well-written description of some experiments in the field; J.H. FLAVELL, *The Developmental Psychology of Jean Piaget* (1963), a competent explanation of Piaget's controversial and provocative ideas; E.B. HUNT, *Concept Learning: An Information Processing Problem* (1962), a comprehensive review and theoretical presentation of major approaches; B.F. SKINNER, *Verbal Behavior* (1957), a challenging proposal for applying laboratory psychology.

Transfer of training: Introductory psychology texts that discuss transfer of training as an experimental phenomenon and its implications for a wide variety of behaviour are H.H. KENDLER, *Basic Psychology*, 2nd ed. (1968); and H.H. and T.S. KENDLER, *Basic Psychology: Brief Edition* (1971). H.C. ELLIS, *The Transfer of Learning* (1965), presents a general analysis of transfer and includes reprints of important journal articles on the topic. Undergraduate texts that review theories and experimental evidence concerning transfer of training are J. DEESE and S.H. HULSE, *The Psychology of Learning*, 3rd ed. (1967); and J.F. HALL, *The Psychology of Learning* (1966). H.W. REESE, *The Perception of Stimulus Relations: Discrimination Learning and Transposition* (1968); and D.A. RILEY, *Discrimination Learning* (1968), are accounts of how transfer of training influences discrimination learning.

Lebanon

The Republic of Lebanon (al-Jumhūrīyah al-Lubnānīyah), a predominantly mountainous country of great scenic beauty, is an Arab republic situated on the eastern shore of the Mediterranean Sea. Consisting of a narrow strip of land about 135 miles (215 kilometres) long from north to south and 20 to 55 miles wide from east to west, it is bounded to the north and east by Syria and to the south by Israel. With an area of 3,950 square miles (10,230 square kilometres), Lebanon is one of the world's smaller sovereign states. The capital is Beirut.

Though Lebanon, particularly its coastal region, was the site of some of the oldest human settlements in the world—the Phoenician ports of Tyre (modern Ṣūr), Sidon (Ṣaydā), and Byblos (Jubayl) were dominant centres of trade and culture in the 3rd millennium BC—it was not until 1920 that the contemporary state came into being. In that year France, which administered Lebanon as a League of Nations mandate, established the state of Greater Lebanon. Lebanon then became a republic in 1926 and achieved independence in 1943.

As an Arab republic, Lebanon shares many of the cultural characteristics of the Arab world, yet it has attributes that differentiate it from many of its Arab neighbours. Its rugged, mountainous terrain has served throughout history as an asylum for diverse religious and ethnic groups and for political dissidents. Lebanon is one of the most densely populated countries in the Mediterranean area. It has one of the highest rates of literacy. Although its prosperity is unevenly distributed, having bypassed large segments of its population, wealth and privilege appear to be evenly distributed among its middle-income group. Notwithstanding its meagre natural resources, Lebanon long managed to serve as a busy commercial and cultural centre for the Middle East.

This outward image of vitality and growth nevertheless disguised serious problems. Not only did Lebanon have to grapple with internal problems of social and economic organization, but also it had to struggle to define its position in relation to Israel, to its Arab neighbours, and to Palestinian refugees living in Lebanon. The Lebanese pluralistic communal structure eventually collapsed under the pressures of this struggle. Communal rivalries over political power became so exacerbated by the complex issues that arose from the Palestinian question that a breakdown of the governmental system resulted from an extremely damaging civil war that began in 1975.

This article is divided into the following sections:

Physical and human geography

THE LAND

Relief. As in any mountainous region, the physical geography of Lebanon is extremely complex and varied. Landforms, climate, soils, and vegetation undergo some sharp and striking changes within short distances. Four distinct physiographic regions may be distinguished: a narrow coastal plain along the Mediterranean Sea, the Lebanon Mountains (Jabal Lubnān), al-Biqāʿ (Beqaa) valley, and the Anti-Lebanon and Hermon ranges running parallel to the coastal mountains.

Physiographic regions

The coastal plain is narrow and discontinuous, almost disappearing in places. It is formed of river-deposited alluvium and marine sediments, which alternate suddenly with rocky beaches and sandy bays, and is generally fertile. In the far north it expands to form the ʿAkkār Plain.

The snowcapped Lebanon Mountains are the most prominent feature of the country's landscape. The range, rising steeply from the coast, forms a ridge of limestone and sandstone, cut by narrow and deep gorges. It is approximately 100 miles long and varies in width from 35 to six miles. Its maximum elevation is at Qurnat as-Sawdā' (10,138 feet [3,090 metres]) in the north, where the renowned cedars of Lebanon grow in the shadow of the peak. The range then gradually slopes to the south, rising again to a second peak, Jabal Ṣannīn, northeast of Beirut. To the south the range gives way to the hills of Galilee, which are lower. The limestone composition of the mountains provides a relatively poor topsoil. The lower and middle slopes, however, are intensively cultivated, the terraced hills standing as a scenic relic of the ingenious tillers of the past. On the coast and in the northern mountains reddish topsoils with a high clay content retain moisture and provide fertile land for agriculture, although they are subject to considerable erosion.

Al-Biqāʿ valley lies between the Lebanon Mountains in the west and the Anti-Lebanon Mountains in the east; its fertile soils consist of alluvial deposits from the mountains on either side. The valley, approximately 110 miles long

Al-Biqāʿ

MAP INDEX

Political subdivisions

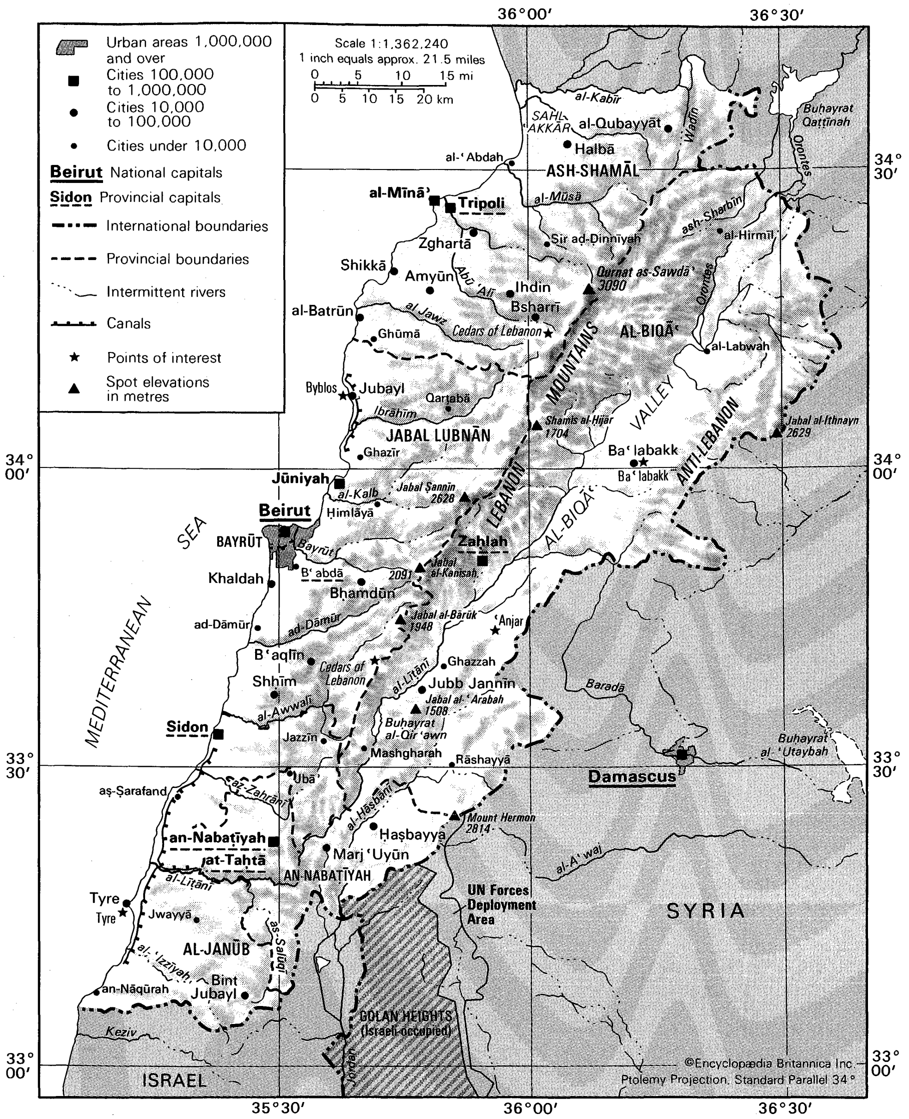
Bayrūt	33 56 N 35 30 E
Biqā', al-	34 00 N 36 10 E
Jabal Lubnān	33 50 N 35 40 E
Janūb, al-	33 20 N 35 18 E
Nabaṭīyah, an-	33 20 N 35 35 E
Shamāl, ash-	34 25 N 36 00 E

Cities and towns

'Abdah, al-	34 31 N 35 58 E
Amyūn	34 18 N 35 49 E
B'abdā	33 50 N 35 32 E
Ba'labakk	34 00 N 36 12 E
B'aqlīn	33 41 N 35 33 E
Batrūn, al-	34 15 N 35 39 E
Beirut (Bayrūt)	33 53 N 35 30 E
Bḥamdūn	33 48 N 35 39 E
Bint Jubayl	33 07 N 35 26 E
Bsharrī	34 15 N 36 01 E
Byblos, see Jubayl	
Dāmūr, ad-	33 44 N 35 27 E
Ghazīr	34 01 N 35 40 E
Ghazzah	33 40 N 35 49 E
Ghūmā	34 13 N 35 42 E
Halbā	34 33 N 36 05 E
Ḥaṣbayya	33 24 N 35 41 E
Ḥimlāyā	33 56 N 35 42 E
Hirmīl, al-	34 23 N 36 23 E
Ihdin	34 17 N 35 58 E
Jazzīn	33 32 N 35 34 E
Jbā'	33 29 N 35 31 E
Jubayl (Byblos)	34 07 N 35 39 E
Jubb Jannīn	33 37 N 35 47 E
Jūniyah	33 59 N 35 38 E
Jwayyā	33 14 N 35 19 E
Khaldah	33 47 N 35 29 E
Labwah, al-	34 12 N 36 21 E
Marj 'Uyūn	33 22 N 35 35 E
Mashgharah	33 32 N 35 39 E
Mīnā', al-	34 27 N 35 49 E
Nabaṭīyah at-Taḥtā, an-	33 23 N 35 29 E
Nāqūrah, an-	33 07 N 35 08 E
Qarṭabā	34 06 N 35 51 E
Qubayyāt, al-	34 34 N 36 17 E
Rāshayyā	33 30 N 35 51 E
Ṣarafand, aṣ-	33 27 N 35 18 E
Ṣayda, see Sidon	
Shḥīm	33 37 N 35 29 E
Shikkā	34 20 N 35 44 E
Sidon (Ṣayda)	33 33 N 35 22 E
Sir aḍ-Ḍinnīyah	34 23 N 36 02 E
Ṣūr, see Tyre	
Tripoli (Ṭarābulus)	34 26 N 35 51 E
Tyre (Ṣūr)	33 16 N 35 11 E
Zaḥlah	33 51 N 35 53 E
Zghartā	34 24 N 35 54 E

Physical features and points of interest

Abū 'Alī, *river*	34 27 N 35 50 E
'Akkār, Sahl, *plain*	34 39 N 36 03 E
'Anjar, *archaeological site*	33 43 N 35 57 E
Anti-Lebanon (Jabal ash-Sharqī, al-), *mountains*	34 00 N 36 30 E
'Arabah, Jabal al-, *mountains*	33 35 N 35 46 E
'Āṣī, al-, see Orontes	
Awwalī, al-, *river*	33 35 N 35 23 E
Ba'labakk, *archaeological site*	34 01 N 36 11 E
Bārūk, Jabal al-, *mountain*	33 43 N 35 44 E
Bayrūt, *river*	33 54 N 35 32 E
Biqā', al-, *valley*	33 50 N 36 00 E
Byblos, *archaeological site*	34 07 N 35 38 E
Cedars of Lebanon, *cedar grove*	33 41 N 35 41 E
Dāmūr, ad-, *river*	33 42 N 35 26 E
Ḥāṣbānī, al-, *river*	33 11 N 35 37 E
Hermon, Mt. (Shaykh, Jabal ash-), *mountain*	33 26 N 35 51 E
Ibrāhīm, *river*	34 04 N 35 38 E
Ithnayn, Jabal al-, *mountain*	34 03 N 36 29 E
'Izzīyah, al-, *river*	33 12 N 35 12 E
Jabal ash-Sharqī, al-, see Anti-Lebanon	
Jawz, al-, *river*	34 16 N 35 39 E
Kabīr, al-, *river*	34 38 N 35 58 E
Kalb, al-, *river*	33 57 N 35 35 E
Kanīsah, Jabal al-, *ridge*	33 50 N 35 47 E
Lebanon Mountains (Lubnān, Jabal)	34 00 N 36 00 E
Līṭānī, al-, *river*	33 20 N 35 14 E
Mediterranean Sea	34 00 N 35 10 E
Mūsā, al-, *river*	34 26 N 36 00 E
Orontes ('Āṣī, al-), *river*	34 15 N 36 20 E
Qir'awn, Buḥayrat al-, *reservoir*	33 34 N 35 42 E
Salūqī, as-, *watercourse*	33 18 N 35 26 E
Ṣannīn, Jabal, *mountain*	33 57 N 35 52 E
Sawdā', Qurnat as-, *mountain*	34 18 N 36 07 E
Shamīs al-Ḥijār, *mountain*	34 04 N 36 01 E
Sharbīn, ash-, *watercourse*	34 27 N 36 26 E
Shaykh, Jabal ash-, see Hermon, Mt.	
Ṣūr, see Tyre	
Tyre (Ṣūr), *archaeological site*	33 16 N 35 11 E
Wadīn, *watercourse*	34 35 N 36 19 E
Zahrānī, az-, *river*	33 29 N 35 20 E



LEBANON

and from six to 16 miles wide, is part of the great East African Rift System. In the south, al-Biqā' becomes hilly and rugged, blending into the foothills of Mt. Hermon (Jabal ash-Shaykh) to form the upper Jordan Valley.

The Anti-Lebanon range (al-Jabal ash-Sharqī) starts with a high peak in the north and slopes southward until it is interrupted by Mt. Hermon (9,232 feet).

Drainage. Lebanese rivers, though numerous, are mostly winter torrents, draining the western slopes of the Lebanon Mountains. The only exception is the Līṭānī (90 miles long), which rises near the famed ruins of Baalbek (Ba'labakk) and flows southward in al-Biqā' to empty into the Mediterranean near historic Tyre. The two other important rivers are the Orontes (Nahr al-'Āṣī), which rises in the north of al-Biqā' and flows northward, and the Kabīr.

Climate. There are sharp local contrasts in climatic conditions. Lebanon is included in the Mediterranean climatic region, which extends westward to the Atlantic Ocean. The winter storms formed over the ocean move eastward through the Mediterranean, bringing rain at that season; in summer the Mediterranean receives no rain. The climate of Lebanon is generally subtropical and is characterized by hot, dry summers and mild, humid winters. Mean daily maximum temperatures on the coast and in al-Biqā' range from 90° F (32° C) in July to 60° F (16° C) on the coast and 50° F (10° C) in al-Biqā' in January. Mean minimum temperatures in January are 50° F (10° C) on the coast and 35° F (2° C) in al-Biqā'. At 5,000 feet, the altitude of the highest settlements, these are reduced by about 15° F (8° C).

Nearly all precipitation falls in winter and averages 30 to 40 inches (750 to 1,000 millimetres) on the coast, rising to more than 50 inches in higher altitudes. Al-Biqāʿ is drier and receives 15 to 25 inches. On the higher mountaintops, this precipitation falls as heavy snow that remains until early summer.

Cedar trees

Plant and animal life. Lebanon was heavily forested in ancient and medieval times, and its timber—particularly its famed cedar—was exported for building and shipbuilding. The natural vegetation, however, has been grazed, burned, and cut for so long that little of it is regenerated. What survives is a wild Mediterranean vegetation of brush and low trees, mostly oaks, pines, cypresses, firs, junipers, and carobs.

Few large wild animals survive in Lebanon, though bears are occasionally seen in the mountains. Among the smaller animals, deer, wildcats, hedgehogs, squirrels, martens, dormice, and hares are found. Numerous migratory birds from Africa and Europe visit Lebanon. Flamingos, pelicans, cormorants, ducks, herons, and snipes frequent the marshes; eagles, buzzards, kites, falcons, and hawks inhabit the mountains; and owls, kingfishers, cuckoos, and woodpeckers are common.

Settlement patterns. Most of the population live on the coastal plain, and progressively fewer people are found farther inland. Rural villages are sited according to water supply and the availability of land, frequently including terraced agriculture in the mountains. Northern villages are relatively prosperous and have some modern architecture. Villages in the south have been generally poorer and less stable; their agricultural land is less fertile and, because of their proximity to Israel, many have been subject to frequent dislocation, invasion, and destruction since 1975. Most cities are located on the coast; they have been inundated by migrants and displaced persons, and numerous, often poor, suburbs have been created as a result. Before 1975 many villages and cities were composed of several different religious groups, usually living together in harmony, and rural architecture reflected a unity of style irrespective of religious identity. Since the civil war began, a realignment has moved thousands of Christians north of Beirut along the coast and thousands of Muslims south or east of Beirut, so that settlement patterns reflect the chasms separating sections of the Lebanese people from each other.

THE PEOPLE

Lebanon has a heterogeneous society composed of numerous ethnic, religious, and kinship groups. Primordial attachments and local communalism antedate the creation of the present territorial and political entity and continue to survive with remarkable tenacity.

Ethnic and linguistic groups. Ethnically, the Lebanese compose a mixture in which Phoenician, Greek, Armenian, and Arab elements are discernible. Arabic is the official language, but French and English are widely spoken. A small percentage of the population is Armenian-speaking, and Syriac is used in some of the churches of the Maronites (Roman Catholics following an Eastern rite).

Religious groups. Perhaps the most distinctive feature of Lebanon's social structure is its varied religious composition. Since the 7th century Lebanon has served as a refuge for persecuted Christian and Muslim sects. The population is estimated to consist of a majority of Muslims and a large minority of Christians. Shīʿite Muslims are the most numerous group. Among the Christians, Maronites form the largest group, and Greek Orthodox and Greek Catholics are the next largest groups. Among the three Muslim denominations, the Shīʿites are followed closely by the Sunnites; the Druzes constitute a small percentage. There is also a very small minority of Jews.

Demography. One of the most salient demographic features of Lebanon is the uneven distribution of its population. The country's overall density is much lower than that of Bayrūt *muḥāfaẓah* (Beirut governorate) but much higher than that of the most sparsely populated, al-Biqāʿ governorate.

Before the civil war began, the movement of people from rural areas was a major factor in the country's soaring rate of urbanization. Most of the internal migration was to Beirut, which accounted for the great majority of Lebanon's urban population. The civil war and postwar fighting led to a substantial return of people to their villages and to a large migration abroad, primarily to the United States, Europe, Latin America, Australia, and the oil states of the Middle East.

THE ECONOMY

Until 1975 Lebanon had an economy characterized by a minimum of government intervention in private enterprise. Since the civil war, the weak central government has exercised little power in economic matters, and local militias have dominated public decision making.

Werner Braun—Camera Press

Cedars of Lebanon on the slopes of the Lebanon Mountains.

The services sector generated the overwhelming proportion of national income before the civil war and employed the largest proportion of the labour force; industry generated the second largest proportion of income and of employment. Agriculture accounted for a smaller proportion of income. The growth of services was related mainly to international transport and trade and to the position of Beirut as a centre of international banking and tourism.

Economic losses after 1974

The war of 1975–76, the Israeli invasion of 1982, and the continuing violence have left deep scars and have led to chaos in the economy. There has been extensive destruction in all sectors, but especially in housing, trade, and public services, and the country's productive capacity has been drastically reduced. The greatest reduction in productive capacity seems to be in services, followed by industry and agriculture.

Resources. The mineral resources of Lebanon are few. There are deposits of high-grade iron ore and lignite; building-stone quarries; high-quality sand, suitable for glass manufacture; and lime. The Līṭānī River hydroelectric project generates electricity and also has increased the amount of irrigated land for agriculture.

Agriculture. Arable land is scarce, but the climate and the relatively abundant water supply from springs favour the intensive cultivation of a variety of crops on mountain slopes and in the coastal region. On the irrigated coastal plain market vegetables, bananas, and citrus crops are grown. In the foothills the principal crops are olives, grapes, tobacco, figs, and almonds. At higher altitudes (about 1,500 feet) peaches, apricots, plums, and cherries are planted, while apples and pears thrive at an altitude of about 3,000 feet. Sugar beets, cereals, and vegetables are the main crops cultivated in al-Biqāʿ. Poultry is a major source of agricultural income, and goats, sheep, and cattle are also raised.

As a result of the continued violence many small farmers have lost their livestock, and there has been a noticeable decrease in the production of many agricultural crops. The production of hemp, the source of hashish, has flourished in al-Biqāʿ valley, however, and the hashish is exported illegally through ports along the coast.

Industry. The majority of the country's industry survived the civil war unscathed. Beirut's industrial belt was razed, but some of the country's large complexes were unharmed. Manufacturing recovered to more than half of the still existing capacity, restrained by limited labour mobility, difficulty in acquiring supplies, insufficient working capital, and difficulty in obtaining credit. The Israeli invasion of 1982, however, with its heavy bombardment of some of Lebanon's major cities and subsequent sabotage by local warring factions, caused further damage to industry and infrastructure.

Beirut's well-developed seaport and airport and the country's free economic and foreign exchange systems, favourable interest rates, and banking secrecy law (modeled upon that of Switzerland) all contributed to the traditional preeminence of trade and services. Prior to the civil war, the country's scenery, its biblical and other historic sites, its hotels, bars, nightclubs, and restaurants, its seaside and mountain resorts, its outdoor sports facilities, and its international cultural festivals made tourism a year-round industry. As the war progressed, the prosperous hotel district in Beirut became the scene of some of the fiercest fighting, and bombing in 1982 caused heavy damage. The closures of Beirut airport, the heavy destruction of the port, and the continued political unrest greatly damaged the service industry.

Finance. The finance sector of Lebanon's economy, including banking and insurance, showed an impressive expansion before the war, and the monetary reserves of Lebanon continued to rise despite political uncertainties. During the two years of civil war and the extended period of domestic instability and economic inflation following that, reserves, which included a considerable portion of gold, nevertheless continued to rise.

Capital inflows

The balance of payments has traditionally shown a surplus. The strength of the Lebanese pound and of the balance-of-payments position reflected large inflows of capital, mostly from Lebanese living abroad (whose numbers rose considerably during and after the civil war) and from the high level of liquidity of commercial banks. By 1983, however, inflows from Lebanese living abroad had begun to decrease, and the value of the Lebanese pound fell dramatically.

Trade. Widespread smuggling, covert foreign aid to armed groups, and illegal drug production have disguised the pattern of trade since 1975. Exports, chiefly vegetable products, textiles, and nonprecious metals are sent mainly to Middle Eastern countries. Imports such as consumer goods, machinery and transport equipment, petroleum products, and food come mostly from western Europe. A huge trade deficit has been partly covered by "invisible" items such as foreign remittances and government loans.

Trade unions. Lebanon is one of the few countries in the Middle East with a comparatively well-developed labour movement. Trade unions have secured some tangible gains, such as fringe benefits, collective bargaining contracts, and better working conditions. During the civil war divisions in many of the trade unions weakened their normal functions, and many of their members joined the warring factions. Many others emigrated.

Transportation. As in antiquity, Lebanon's situation makes it a vital crossroads between East and West. The road network traversing Lebanon includes international highways, which form part of major land routes connecting Europe with the Arab countries and the East. There are also national highways, paved secondary roads, and unpaved roads. The railway system, which includes lines along the coast and up al-Biqāʿ valley and a cog railway across the Lebanon Mountains, connects with the rail system of Syria. In the past the system was linked with rail systems of other Arab countries and with Europe and was used mainly for long-distance bulk transport.

Port of Beirut

There are numerous ports along the Lebanese seacoast. Berths for oil tankers have been built offshore at Tripoli and at az-Zahrānī, near Sidon, where pipeline terminals and refineries also are located. The principal cargo and passenger port is that of Beirut, which has a free zone and storage facilities for transit shipments. The port has been expanded and deepened, and a large storage silo (for wheat and other grains) has been built there, but port facilities were severely damaged during the civil war and the postwar fighting. The harbour at Jūniyah has grown in importance.

Beirut International Airport was one of the busiest airports in the Middle East before the civil war. Its runways were built to handle the largest jet airplanes in service, and a number of international airlines used Beirut regularly.

GOVERNMENT AND SOCIAL CONDITIONS

Parliament

Government. *The constitutional framework.* Modern Lebanon is a republic with a parliamentary system of government. Its constitution, promulgated in 1926 during the French mandate and modified by several subsequent amendments, provides for a unicameral Chamber of Deputies (renamed National Assembly in 1979) elected, for a term of four years, by universal adult suffrage (women attained the right to vote and eligibility to run for office in 1953). Parliamentary seats are apportioned on a religious basis in the ratio of six Christians to five Muslims, making the total number always a multiple of 11. This sectarian distribution is also to be observed in administrative appointments for public office.

The president of the republic is elected by a two-thirds majority of the National Assembly for a term of six years and is eligible for reelection only after the lapse of a further six years. By an unwritten convention, the president must be a Maronite, the premier a Sunnite Muslim, and the speaker of the National Assembly a Shīʿite. The president, who is constitutionally invested with the power of chief executive, calls upon a Sunnite Muslim to form a Cabinet, and the Cabinet members' portfolios must reflect the sectarian balance. The Cabinet, to remain in power, requires a vote of confidence from the Assembly. A vote of no confidence, however, is a constitutional right that is rarely exercised in practice. A Cabinet usually falls because of internal dissension or because the president withdraws his support.

Local government. For administrative purposes Lebanon is divided into *muḥāfaẓāt* (governorates): Bayrūt (Beirut), Jabal Lubnān, ash-Shamāl, al-Janūb, al-Biqāʿ, and an-Nabaṭīah. These are administered by the *muḥāfiẓ* (governor), who represents the central government. The *muḥāfaẓāt* are further divided into *qaḍāʾ*s (districts), each of which is presided over by a *qāʾim-maqām* (district chief), who, along with the *muḥāfiẓ*, supervises local government. Municipalities (communities with at least 500 inhabitants) elect their own councils, which in turn elect mayors and vice mayors. Villages and towns (more than 50 and fewer than 500 inhabitants) elect a *mukhṭār* (headman) and a council of elders, who serve on an honorary basis. All officers of local governments serve four-year terms.

The political process. The political system in Lebanon remains a curious blend of secular and traditional features. Until 1975 the country appeared to support liberal and democratic institutions, yet, in effect, it had hardly any of the political instruments of a civil polity. Its political parties, parliamentary blocs, and pressure groups were so closely identified with parochial, communal, and personal loyalties that they often failed to serve the larger national purpose of the society. The National Pact of 1943, a sort of Christian–Muslim entente, sustained the national entity (*al-kiyān*), yet this sense of identity was neither national nor civic.

National Pact of 1943

In April 1975 the political process collapsed. The war that had engulfed the Lebanese exposed the vulnerability of the political system. The legitimate authority continues to maintain the facade of continuity, while the process on which it is based was destroyed by the contending forces in the conflict that has continued to ravage Lebanon. Real power rests in the sectarian militias, who are not under the control of the official central government.

Justice. The system of law and justice is mostly modeled on French concepts. The judiciary consists of courts of the first instance, courts of appeal, courts of cassation, and a Court of Justice that handles cases affecting state security. The Council of State is a court that deals with administrative affairs. In addition, there are religious courts that deal with matters of personal status (such as inheritance, marriage, and property matters) as they pertain to autonomous communities. Despite the country's well-developed legal system and a very high proportion of lawyers, significant numbers of disputes and personal grievances are resolved outside the courts. Justice by feud and vendetta continues.

Armed forces. The armed forces consist of an army, air force, and navy. Lebanon also has a paramilitary gendarmerie and police force. During the civil war the army practically disintegrated as splinter groups joined the different warring factions. Reconstruction of the Lebanese armed forces has been attempted, particularly with the assistance first of the United States and then of Syria, but to very little effect. Responsibility for maintaining security and order has fallen to the various political and religious factions and to foreign occupiers.

Education. A well-developed system of education reaches all levels of the population. Literacy is among the highest in the Middle East. Education was once almost exclusively the responsibility of religious communities or foreign groups, but the number of students in public schools has risen to about two-fifths of the total school enrollment.

The compulsory five-year primary school program is followed either by a seven-year secondary program (leading to the official baccalaureate certificate) or by a four-year program of technical or vocational training. Major universities include the American University of Beirut, the Universitê Saint-Joseph (subsidized by the French government and administered by the Jesuit order), the Lebanese University (Universitê Libanaise), and the Beirut Arab University (an affiliate of the University of Alexandria).

Universities

Health and welfare. Public health services are largely concentrated in the cities, although the government increasingly directs medical aid into rural areas. As in the field of social welfare, nongovernmental voluntary associations—mostly religious, communal, or ethnic—are active. The Lebanese diet is generally satisfactory, and the high standard of living and the favourable climate have served to reduce the incidence of many diseases that are still common in other Middle Eastern countries.

Lebanon has a large number of skilled medical personnel, and hospital facilities are also adequate under normal circumstances. The thousands of casualties and deaths caused by almost constant warfare since 1975 have overburdened the country's health and medical facilities and hampered the treatment of routine illnesses.

The National Social Security Fund, which is not fully implemented, provides sickness and maternity insurance, labour accident and occupational disease insurance, family benefits, and termination-of-service benefits.

Housing. In response to the need for low-cost housing, the Popular Housing Law was enacted, providing for the rehabilitation of substandard housing. Prior to the civil war a substantial percentage of homes were without bathrooms, and thousands of families, including Palestinian refugees, were living in improvised accommodations. When an economic boom attracted villagers to the capital, the housing shortage worsened considerably. The civil war drastically increased the problem. Thousands of homes in battle zones were destroyed, and entire villages were evacuated and others occupied. The result was chaos in which property rights were violated as a matter of course. The government, in an attempt to remedy the situation, set up a Housing Bank to make housing loans.

Housing Bank

Wages and cost of living. A minimum wage is set by the Labour Code, and legislation provides for cost-of-living increases. The cost of living increased sharply prior to, during, and after the civil war, mainly because of a substantial rise in the cost of rent, education, food, and petroleum products.

Social and economic division. Lebanese society was able for a long time to give a semblance of relative economic stability. The existence of a large middle-income group, in addition to the political and social legitimacy that kinship ties and religious and communal attachments had, reinforced the false veneer that masked the growing socioeconomic dislocations. The interaction of these factors covered up the growing class polarization, especially around the industrial belt that encircled Beirut. The eruption of civil conflict in 1975 and the ensuing state of chaos is attributable in part to the fact that the system of government was unresponsive to the acute social problems and grievances.

CULTURAL LIFE

The cultural milieu. Historically, Lebanon is the heir of a long succession of Mediterranean cultures—Phoenician, Greek, and Arab. Its cultural milieu continues to show clear manifestations of a rich and diverse heritage. As an Arab country, Lebanon shares more than a common language with neighbouring Arab states; it also has a similar cultural heritage and common interests.

In the 19th century Lebanese linguists were in the vanguard of the Arabic literary awakening. In more recent times, writers of the calibre of Khalil Gibran, Georges Shehade, and Michel Chiha have been widely translated and have reached an international audience.

While for a time cultural life in Lebanon was predominantly centred around universities and affiliated institutions, there has been an impressive proliferation of cultural activities under other auspices. Beirut has several museums and a number of private libraries, learned societies, and research institutions.

The state of the arts. Lebanon's antiquities and ruins have provided not only inspiration for artists but also magnificent backdrops for annual music festivals, most notably the Baalbek International Festival. At one time, international opera, ballet, symphony, and drama companies, of nearly all nationalities, competed to enrich the cultural life of Beirut. Lebanon has produced a number of gifted young artists who have shown a refreshing readiness to experiment with new expressive forms. Some Lebanese are active in European opera and theatre companies, while others are intent on creating a wider audience for classical Arabic music and theatre.

Music festivals

The cultural awakening encouraged the revival of national folk arts, particularly song, *dabkah* (the national dance), and *zajal* (folk poetry), and the refinement of traditional crafts. Although the Baalbek International Festival was suspended during the civil war, popular theatre and radio satires continued to flourish in the war-ridden country.

The communications media. In addition to the wide variety of foreign newspapers and magazines that can be found in Beirut, Lebanon has registered publications in Arabic, English, French, and Armenian, including a number of daily newspapers.

Because of the wide choice of films, comfortable surroundings, and low prices, going to view motion pictures emerged as a popular form of entertainment among the Lebanese. Moviegoing has been supplanted to some extent, however, by private viewing of videocassettes.

Television programming is offered by Beirut's private companies, and Egyptian and Syrian broadcasts are received. With the advent of the transistor, Lebanon became virtually saturated with radios. The government-run radio station broadcasts Arabic, French, English, and Armenian programs, and clandestine radio stations broadcast the news and views of the warring parties.

For statistical data on the land and people of Lebanon, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (S.G.K./C.F.M./W.L.O.)

History

PHOENICIA

Origins and relations with Egypt. The evidence of tools found in caves along the coast of Lebanon shows that the area was inhabited from the Paleolithic through the Neolithic periods. Village life followed the domestication of plants and animals (the Neolithic Revolution, after about 10,000 BC), with Byblos (modern Jubayl) apparently taking the lead. At this site also appear the first traces in Lebanon of pottery and metallurgy (first copper, then bronze, an alloy of tin and copper) by the 4th millennium BC. The Phoenicians, indistinguishable from the Canaanites of Palestine, probably arrived in the land that became Phoenicia (a Greek term applied to the coast of Lebanon) in about 3000 BC. Herodotus and other Classical writers preserve a tradition that they came from the coast of the Erythraean Sea (*i.e.*, the Persian Gulf), but in fact nothing certain is known of their original homeland.

Arrival of the Phoenicians

Except at Byblos, no excavations have produced any information concerning the 3rd millennium in Phoenicia before the advent of the Phoenicians. At Byblos, the first urban settlement is dated *c.* 3050–2850 BC. Commercial and religious connections, probably by sea, with Egypt are attested from the Egyptian 4th dynasty (*c.* 2613–*c.* 2494 BC). The earliest artistic representations of Phoenicians are found in a damaged relief at Memphis of Pharaoh Sahure of the 5th dynasty (early 25th century BC). This shows the arrival of an Asiatic princess to be the Pharaoh's bride; her escort is a fleet of seagoing ships, probably of the type known to the Egyptians as "Byblos ships," manned by crews of Asiatics, evidently Phoenicians.

Byblos was destroyed by fire about 2150 BC, probably by the invading Amorites. The Amorites rebuilt on the site, and a period of close contact with Egypt was begun. Costly gifts were given by the pharaohs to those Phoenician and Syrian princes, such as the rulers of Ugarit and Katna, who were loyal to Egypt. Whether this attests to Egypt's political dominion over Phoenicia at this time or simply to strong diplomatic and commercial relations is not entirely clear.

In the 18th century BC new invaders, called Hyksos, destroyed the Amorite rule in Byblos and, passing on to Egypt, brought the Middle Kingdom to an end (*c.* 1720 BC). Little is known about the Hyksos' origin, but they seem to have been ethnically mixed, including a considerable Semitic element, since the Phoenician deities El, Baal, and Anath figured in their pantheon. The rule of the Hyksos in Egypt was brief and their cultural achievement slight, but in this period the links with Phoenicia and Syria were strengthened by the presence of Hyksos aristocracies throughout the region. Pharaoh Ahmose I expelled the Hyksos in about 1567 BC and instituted the New Kingdom policy of conquest in Palestine and Syria. In his annals Ahmose records capturing oxen from the Fenkhw, a term here perhaps referring to the Phoenicians. In the annals of the greatest Egyptian conqueror, Thutmose III (*c.* 1504–1450 BC), the coastal plain of Lebanon, called Djahy, is described as rich with fruit, wine, and grain. Of particular importance to the New Kingdom pharaohs was the timber, notably the cedar, of the Lebanese forests. A temple relief at Karnak depicts the chiefs of Lebanon felling cedars for the Egyptian officers of Seti I (*c.* 1300 BC).

Fuller information about the state of Phoenicia in the 14th century BC comes from the Amarna Letters, diplomatic texts belonging to the Egyptian foreign office, written in cuneiform and found at Tell el-Amarna in Middle Egypt. These archives reveal that the Land of Retenu (Syria–Palestine) was divided into three administrative districts, each under an Egyptian governor. The northernmost district (Amurru) included the coastal region from Ugarit to Byblos; the central (Upi) included the southern al-Biqāʿ valley and Anti-Lebanon; and the third district (Canaan) included all of Palestine from the Egyptian border to Byblos. Also among the letters are many documents addressed by the subject princes of Phoenicia and their Egyptian governors to the pharaoh. It was a time of much political unrest. The Hittites from central Anatolia were invading Syria; nomads from the desert supported the invasion, and many of the local chiefs were ready to seize the opportunity to throw off the yoke of Egypt. The tablets that reveal this state of affairs are written in the language and script of Babylonia (*i.e.*, Akkadian) and thus show the extent to which Babylonian culture had penetrated Palestine and Phoenicia; at the same time they illustrate the closeness of the relations between the Canaanite towns (*i.e.*, those in Palestine) and the dominant power of Egypt.

The Hittite invasion

After the reign of Akhenaton (Amenhotep IV; 1379–*c.* 1362), that power collapsed altogether; but his successors attempted to recover it, and Ramses II reconquered Phoenicia as far as the Nahr el Kelb. In the reign of Ramses III (1198–66) many great changes began to occur as a result of the invasion of Syria by peoples from Asia Minor and Europe. The successors of Ramses III lost their hold over Canaan; the 21st dynasty no longer intervened in the affairs of Syria. In *The Story of Wen-Amon,* a tale of an Egyptian religious functionary sent to Byblos to secure cedar around 1100 BC, the episode of the functionary's inhospitable reception shows the extent of the decline of Egypt's authority in Phoenicia at this time. Sheshonk (Shishak) I, the founder of the 22nd dynasty, in about 928 BC endeavoured to assert the ancient supremacy of Egypt. His successes, however, were not lasting, and, as is clear from the Old Testament, the power of Egypt thereafter became ineffective.

Phoenicia as a colonial and commercial power. Kingship appears to have been the oldest form of Phoenician government. The royal houses claimed divine descent, and the king could not be chosen outside their members. His power, however, was limited by the powerful merchant families, who wielded great influence in public affairs. Associated with the king was a council of elders; such at least was the case at Byblos, Sidon, and perhaps Tyre. During Nebuchadrezzar II's reign (605–562 BC) a republic took the place of the monarchy at Tyre, and the government was administered by a succession of suffetes (judges); they held office for short terms, and in one instance two ruled together for six years. Much later, in the 3rd century BC, an inscription from Tyre also mentions a suffete. Carthage was governed by two suffetes, and these officers are frequently named in connection with the Carthaginian colonies. But this does not justify any inference that Phoenicia itself had such magistrates. Under the Persians a federal bond was formed linking Sidon, Tyre, and Aradus. Federation on a larger scale was never possible in Phoenicia, for the reason that no sense of political unity existed to bind the different states together.

Colonies. By the 2nd millennium BC the Phoenicians had already extended their influence along the coast of the Levant by a series of settlements, some well known, some virtually nothing but names. Well known throughout history are Joppa (Tel Aviv–Yafo) and Dor (later Tantura, modern Nasholim) in the south. The earliest site, however, outside the Phoenician homeland known to possess important aspects of Phoenician culture is Ugarit (Ras Shamra), about six miles north of Latakia. The site was already occupied before the 4th millennium BC, but the Phoenicians only became prominent there in the Egyptian 12th dynasty (1991–1786 BC).

Evidence remains of two temples dedicated to the Phoenician gods Baal and Dagon, although the ruling family appears to have been of different, non-Phoenician stock. The 15th century BC shows strong cultural influences already established there from Cyprus and the world of Mycenaean Greece. A splendid archive of literary and administrative documents found at Ugarit from this period provides evidence of an early form of alphabetic script, arguably the most important Phoenician contribution to Western civilization. In the latter part of the 13th century BC, a flood of land and sea raiders (the Sea Peoples) descended on the Levant coast, destroying many of the Phoenician cities and rolling onward to the frontier of Egypt, from which they were beaten back by the Pharaoh Ramses III. Ugarit was destroyed, together with Aradus and Byblos, though the latter were afterward rebuilt. Though Sidon was destroyed only in part, its inhabitants fled to Tyre, which from this time was regarded as the principal city of Phoenicia and began its period of prosperity and expansion.

Phoenician expansion

Tyre's first colony, Utica in North Africa, was founded perhaps as early as the 10th century BC. It is likely that the expansion of the Phoenicians at the beginning of the 1st millennium BC is to be connected with the alliance of Hiram of Tyre with Solomon of Israel in the second half of the 10th century BC. In the following century, Phoenician presence in the north is shown by inscriptions at Samal (Zincirli Hüyük) in eastern Cilicia, and in the 8th century at Karatepe in the Taurus Mountains, but there is no evidence of direct colonization. Both these cities acted as fortresses commanding the routes through the mountains to the mineral and other wealth of Anatolia.

Cyprus had Phoenician settlements by the 9th century BC. Citium, known to the Greeks as Kition (biblical Kittim), in the southeast corner of the island, became the principal colony of the Phoenicians in Cyprus. Elsewhere in the Mediterranean, several smaller settlements were planted as stepping-stones along the route to Spain and its mineral wealth in silver and copper: at Malta, early remains go back to the 7th century BC, and at Sulcis and Nora in Sardinia and Motya in Sicily, perhaps a century earlier. According to Thucydides, the Phoenicians controlled a large part of the island but withdrew to the northwest corner under pressure from the Greeks. Modern scholars, however, disbelieve this and contend that the Phoenicians arrived only after the Greeks were established.

In North Africa the next site colonized after Utica was Carthage (near Tunis). Carthage in turn seems to have established (or, in some cases, reestablished) a number of settlements in Tunisia, Algeria, Morocco, the Balearic Islands, and southern Spain, eventually making this city the acknowledged leader of the western Phoenicians.

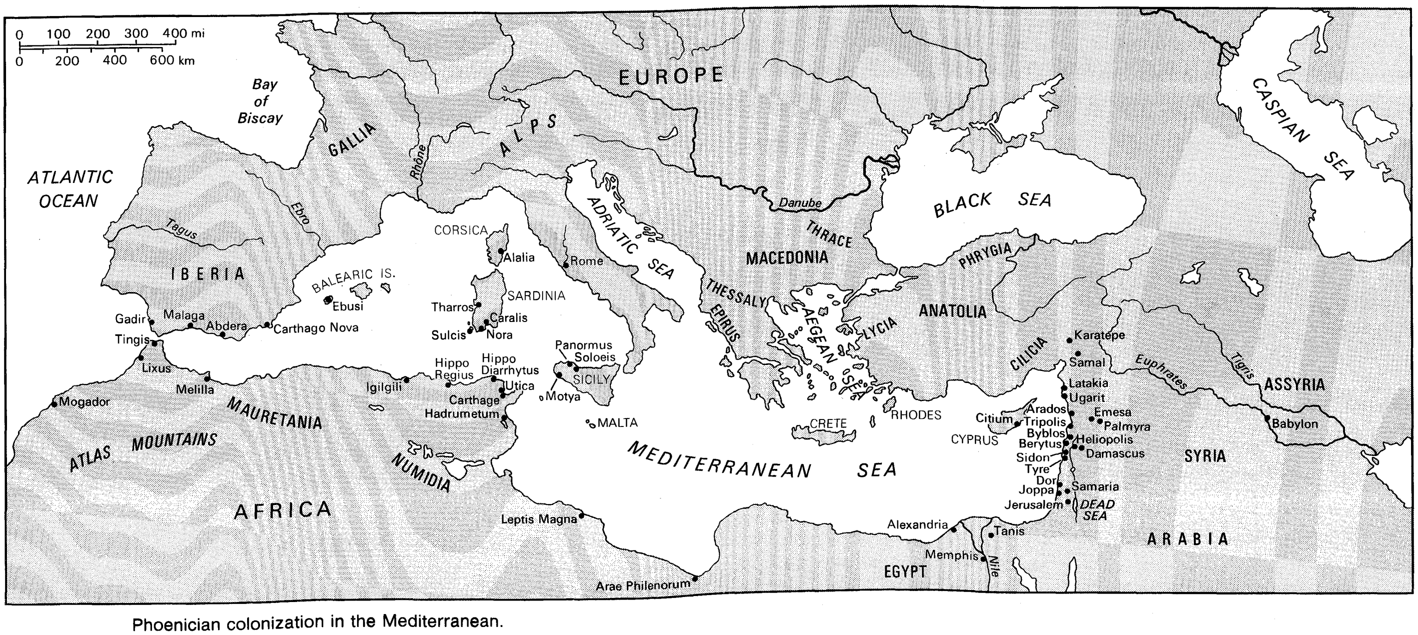
There is little factual evidence to confirm the presence of any settlement in Spain earlier than the 7th century BC, or perhaps the 8th century, and many of these settlements should be viewed as Punic (Carthaginian) rather than Phoenician, though it is likely that the colonizing expeditions of the Carthaginians were supported by many emigrants from the Phoenician homeland. It is very probable that the tremendous colonial activity of the Phoenicians and Carthaginians was stimulated in the 8th to 6th centuries BC by the military blows that were wrecking the trade of the Phoenician homeland in the Levant. Also, competition with the synchronous Greek colonization of the western Mediterranean cannot be ignored as a contributing factor.

In the 3rd century BC Carthage, defeated by the Romans, embarked on a further imperialistic phase in Spain to recoup its losses. Rome responded, defeated Carthage a second time, and annexed Spain. Finally, in 146 BC, after a third war with Rome, Carthage suffered total destruction. It was rebuilt as a Roman colony in 44 BC. The ancient Phoenician language survived in use as a vernacular in some of the smaller cities of North Africa at least until the time of St. Augustine, bishop of Hippo (5th century AD).

Commerce under Egyptian protection

Commerce. The role that tradition especially assigns to the Phoenicians as the merchants of the Levant was first developed on a considerable scale at the time of the Egyptian 18th dynasty. The position of Phoenicia, at a junction of both land and sea routes, under the protection of Egypt, favoured this development, and the discovery of the alphabet and its use and adaptation for commercial purposes assisted the rise of a mercantile society. A fresco in an Egyptian tomb of the 18th dynasty depicted seven Phoenician merchant ships that had just put in at an Egyptian port to sell their goods, including the distinctive Canaanite wine jars in which wine, a drink foreign to the Egyptians, was imported. *The Story of Wen-Amon* recounts the tale of a Phoenician merchant, Werket-el of Tanis in the Nile Delta, who was the owner of "50 ships" that sailed between Tanis and Sidon. The Sidonians are also famous in the poems of Homer as craftsmen, traders, pirates, and slave dealers. The prophet Ezekiel (chapters 27 and 28), in a famous denunciation of the city of Tyre, catalogs the vast extent of its commerce, covering most of the then-known world.

The exports of Phoenicia as a whole included particularly cedar and pine wood from Lebanon, fine linen from Tyre,



Phoenician colonization in the Mediterranean.

Byblos, and Berytos, cloths dyed with the famous Tyrian purple (made from the snail *Murex*), embroideries from Sidon, metalwork and glass, glazed faience, wine, salt, and dried fish. They received in return raw materials, such as papyrus, ivory, ebony, silk, amber, ostrich eggs, spices, incense, horses, gold, silver, copper, iron, tin, jewels, and precious stones.

In addition to these exports and imports, the Phoenicians also conducted an important transit trade, especially in the manufactured goods of Egypt and Babylonia (Herodotus, i, 1). From the lands of the Euphrates and Tigris regular trade routes led to the Mediterranean. In Egypt the Phoenician merchants soon gained a foothold; they alone were able to maintain a profitable trade in the anarchic times of the 22nd and 23rd dynasties (*c.* 945–*c.* 730 BC). Though there were never any regular colonies of Phoenicians in Egypt, the Tyrians had a quarter of their own in Memphis (Herodotus, ii, 112). The Arabian caravan trade in perfume, spices, and incense passed through Phoenician hands on its way to Greece and the West (Herodotus, iii, 107).

Caravan trade

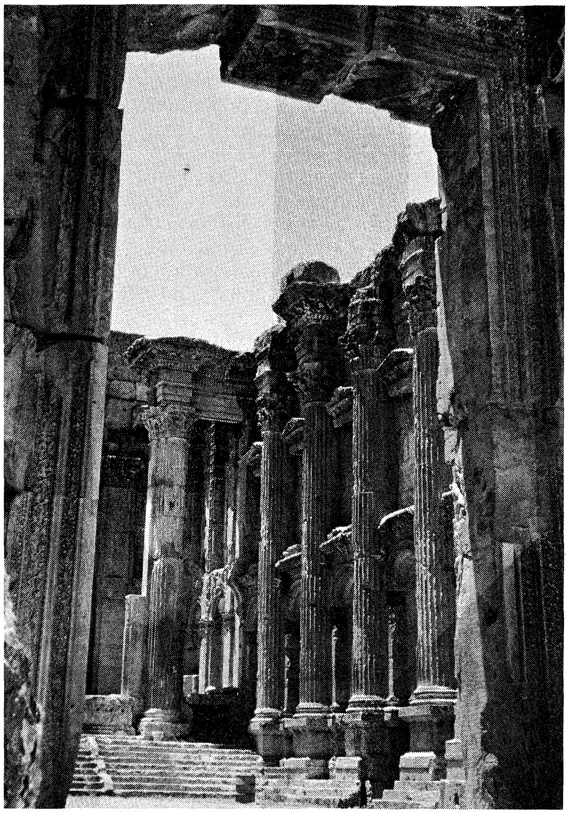
The Phoenicians were not mere passive peddlers in art or commerce. Their achievement in history was a positive contribution, even if it was only that of an intermediary. For example, the extent of the debt of Greece alone to Phoenicia may be fully measured by its adoption, probably in the 8th century BC, of the Phoenician alphabet with very little variation (along with Semitic loan words); by "orientalizing" decorative motifs on pottery and by architectural paradigms; and by the universal use in Greece of the Phoenician standards of weights and measures.

Navigation and seafaring. For the establishment of commercial supremacy, an essential constituent was the Phoenician skill in navigation and seafaring. The Phoenicians are credited with the discovery and use of Polaris (the Pole Star). Fearless and patient navigators, they ventured into regions where no one else dared to go, and always, with an eye to their monopoly, they carefully guarded the secretş of their trade routes and discoveries and their knowledge of winds and currents. Pharaoh Necho II (610–595 BC) organized the Phoenician circumnavigation of Africa (Herodotus, iv, 42). Hanno, a Carthaginian, led another in the mid-5th century. The Carthaginians seem to have reached the island of Corvo in the Azores; and they may even have reached Britain, for many Carthaginian coins have been found there.

Assyrian and Babylonian domination of Phoenicia. Between the withdrawal of Egyptian rule in Syria and the western advance of Assyria there was an interval during which the city-states of Phoenicia owned no suzerain. Byblos had kings of its own, among them Ahiram, Abi-baal, and Ethbaal (Ittobaʿal) in the 10th century, as excavations have shown. The history of this time period is mainly a history of Tyre, which not only rose to a hegemony among the Phoenician states but also founded colonies beyond the seas. Unfortunately, the native historical records of the Phoenicians have not survived, but it is clear from the Bible that the Phoenicians lived on friendly terms with the Israelites. In the 10th century Hiram, king of Tyre, built the Temple of Solomon at Jerusalem in return for rich gifts of oil, wine, and territory. In the following century Ethbaal of Tyre married his daughter Jezebel to Ahab, king of Israel, and Jezebel's daughter in turn married the King of Judah.

The Assyrian intrusion

In the 9th century, however, the independence of Phoenicia was increasingly threatened by the advance of Assyria. In 868 BC Ashurnasirpal II reached the Mediterranean and exacted tribute from the Phoenician cities. His son, Shalmaneser III, took tribute from the Tyrians and Sidonians and established a supremacy over Phoenicia, at any rate in theory, which was acknowledged by occasional payments of tribute to him and his successors. In 734 BC Tiglath-pileser III in his western campaign established his authority over Byblos, Arados, and Tyre. A fresh invasion, by Shalmaneser V, took place in 725 when he was on his way to Samaria, and in 701 Sennacherib, facing a rebellion of Philistia, Judah, and Phoenicia, drove out and deposed Luli, identified as king of both Sidon and Tyre. In 678 Sidon rebelled against the Assyrians, who marched down and annihilated the city, rebuilding it on the mainland. Sieges of Tyre took place in 672 and 668, but it resisted both, only submitting in the later years of Ashurbanipal.



Ruins of the Temple of Bacchus at Baalbek.

Peter Fenwick—The J. Allan Cash Photolibrary

During the period of Neo-Babylonian power, which followed the fall of Nineveh in 612 BC, the pharaohs made attempts to seize the Phoenician and Palestinian seaboard. Nebuchadrezzar II, king of Babylon, having sacked Jerusalem, marched against Phoenicia and besieged Tyre, but it held out successfully for 13 years, after which it capitulated, seemingly on favourable terms.

Persian period. Phoenicia passed from the suzerainty of the Babylonians to that of their conquerors, the Persians, in 538 BC. Not surprisingly, the Phoenicians turned as loyal supporters to the Persians, who had overthrown their oppressors, and reopened to them the trade of the Orient. Lebanon, Syria–Palestine, and Cyprus were organized as the fifth satrapy (province) of the Persian Empire. At the time of Xerxes I's invasion of Greece, the city of Sidon was considered to be the principal city of Phoenicia, and the ships of Sidon were considered the finest part of Xerxes' fleet, its king ranking next to Xerxes and before the king of Tyre. (Phoenician coins have been used to supplement historical sources on the period. From the reign of Darius I [522–486 BC], the Persian monarchs had allowed their satraps and vassal states to coin silver and copper money. Arados, Byblos, Sidon, and Tyre therefore issued a coinage of their own.) In the 4th century Tyre, and later Sidon, revolted against the Persian king. The revolt was suppressed in 345 BC.

Greek and Roman periods. In 332 BC Tyre resisted Alexander the Great in a siege of eight months. Alexander finally captured the city by driving a mole into the sea from the mainland to the island. As a result Tyre, the inhabitants of which were largely sold into slavery, lost all importance, soon being replaced in the leadership of the Oriental markets by Alexandria, the conqueror's newly founded city in Egypt. In the Hellenistic period (323–30 BC) the cities of Phoenicia became the prize for the competing Macedonian dynasties, controlled first by the Ptolemies of Egypt in the 3rd century BC and then by the Seleucids of Syria in the 2nd and early decades of the 1st century BC. The Seleucids apparently permitted a good measure of autonomy to the Phoenician cities. Tigranes II of Armenia brought an end to the Seleucid dynasty in 83 BC and extended his realm to Mt. Lebanon. The Romans eventually intervened to restore Seleucid sovereignty, but when anarchy prevailed they imposed peace and assumed direct rule in 64 BC.

Phoenicia was incorporated into the Roman province of Syria, though Aradus, Sidon, and Tyre retained self-government. Berytus (Beirut), relatively obscure to this point, rose to prominence by virtue of Augustus' grant of Roman colonial status and by the lavish building program financed by Herod the Great (and in turn by his grandson and great-grandson). Under the Severan dynasty (AD 193–235) Sidon, Tyre, and probably Heliopolis (Baalbek) also received colonial status. Under this dynasty the province of Syria was partitioned into two parts: Syria Coele ("Hollow Syria"), comprising a large region loosely defined as north and east Syria; and Syria Phoenice in the southwestern region, which included not only coastal Phoenicia but also the territory beyond the mountains and into the Syrian desert. Under the provincial reorganization of the Eastern Roman emperor Theodosius II in the early 5th century AD, Syria Phoenice was expanded into two provinces: Phoenice Prima (Maritima), basically ancient Phoenicia; and Phoenice Secunda (Libanesia), an area extending to Mt. Lebanon on the west and deep into the Syrian desert on the east. Phoenice Secunda included the cities of Emesa (its capital), Heliopolis, Damascus, and Palmyra.

During the period of the Roman Empire the native Phoenician language died out in Lebanon and was replaced by Aramaic as the vernacular. Latin, the language of the soldiers and administrators, in turn fell before Greek, the language of letters of the eastern Mediterranean, by the 5th century AD. Lebanon produced a number of important writers in Greek, most notably Philo of Byblos (64–141), and in the 3rd century Porphyry of Tyre and Iamblichus of Chalcis in Syria Coele. Porphyry played a key role in disseminating the Neoplatonic philosophy of his master Plotinus, which would influence both pagan and Christian thought in the later Roman Empire.

In many respects, the two most important cities of Lebanon during the time of the Roman Empire were Heliopolis and Berytus. At Heliopolis the Roman emperors, particularly the Severans, constructed a monumental temple complex, the most spectacular elements of which were the Temple of Jupiter Heliopolitanus and the Temple of Bacchus. Berytus, on the other hand, became the seat of the most famous provincial school of Roman law. The school, which probably was founded by Septimius Severus, lasted until the destruction of Berytus itself by a sequence of earthquakes, tidal wave, and fire in the mid-6th century. Two of Rome's most famous jurists, Papinian and Ulpian, both natives of Lebanon, taught as professors at the law school under the Severans. Their judicial opinions constitute well over a third of the Pandects (Digest) contained in the great compilation of Roman law commissioned by the emperor Justinian I in the 6th century AD.

In 608–609 the Persian king Khosrow II pillaged Syria and Lebanon and reorganized the area into a new satrapy, excluding only Phoenicia Maritima. Between 622 and 629 the Byzantine emperor Heraclius mounted an offensive and restored Syria–Lebanon to his empire. This success was short-lived; in the 630s Muslim Arabs conquered Palestine and Lebanon, and the old Phoenician cities offered only token resistance to the invader.

(R.D.B./W.L.O./G.R.B.)

LEBANON IN THE MIDDLE AGES

The Maronites

The population of Lebanon did not begin to take its present form until the 7th century AD. At some time in the Byzantine period, a military group of uncertain origin, the Mardaites, established themselves in the north among the indigenous population. From the 7th century onward another group entered the country, the Maronites, a Christian community adhering to the Monothelite doctrine. Forced by persecution to leave their homes in northern Syria, they settled in the northern part of the mountain and absorbed the Mardaites and indigenous peasants to form the present Maronite Church. Originally Syriac-speaking, they gradually adopted the Arabic language although keeping Syriac for liturgical purposes. In south Lebanon Arab tribesmen came in after the Muslim conquest of Syria in the 7th century and settled among the indigenous people. In the 11th century many were converted to the Druze faith, an esoteric offshoot of Shīʿite Islām. South Lebanon became the headquarters of the faith. Groups of Shīʿite Muslims settled on the northern and southern fringes of the mountain and in al-Biqāʿ. In the coastal towns the population became mainly Sunnite Muslim, but in town and country alike there remained considerable numbers of Christians of various sects. In course of time, virtually all sections of the population adopted Arabic, the language of the Muslim states in which Lebanon was included.

Beirut and Mt. Lebanon were ruled by the Umayyads (661–750) as part of the district of Damascus. Despite the occasional rising by the Maronites, Lebanon provided naval forces to the Umayyads in their interminable warfare with the Byzantines. The 8th-century Beirut legist al-Awzāʿī established a school of Islāmic law that heavily influenced Lebanon and Syria. From the 9th to the 11th century coastal Lebanon was usually under the sway of independent Egyptian Muslim dynasties, although the Byzantine Empire attempted to gain portions of the north.

At the end of the 11th century Lebanon became a part of the crusaders' states, the north being incorporated in the county of Tripolis, the south in the kingdom of Jerusalem. The Maronite Church began to accept papal supremacy, while keeping its own patriarch and liturgy.

Despite the strong fortresses of the crusaders, a Muslim reconquest of Lebanon, under the leadership of Egypt, began with the fall of Beirut in 1187. Mongol raids against al-Biqāʿ valley were defeated. Lebanon became part of the Mamlūk state of Egypt and Syria in the 1280s and 1290s and was divided among several provinces. Mamlūk rule, which allowed limited local autonomy to regional leaders, encouraged commerce. The coastal cities, and especially Tripoli, flourished, and the people of the interior were left largely free to manage their own affairs.

OTTOMAN PERIOD

Ottoman expansion began in the area under Selim I (1512–20). He defeated the Mamlūks in 1516–17 and added Lebanon (as part of Mamlūk Syria and Egypt) to his empire. Between the 16th and 18th centuries Ottoman Lebanon evolved a social and political system of its own. Ottoman Aleppo or Tripoli governed the north, Damascus the centre, and Sidon (after 1660) the south. Coastal Lebanon and al-Biqāʿ valley were usually ruled more directly by Istanbul, while Mt. Lebanon enjoyed semiautonomous status. The population took up its present position: the Shīʿites were driven out of the north but increased their strength in the south; many Druze moved from south Lebanon to Jebel Druze (Jabal ad-Durūz) in southern Syria; Maronite peasants, increasing in numbers, moved south into districts mainly populated by Druze. Monasteries acquired more land and wealth. In all parts of the mountains there grew up families of notables, who controlled the land and established a feudal relation with the cultivators; some were Christian, some Druze, who were politically dominant. From them arose the House of Mʿan, which established a princedom over the whole of Mt. Lebanon and was accepted by Christians and Druze alike. Fakhr ad-Dīn II ruled most of Lebanon from 1593 to 1633 and encouraged commerce. When the House of Maʿn died out in 1697, the notables elected as prince a member of the Shihāb family who were Sunnite Muslims but with Druze followers, and this family ruled until 1842. Throughout this period European influence was growing. European trading colonies were established in Saïda and other coastal towns, mainly to trade in silk, the major Lebanese export from the 17th to the 20th century. French political influence was great, particularly among the Maronites, who formally united with the Roman Catholic Church in 1736.

European influence

The 19th century was marked by economic growth, social change, and political crisis. The growing Christian population moved southward and into the towns, and toward the end of the century many of these Christians emigrated to North America, South America, and Egypt. French Catholic and U.S. Protestant mission schools, as well as schools of the local communities, multiplied: in 1866 the American mission established the Syrian Protestant College (later the American University of Beirut), and in 1881 the Jesuits started the Université Saint-Joseph. Such schools produced a literate class, particularly among the Christians, that found employment as professionals. Beirut became a great international port, and its merchant houses established connections with Egypt, the Mediterranean countries, and England.

The growth of the Christian communities upset the traditional balance of Lebanon. The Shihāb princes inclined more and more toward them, and part of the family indeed became Maronites. The greatest of them, Bashīr II (reigned 1788–1840), after establishing his power with the help of Druze notables, tried to weaken them. When the Egyptian troops of Ibrāhīm Pasha occupied Lebanon and Syria in 1831, Bashīr formed an alliance with him to limit the power of the ruling families and to preserve his own power. But Egyptian rule was ended by Anglo-Ottoman intervention, aided by a popular rising in 1840, and Bashīr was deposed. With him the princedom virtually ended; his weak successor was deposed by the Ottomans in 1842, and from that time relations grew worse between the Maronites, led by their patriarch, and the Druze, trying to retain their traditional supremacy. The French supported the Maronites and the British supported a section of the Druze, while the Ottoman government encouraged the collapse of the traditional structure, which would enable it to impose its own direct authority. The conflict culminated in the massacre of Maronites by the Druze in 1860. The complacent attitude of the Ottoman authorities led to direct French intervention on behalf of the Christians. The powers jointly imposed the Organic Regulation of 1861 (modified in 1864), which gave Mt. Lebanon, the axial mountain region, autonomy under a Christian governor appointed by the Ottoman sultan, assisted by a council representing the various communities. Mt. Lebanon prospered under this regime until World War I, when the Ottoman government placed it under strict control, similar to that already established for the coast and al-Biqāʿ valley.

French intervention

FRENCH MANDATE

At the end of the war Lebanon was occupied by Allied forces and placed under a French military administration. In 1920 Beirut and other coastal towns, al-Biqāʿ, and certain other districts were added to the autonomous territory Mt. Lebanon as defined in 1861, to form Greater Lebanon (Grand Liban; subsequently called the Lebanese Republic). In 1923 the League of Nations formally gave the mandate for Lebanon and Syria to France. The Maronites, strongly pro-French by tradition, welcomed this, and during the next 20 years, while France held the mandate, the Maronites were favoured. The expansion of prewar Lebanon into Greater Lebanon, however, changed the balance of the population. Although the Maronites were the largest single element, they no longer formed a majority. The population was more or less equally divided between Christians and Muslims, and a large section of it wanted neither to be ruled by France nor to be part of an independent Lebanon, but rather to form part of a larger Syrian or Arab state. To ease tensions between the communities, the constitution of 1926 provided that each should be equitably represented in public offices. Thus by convention the president of the republic was normally a Maronite, the prime minister a Sunnite Muslim, and the speaker of the chamber a Shīʿite Muslim.

French administration was reasonably efficient. Public utilities and communications were improved, and education was expanded (although higher education was left almost wholly in the hands of religious bodies). Beirut prospered as a centre of trade with surrounding countries, but agriculture was depressed by the decline of the silk industry and the worldwide economic depression. As the middle class of Beirut grew and a real if fragile sense of common national interest sprang up alongside communal loyalties, there grew also the desire for more independence. A Franco-Lebanese treaty of independence and friendship was signed in 1936 but was not ratified by the French government. Lebanon was controlled by the Vichy authorities after the fall of France in 1940 but was occupied by British and Free French troops in 1941. The Free French representative proclaimed the independence of Lebanon and Syria, which was underwritten by the British government. Because of their own precarious position, however, the Free French were unwilling to relax control. In 1943, however, they held elections, which resulted in victory for the Nationalists. Their leader, Bishara al-Khuri, was elected president. The new government passed legislation introducing certain constitutional changes that eliminated all traces of French influence, to which the French objected. On Nov. 11, 1943, the President and almost the entire government were arrested by the French. This led to an insurrection followed by British diplomatic intervention; the French restored the government and transferred powers to it. After another crisis in 1945, agreement was reached on a simultaneous withdrawal of British and French troops. This was completed by the end of 1946, and Lebanon became wholly independent; it had already become a member of the United Nations and the Arab League. (R.D.B./W.L.O.)

LEBANON AFTER INDEPENDENCE

For many years Lebanon maintained its parliamentary democracy, despite serious trials. The main problem for Lebanon was to implement the unwritten power-sharing National Pact of 1943 between the Christians and Muslims. In the early years of independence, so long as no urgent call for pan-Arab unity came from outside, the National Pact faced no serious strains.

The Khuri regime, 1943–52. The Maronite president Bishara al-Khuri closely cooperated with the Sunnite leader Riad as-Sulh, who was premier most of the time. A temporary amendment of the constitution permitted the president, in 1949, a second six-year term. The parliamentary elections of 1947 were blatantly rigged to produce a Parliament favourable to the amendment. This, together with the open favouritism of the President toward his friends and the gross corruption he allegedly condoned, made Khuri increasingly unpopular after his reelection in 1949.

The military coup that overthrew the Kuwatli regime in Syria in March 1949 encouraged the opponents of Khuri in Lebanon. In July 1949 the Syrian Social Nationalist Party (PPS) tried to overthrow the regime by force. The coup failed and its leaders were seized and shot. The PPS took its revenge by securing the assassination of Khuri's premier in 1951. The mounting opposition to the Khuri regime culminated in September 1952 in a general strike that forced his resignation. Camille Chamoun was elected by the Parliament to succeed him.

Lebanon's difficult position

The Chamoun regime and the 1958 crisis. The presidency of Camille Chamoun coincided with the rise of Nasser in Egypt. During the Suez war (October–December 1956), Chamoun earned Nasser's enmity by refusing to break off diplomatic relations with Britain and France, which had joined Israel in attacking Egypt. Chamoun was accused of seeking to align Lebanon with the Western-sponsored Baghdad Pact.

Matters came to a head following the parliamentary elections of 1957, which allegedly were manipulated to produce a Parliament favourable to the reelection of Chamoun. When Syria entered into a union with Egypt, as the United Arab Republic, in February 1958, the Muslim opposition to Chamoun in Lebanon hailed the union as a triumph for pan-Arabism, and there were widespread demands that Lebanon be associated in the union. In May a general strike was proclaimed, and the Muslims of Tripoli rose in armed insurrection. The insurrection spread, and the army was asked to take action against the insurgents. The commanding general, Fuad Chehab, refused to attack them for fear that the army, which was composed of Christians and Muslims, would split apart. The Chamoun government took the issue of external intervention to the United Nations, accusing the United Arab

Republic of intervention, and UN observers were sent to Lebanon. When in July the pro-Western regime in Iraq was toppled in a coup, President Chamoun immediately requested U.S. military intervention, and on the following day U.S. marines landed outside Beirut. The presence of U.S. troops had little immediate effect on the internal situation, but the insurrection slowly faded out. Parliament turned to the commander of the army, General Chehab, to succeed Chamoun as his term ended; Rashid Karami became the new premier.

National unity of the Lebanese people

The Chehab, Hélou, and Franjieh regimes, 1958–76. The crisis had been resolved by compromise, and the Chehab regime was successful in maintaining the compromise and promoting the national unity of the Lebanese people. By his refusal as army commander to take offensive action against the insurgents in 1958, Chehab had earned the confidence of the Muslims. Once in power, he proceeded to allay long-standing Muslim grievances by associating Muslims more closely in the administration and by attending to neglected areas of Lebanon where Muslims predominated. Internal stability was further promoted by the maintenance of good relations with the United Arab Republic, which, even after the Syrian secession in 1961, remained highly popular with the Muslim Lebanese. The economic boom that had begun under the Chamoun regime as the result of the flight of capital from the unstable Arab world into Lebanon continued under the Chehab regime.

Charles Hélou, a former journalist and member of Khuri's Constitutional Bloc, was elected to succeed Chehab in 1964. His government, essentially a weaker version of the Chehab regime, was followed in 1970 by the troubled regime of Suleiman Franjieh.

Despite the apparent calm of the years 1958–69, events in Lebanon moved toward the outbreak of one of the most destructive civil wars in modern history. The essential issues were separate but strongly and immediately affected one another. First came the fact that the political structure of Lebanon was based on a sort of floating consensus—an agreement among leaders of the various religious and ethnic minorities over their respective roles in the state. Under the regimes of Chehab and Hélou, the personal and group ambitions of the various sectors of society were held in check, and the organs of the state, especially the army and the security police, were employed to prevent recourse to violence.

Lebanese social change

During that period, major changes took place in Lebanese society, most significantly urbanization, which brought 40 percent of the Lebanese population to the city of Beirut. But the city, like the country, failed to make its people homogeneous; Beirut became a reflection of Lebanon as a whole. Each quarter took on a religious affiliation, with virtually every village having an enclave established in one suburb or more. And the newcomers suffered from deep and growing social and economic contrasts with their more affluent neighbours. Those who remained in the rural areas lost even more ground, relatively speaking. By the early 1970s, agriculture, in which nearly half of the population was employed, accounted for less than 11 percent of the country's gross domestic product, while the share of urban commerce was rising. An already existing schism was deepened between the groups, which increasingly took on an urban Christian versus rural Muslim coloration.

A second factor, the role of Lebanon in the Arab world, was also a complex issue. Many Syrians still felt that the French decision to separate Lebanon from Syria in 1920 was invalid; many in Lebanon agreed. Lebanon's noninvolvement in the Arab–Israeli wars of 1967 and 1973, its strong and often heavy-handed security policies, and rumours of its secret understandings with Israel all directed attention on this issue.

Palestinian radicalization

Third, the Palestinians thought of Lebanon, after the ruinous Jordanian campaign against them in September 1970, as their last refuge, and by 1973 roughly one person in each 10 in Lebanon was a Palestinian. Yet the Palestinians were constantly made aware of their separate and inferior status; landless and mostly poor, they were exploited as a source of cheap labour. Increasingly their politics became radicalized, and they found common cause with those Lebanese who were poor, rural, and mainly Muslim. As they acquired structure, motivation, and arms during the period after the fall of their base in Jordan, they were sought out as allies by groups in Lebanon. Faced in September 1975 by an Egyptian–Israeli interim agreement on Sinai, the Palestinians concluded that they were being deserted by the Arab states and would shortly be suppressed in Lebanon.

The civil war, 1975–76. *Beginning of the war.* Toward the end of the presidency of Charles Hélou, the Palestinians began to clash with Lebanese security forces. The situation was increasingly complicated by the inability of the Palestine Liberation Organization (PLO) to control the more radical of the factions into which the Palestinians were divided. Under an agreement announced in Cairo on Nov. 3, 1969, the Lebanese government gave the Palestinians virtually a free hand in the refugee camps and at forward posts in the south along the Israeli frontier. In return, the PLO promised not to intervene in Lebanese politics; this was impossible for the Palestinians and not desired by the Lebanese left. When the Lebanese failed to restrain the Palestinians, the Israelis began to raid the south with increasing severity, and this encouraged the Lebanese Christian right, particularly the militant Phalangist Party (with discreet army backing), to attack the Palestinians with its well-organized and well-armed militia.

Franjieh election

It was in this atmosphere that Suleiman Franjieh was elected president, by only one vote in the Parliament, on Aug. 17, 1970. Franjieh was known as a man of personally violent tendencies, and the Lebanese right looked to him to suppress the left and their Palestinian allies. Economic events, however, soon got out of control. Lebanon was caught in a severe inflation that enriched those on one side of the social chasm while it exacerbated the distress and bitterness of those on the other side. The government appeared incapable of addressing this problem; in its attempt to make an appearance of reform, it discarded or destroyed the strong instruments of state control that it had inherited from the aftermath of the civil war of 1958. The more radical of the guerrilla groups, angered by failure of the Lebanese to prevent Israeli incursions and encouraged by the vacillation of the regime, determined upon a series of dramatic attacks. When their plots were foiled, they retaliated by abducting several Lebanese soldiers. Heavy fighting then broke out between the army and the armed refugees in May 1973, and thereafter, despite a series of truces, events moved rapidly toward civil war. By early 1974 the Palestinians were providing the force for the Lebanese National Movement led by Kamal Jumblatt. Into this arena then stepped the relatively deprived Shīʿite Muslims, especially those of the rural al-Biqāʿ valley, under the leadership of a Muslim *imām,* Musa as-Sadr; his Movement of the Deprived was based on a call for Shīʿite political rights, economic justice, and defense against Israeli raids in the mainly Shīʿite south.

Hardly a day passed after the beginning of full civil war in April 1975 without a battle somewhere in Lebanon. The country was torn apart, and the central government virtually ceased to exist. The army, long the mainstay of the government, was immobilized by the nature of the conflict. And the combatants, amply supplied by various foreign groups, turned upon one another with a ferocity—and firepower—almost unequaled in such a small area of the world.

Syrian reversal of position

Final phases of the war. During the early phases of the civil war, the Christians had received encouragement from Pres. Anwar el-Sādāt of Egypt, while President Assad of Syria had encouraged and aided the left and the Palestinians. Gradually the left and the Palestinians began to win the war. By the early months of 1976, it seemed clear that the Christians were losing and that either they would be defeated (so that Lebanon would be reconstituted as a left-dominated, Palestine-oriented state) or Lebanon would be partitioned. Either case appeared to the Syrians likely to bring Israeli intervention. This realization forced a reversal of Syrian policy, first toward restraint on the left and then toward support for the Christians. Ironically, both the Syrians and the Israelis, so opposed to one another on other issues, took up the cause of the Lebanese Christians with essentially the same tools. Syria supplied arms and prevented the Palestinians from taking strategic points, while Israel blockaded Sidon and Tyre (through which arms to the left and the Palestinians had arrived), trained a large military contingent of Lebanese in Israel and supplied it with tanks, and shipped equipment to the Christian sector. During the summer of 1976 the Israeli Army occupied sizable parts of southern Lebanon; from the east, beginning in early June, Syrian military units entered the country with about 450 tanks and 20,000 soldiers.

This realignment led to two new developments. First, the non-Syrian left, represented by Iraq and Libya, resumed close contacts with the PLO, which earlier it had denounced as conservative and ineffectual, and this, in turn, caused President Sādāt of Egypt to seek closer relations with the PLO leader Yāsir ʿArafāt in order to win Palestinian support for his claim to Arab leadership and to prevent attacks on the Egyptians' separate disengagement agreement (September 1975) with the Israelis. Second, the Christians, with strong support from Syrian tanks and soldiers, began to win the civil war. In late June 1976, Christian forces launched attacks on the two Palestinian refugee camps of Jisr al-Bāshā, which fell after one week, and Tall az-Zaatar, which, becoming a symbol of Palestinian resistance, withstood a tragic and bloody siege until mid-August.

In April President Franjieh had signed a constitutional amendment to permit the Parliament to elect a new president. Elias Sarkis was elected on May 8 with the support of Syria but was not inaugurated until September 23. Meanwhile, Lebanon was effectively partitioned along the "Green Line," which passed through the centre of Beirut (east–west) and along the main road to Damascus: to the north was a Christian government, to the south a leftist (Druze–Muslim–Palestinian) government led by Kamal Jumblatt until he was murdered in March 1977. Repeated attempts were made to bring the fighting to an end, until finally a formal "summit" meeting, held on Oct. 25–26, 1976, established an Arab League peacekeeping force of 30,000 troops, mostly Syrian, under the overall command of President Sarkis. By the end of November, despite continued minor clashes, the civil war came to a close.

Postwar fighting

Continued fighting among the Lebanese factions led to the loss of prestige of the former political elite except for that of the Phalangist Party, whose successful leadership of the Christian-rightist coalition so alarmed the Syrian army of occupation that it once again began to support the Muslim–leftist–Palestinian groups in 1978. The destruction and violence had caused hundreds of thousands of Lebanese to flee their homes, particularly from southern Lebanon, where the threat of Israeli intervention stopped Syria from imposing a peace. Israel invaded the area with 20,000 troops on March 14–15, 1978, to destroy Palestinian military bases and to force Lebanon to curb future raids by the PLO into Israel. A small contingent of UN forces replaced the Israelis by June; however, Israel continued to supply arms, money, and troops to the Christians in the south, while the Palestinians soon returned to the same region.

Consequences of the war. The civil war was a catastrophe for the Lebanese, whose country lay in ruins. There seemed to be no compromise acceptable to both the Muslims, who numbered more than one-half of the population, and the Christians, who were determined to keep their control of key government institutions. Foreign intervention merely restrained open, full-scale warfare. Economic destruction was massive, but the chief political problem was the bitterness caused by the thousands of deaths and the ensuing hatreds that promised to destroy the possibility of Lebanese living together again in one nation with one government.

For the Palestinians the war cost perhaps 20,000 killed and twice that many wounded. The Syrians appeared stronger than before, but having got into Lebanon, they faced the problem of extricating themselves. Only Israel among the states of the Middle East appeared to have "won," and the unity of its enemies appeared to be shattered. The Palestinians lost their major bid; Syria acted, apparently, out of fear of Israeli intervention and against its fellow Arabs; and the Lebanese Christians, likely to regain much of their former power, were in Israel's debt. More important, the revulsion and horror of the war had caused Arabs everywhere to question, as never before, the very dream of pan-Arabism.

The Israeli invasion of 1982. The political disintegration of Lebanon led directly to intensified external intervention. Bashir Gemayel, the leader of the Phalangist militia, whose strength derived in part from extensive Israeli aid, forcibly united under his control all the Maronite private armies and thereby created a ministate in East Beirut and the northern coastal sector of Lebanon. The Syrian army was dominant in most of the rest of Lebanon, but a jumble of factions, many of which were armed and paid by outsiders, disputed Syria's power and wreaked havoc because of their internecine quarrels.

Israeli forces bombed PLO headquarters in West Beirut on July 17, 1981, causing in the process more than 300 civilian deaths. This attack led the United States to arrange a cease-fire between the Israelis and the PLO, which, it was hoped, would end raids into northern Israel and would provide the opportunity for President Sarkis of Lebanon to try once again for national reconciliation. The government was again unable to exert its authority, and Palestinian attacks and Israeli counterattacks intensified. The situation erupted on June 6, 1982, when an estimated 60,000 Israeli troops invaded Lebanon.

Although the stated goal of Israel was only to secure the territory north of its border with Lebanon so as to stop PLO raids, Israeli Prime Minister Menachem Begin sought to destroy the PLO and establish in power a Lebanese government that would conclude a peace treaty with Israel along the lines of the Egyptian–Israeli peace of 1979. The invasion was successful: Syrian forces were defeated, the PLO retreated to West Beirut, and Egypt and the other Arab states did little but protest. From late June to August, Israel hesitated to attack PLO and leftist Muslim troops in densely populated West Beirut. Instead, Israel shelled, bombed, and blockaded the area to pressure the PLO and Syrian garrisons to evacuate their forces.

Under supervision by an international (U.S., French, and Italian) force, PLO leaders and troops left Beirut for a number of Arab countries in late August. Because Syria supported the PLO forces remaining in northern Lebanon and in al-Biqāʿ valley, the forces could not be compelled by Israel to leave, but the Syrian backing was used to foster a PLO leadership that opposed Arafat. (In heavy fighting near Tripoli, Arafat was forced into exile in December 1983 for a second time, on this occasion at the instigation of the Syrians.) The Israeli victory in the south and centre was shared by the Phalangists, who then had no barrier to electing their leader president of Lebanon. Bashir Gemayel, however, was assassinated before his inauguration. The Phalangists then secured the election of Bashir's brother, Amin Gemayel, to replace the exhausted and ineffectual Sarkis as president. After West Beirut was occupied by the Israelis, Phalangist militiamen massacred perhaps as many as 1,000 Palestinians in two refugee camps in Beirut in revenge for the death of Bashir Gemayel.

On May 17, 1983, Israel and Lebanon concluded what was very nearly a peace treaty. It called for the withdrawal of Israeli forces and the establishment of bilateral relations. Israel's power in Lebanon deteriorated and Israeli casualties mounted; in September 1983 Israel began withdrawing its forces, first from central Lebanon and then from the south. The international peacekeeping force left Beirut in February 1984 after suffering heavy casualties, and in March Syria and Lebanese Muslims and leftists forced Pres. Amin Gemayel to abrogate the Lebanon–Israel agreement. By June 1985 Israel had withdrawn most of its military from Lebanon, leaving Gemayel and the central government dependent on the dominant Syrians.

This abrupt reversal among the intervening foreign states exacerbated political instability inside Lebanon. The Christian and rightist movement, the Shīʿite–Druze alliance, and the PLO all split asunder over the question of accepting or rejecting Syria's leadership. The cabinet announced on April 30, 1984, represented all Lebanese factions, but it was unable to bring about substantial political reforms or the permanent disarming of private militias, despite intense Syrian pressure to do so. Although the interminable quarrels of the Lebanese continued, with deadly consequences for themselves, external intervention lessened, as Syria attempted to resolve the basic issues that had led to the first outbreak of civil war in 1975.

For later developments in the political history of Lebanon, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (W.L.O.)

BIBLIOGRAPHY. For general discussions of the land and people, see W.B. FISHER, *The Middle East,* 7th ed. (1978); DAVID C. GORDON, *The Republic of Lebanon: Nation in Jeopardy* (1983); and, for bibliography, SHEREEN KHAIRALLAH, *Lebanon* (1979). Economic and social matters are discussed by ABDUL-AMIR BADRUD-DIN, *The Bank of Lebanon* (1984); NADIM G. KHALAF, *The Economic Implications of the Size of Nations; with Special Reference to Lebanon* (1971); HUDA C. ZURAYK and HAROUTUNE K. ARMENIAN, *Beirut 1984* (1985); JOSEPH CHAMIE, *Religion and Fertility: Arab Christian-Muslim Differentials* (1981); YUSIF A. SAYIGH, *Entrepreneurs of Lebanon* (1962), a study of the role of entrepreneurs in the national development of the country; ANNE H. FULLER, *Buarij: Portrait of a Lebanese Muslim Village* (1961, reissued 1968); LILIANE GERMANOS-GHAZALY, *Le Paysan, la terre et la femme: organisation sociale d'un village du Mont-Liban* (1978); FRIEDRICH RAGETTE (ed.), *Beirut of Tomorrow: Planning for Reconstruction* (1983); and JOHN GULICK, *Tripoli: A Modern Arab City* (1967). Useful discussions of Lebanese government include MICHAEL W. SULEIMAN, *Political Parties in Lebanon: The Challenge of a Fragmented Political Culture* (1967); GEORGE GRASSMUCK and KAMAL SALIBI, *Reformed Administration in Lebanon,* 2nd ed. (1964); ADEL A. FREIHA, *L'Armée et l'état au Liban, 1945–1980* (1980); and R.D. MCLAURIN, "Lebanon and Its Army: Past, Present, and Future," in EDWARD E. AZAR *et al., The Emergence of a New Lebanon: Fantasy or Reality?,* pp. 79–114 (1984). Cultural matters are discussed by LAWRENCE I. CONRAD, "Culture and Learning in Beirut," *The American Scholar,* 52:463–478 (autumn 1983); and FRIEDRICH RAGETTE, *Architecture in Lebanon: The Lebanese House During the 18th and 19th Centuries* (1974, reprinted 1980).

For ancient history, see *Cambridge Ancient History,* especially vol. 1, part 1, *Prolegomena and Prehistory,* 3rd ed. (1970); vol. 1, part 2, *Early History of the Middle East,* 3rd ed. (1971); vol. 2, part 1, *History of the Middle East and the Aegean Region c. 1800–1380 B.C.,* 3rd ed. (1973); and vol. 3, part 3, *The Expansion of the Greek World, Eighth to Sixth Centuries B.C.,* 2nd ed. (1982); and DONALD HARDEN, *The Phoenicians,* rev. ed. (1971). See also MAURICE DUNAND, *Byblos: Its History, Ruins and Legends,* 2nd ed. (1968; originally published in French, 2nd ed., 1968, reissued 1973); FRIEDRICH RAGETTE, *Baalbek* (1980); F.M. HEICHELHEIM, "Roman Syria," in TENNEY FRANK (ed.), *An Economic Survey of Ancient Rome,* vol. 4, pp. 121–257 (1938, reprinted 1975); J.-P. REY-COQUAIS, "Syrie Romaine, de Pompée à Dioclétien," *Journal of Roman Studies,* 68:44–73 (1978); HILDEGARD TEMPORINI and WOLFGANG HAASE (eds.), *Aufstieg und Niedergang der Römischen Welt,* vol. 2, part 8, *Politische Geschichte: Provinzen und Rundvölker: Syrien, Palästina, Arabien,* pp. 3–294 (1977); and NINA JIDEJIAN, *Byblos Through the Ages* (1968), *Tyre Through the Ages* (1969), *Sidon Through the Ages* (1971), *Beirut Through the Ages* (1973), and *Baalbek: Heliopolis, City of the Sun* (1975).

For medieval and modern history the two most important works are PHILIP K. HITTI, *Lebanon in History: From the Earliest Times to the Present,* 3rd ed. (1967); and KAMAL S. SALIBI, *The Modern History of Lebanon* (1965, reissued 1977). The Ottoman period is discussed by ABDUL-RAHIM ABU-HUSAYN, *Provincial Leaderships in Syria, 1575–1650* (1985); DOMINIQUE CHEVALLIER, *La Société du Mont Liban à l'époque de la révolution industrielle en Europe* (1971, reissued 1982); ILIYA F. HARIK, *Politics and Change in a Traditional Society: Lebanon, 1711–1845* (1968); and P.M. HOLT, *Egypt and the Fertile Crescent: A Political History, 1516–1922* (1966, reprinted 1969).

For the 20th century, see ALBERT H. HOURANI, *Arabic Thought in the Liberal Age, 1798–1939* (1962, reissued 1983), and *Syria and Lebanon: A Political Essay* (1946, reprinted 1968); STEPHEN HELMSLEY LONGRIGG, *Syria and Lebanon Under French Mandate* (1958, reissued 1972); and MICHAEL C. HUDSON, *The Precarious Republic: Political Modernization in Lebanon* (1968, reissued 1985). On the civil war and subsequent events, see WALID KHALIDI, *Conflict and Violence in Lebanon* (1979, reprinted 1983); KAMAL S. SALIBI, *Crossroads to Civil War: Lebanon, 1958–1976* (1976); HELENA COBBAN, *The Making of Modern Lebanon* (1985); DAVID GILMOUR, *Lebanon, the Fractured Country,* rev. ed. (1984); and ITAMAR RABINOVICH, *The War for Lebanon, 1970–1985,* rev. ed. (1985).

(S.G.K./C.F.M./R.D.B./W.L.O./G.R.B.)

The Evolution of Modern Western Legal Systems

The three great law families of modern Western civilization are civil law (also called Romano-Germanic law), common law (also called Anglo-American law), and Socialist law. They are descended from ancient Roman law and ancient Germanic tribal law and have been altered by various customary, ecclesiastical, feudal, commercial, and modern sociopolitical influences.

Most of the laws of the legal systems of continental Europe are traditionally classified as civil law. This type of law spread to Latin America and, later, to those countries of Asia and Africa that found it necessary to westernize their laws, such as Japan (in which American law has also had great influence), Thailand, Turkey, and Ethiopia. It also prevails, supplemented by religious or customary laws, in those regions that were colonies, protectorates, or trust territories of France, Belgium, The Netherlands, Spain, Portugal, and Italy, including Morocco, Algeria, Zaire, Indonesia, and Somalia, as well as in some dependencies or outlying parts of European countries, such as Martinique and Curaçao in the West Indies. Civil law, supplemented increasingly by Islāmic law, has come to prevail in several countries of the Middle East and North Africa.

Common law is, essentially, the law of England and the law of those countries in which the law of England has been received or implanted. Although common law was often transformed in certain respects or supplemented by local or religious traditions, its principal features have been preserved. Common-law countries that have remained closest to the English tradition include Canada, Australia, New Zealand, the Republic of Ireland, and the present and former British areas of the West Indies. Common law also prevails in India, Pakistan, Burma, Malaysia, and Singapore, where it is supplemented in matters of personal status by religious laws. In Liberia and in most of those parts of Africa and Oceania that were British colonies, protectorates, or trust territories, common law is supplemented by native customs. The United States has developed a type of common law that differs in many respects from the English.

Civil law and common law have merged in the legal systems of several countries (Scotland, Puerto Rico, the Philippines, Israel, South Africa, Zimbabwe, Sri Lanka, Mauritius, and the Seychelles) and in other localized regions (Quebec, in Canada, and Louisiana, in the United States). In most of these places private law still follows a civil-law pattern. The laws of the Nordic countries (Sweden, Finland, Denmark, Norway, and Iceland) are also closer to civil law than to common law.

In the Soviet Union and the other Socialist countries of eastern Europe, all law is permeated with the collectivist spirit, and the aim of law is seen as assisting in the creation of a new social order. These systems have diverged from their Romano-Germanic roots in important respects, but many concepts of civil law are still used in Socialist legal thinking.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Five, Division V, especially Section 551.

This article is divided into the following sections:

The Western legal heritage

ROMAN LAW

In its strictest sense the term Roman law denotes the law of the city of Rome and of the Roman Empire that was in force during the period from the foundation of the city (traditional date 753 BC) until the fall of the Western Empire in the 5th century AD and the fall of the Eastern Empire in 1453. The term Roman law today, however, often refers to more than the laws of Roman society. The legal institutions evolved by the Romans had influence on the laws of other peoples in times long after the disappearance of the Roman Empire and in countries that were never subject to Roman rule. To take the most striking example, in a large part of Germany, until the adoption of a common code for the whole empire in 1900, the Roman law was in force as "subsidiary law"; that is, it was applied unless excluded by contrary local provisions. This law, however, which was in force in parts of Europe long after the fall of the Roman Empire, was not the Roman law in its original form. Although its basis was indeed the Corpus Juris Civilis—the codifying legislation of the emperor Justinian I—this legislation had been interpreted, developed, and adapted to later conditions by generations of jurists from the 11th century onward and had received additions from non-Roman sources.

The range of influence of Roman law

Development of the jus civile and jus gentium. In the great span of time during which the Roman Republic and Empire existed, there were many phases of legalistic development. During the period of the republic (753–31 BC), the *jus civile* (civil law) developed. Based on custom or legislation, it applied exclusively to Roman citizens. By the middle of the 3rd century BC, however, another type of law, *jus gentium* (international law), was developed by the Romans to be applied both to themselves and to foreigners. *Jus gentium* was not the result of legislation, but was, instead, a development of the magistrates and governors who were responsible for administering justice in cases in which foreigners were involved. The *jus gentium* became, to a large extent, part of the massive body of law that was applied by magistrates to citizens, as well as to foreigners, as a flexible alternative to *jus civile.*

Roman law, like other ancient systems, originally adopted the principle of personality—that is, that the law of the state applied only to its citizens. Foreigners had no rights and, unless protected by some treaty between their state and Rome, they could be seized like ownerless pieces of property by any Roman. But from early times there were treaties with foreign states guaranteeing mutual protection. Even in cases in which there was no treaty, the increasing commercial interests of Rome forced it to protect, by some form of justice, the foreigners who came within its borders. A magistrate could not simply apply Roman law because that was the privilege of citizens; even had there not been this difficulty, foreigners would probably have objected to the cumbersome formalism that characterized the early *jus civile.*

The law that the magistrates applied probably consisted of three elements: (1) an existing mercantile law that was used by the Mediterranean traders; (2) those institutions of the Roman law that, after being purged of their formalistic elements, could be applied universally to any litigant, Roman or foreigner; and (3) in the last resort, a magistrate's own sense of what was fair and just. This system of *jus gentium* was also adopted when Rome began to acquire provinces so that provincial governors could administer justice to the *peregrini* (foreigners). This word came to mean not so much persons living under another government (of which, with the expansion of Roman power, there came to be fewer and fewer) as Roman subjects who were not citizens. In general, disputes between members of the same subject state were settled by that state's own courts according to its own law, whereas disputes between provincials of different states or between provincials and Romans were resolved by the governor's court applying *jus gentium.* By the 3rd century AD, when citizenship was extended throughout the empire, the practical differences between *jus civile* and *jus gentium* ceased to exist. Even before this, when a Roman lawyer said that a contract of sale was *juris gentium,* he meant that it was formed in the same way and had the same legal results whether the parties to it were citizens or not. This became the practical meaning of *jus gentium.* Because of the universality of its application, however, the idea was also linked with the theoretical notion that it was the law common to all peoples and was dictated by nature—an idea that the Romans took from Greek philosophy.

Written and unwritten law. The Romans divided their law into *jus scriptum* (written law) and *jus non scriptum* (unwritten law). By "unwritten law" they meant custom; by "written law" they meant not only the laws derived from legislation but, literally, laws based on any written source.

Types of written law

There were various types of written law, the first of which consisted of *leges* (singular *lex*), or enactments of one of the assemblies of the whole Roman people. Although the wealthier classes, or patricians, dominated these assemblies, the common people, or plebeians, had their own council in which they enacted resolutions called *plebiscita.* Only after the passage of the Lex Hortensia in 287 BC, however, did *plebiscita* become binding on all classes of citizens; thereafter, *plebiscita* were generally termed *leges* along with other enactments. In general, legislation was a source of law only during the republic. When Augustus established the empire in 31 BC, the assemblies did not at once cease to function, but their assent to any proposal became merely a formal ratification of the emperor's wishes. The last known *lex* was passed during the reign of Nerva (AD 96–98).

The earliest and most important legislation, or body of *leges,* was the Twelve Tables, enacted in 451–450 BC during the struggle of the plebeians for political equality. It represented an effort to obtain a written and public code that patrician magistrates could not alter at will against plebeian litigants. Little is known of the actual content of the Twelve Tables; the text of the code has not survived, and only a few fragments are extant, collected from allusions and quotations in the works of authors such as Cicero. From the fragments it is apparent that numerous matters were treated, among them family law, delict (tort, or offense against the law), and legal procedure.

A second type of written law consisted of the *edicta* (edicts), or proclamations issued by a superior magistrate (praetor) on judicial matters. The office of praetor was created in 367 BC to take over the expanding legal work involving citizens; later, a separate praetor was created to deal with foreigners. Upon taking office, a praetor issued an edict that was, in effect, the program for his year in office. The curule aediles, who were the magistrates responsible for the care and supervision of the markets, also issued edicts. During the later stages of the republic, these praetorian and magisterial edicts became an instrument of legal reform, and *leges* ceased to be a major source of private law.

Law-making power of magistrates

The Roman system of procedure gave the magistrate great powers for providing or refusing judicial remedies, as well as for determining the form that such remedies should take. The result of this magisterial system was the development of a new body of rules that existed alongside, and often superseded, the civil law. The *edicta* remained a source of law until about AD 131, when the emperor Hadrian commissioned their reorganization and consolidation and declared the resulting set of laws to be unalterable, except by the emperor himself.

A third type of written law was the *senatus consulta,* or resolutions of the Roman senate. Although these suggestions to various magistrates had no legislative force during the republic, they could be given force by the magistrates' edicts. In the early empire, as the power of the assemblies declined and the position of the emperor increased, *senatus consulta* became resolutions that endorsed the proposals of the emperor. As the approval of the Senate became increasingly automatic, the emperor's proposals became the true instrument of power. Consequently, emperors ceased referring proposals to the Senate and, not long after the early imperial period, ended the practice of legislating through the Senate.

A fourth type of written law consisted of the *constitutiones principum,* which were, in effect, expressions of the legislative power of the emperor. By the middle of the 2nd century AD, the emperor was, essentially, the sole creator of the law. The chief forms of imperial legislation were edicts or proclamations; instructions to subordinates, especially provincial governors; written answers to officials or others who consulted the emperor; and decisions of the emperor sitting as a judge.

The last type of written law was the *responsa prudentium,* or answers to legal questions given by learned lawyers to those who consulted them. Although law, written and unwritten, was originally a rather secretive monopoly of the college of pontiffs, or priests, a recognizable class of legal advisers, *juris consulti* or *prudentes,* had developed by the early 3rd century BC. These legal advisers were not professionals as such but men of rank who sought popularity and advancement in their public careers by giving free legal advice. They interpreted statutes and points of law, especially unwritten law, advised the praetor on the content of his edict, and assisted parties and judges in litigation. Augustus empowered certain jurists to give *responsa* with the emperor's authority; this increased their prestige, but the practice lapsed as early as AD 200.

During the early empire, numerous commentaries were written by the great jurists on individual *leges,* on civil law, on the edict, and on law as a whole. In the 5th century a law was passed stipulating that only the works of certain jurists could be cited. Legal scholarship declined in the postclassical period.

The law of Justinian. When the Byzantine emperor Justinian I assumed rule in AD 527, he found the law of the Roman Empire in a state of great confusion. It consisted of two masses that were usually distinguished as old law and new law.

Old law and new law under Justinian

The old law comprised (1) all of the statutes passed under the republic and early empire that had not become obsolete; (2) the decrees of the Senate passed at the end of the republic and during the first two centuries of the empire; and (3) the writings of jurists and, more particularly, of those jurists to whom the emperors had given the right of declaring the law with their authority. These jurists, in their commentaries, had incorporated practically all that was of importance. Of these numerous records and writings of old law, many had become scarce or had been lost altogether, and some were of doubtful authenticity. The entire mass of work was so costly to produce that even the public libraries did not contain complete collections. Moreover, these writings contained many inconsistencies.

The new law, which consisted of the ordinances of the emperors promulgated during the middle and later stages of the empire, was in a similarly disorganized condition. These ordinances or constitutions were extremely numerous and contradictory. Because no complete collection existed (earlier codices were not comprehensive), other ordinances had to be obtained separately. It was thus necessary to collect into a reasonable corpus as much of the law, both new and old, as was regarded as binding and to purge its contradictions and inconsistencies.

Immediately after his accession, Justinian appointed a commission to deal with the imperial constitutions. The 10 commissioners went through all of the constitutions of which copies existed, selected those that had practical value, cut all unnecessary matter, eliminated contradictions by omitting one or the other of the conflicting passages, and adapted all the provisions to the circumstances of Justinian's own time. The resulting Codex Constitutionum was formally promulgated in 529, and all imperial ordinances not included in it were repealed. This Codex has been lost, but a revised edition of 534 exists as part of the so-called Corpus Juris Civilis.

The success of this first experiment encouraged the Emperor to attempt the more difficult enterprise of simplifying and digesting the writings of the jurists. Thus, beginning in 530, a new commission of 16 eminent lawyers set about this task of compiling, clarifying, simplifying, and ordering; the results were published in 533 in 50 books that became known as the Digest (Digesta) or Pandects (Pandectae). After enacting the Digest as a lawbook, Justinian repealed all of the other law contained in the treatises of the jurists and directed that those treatises should never be cited in the future, even by way of illustration; at the same time, he abrogated all of the statutes that had formed a part of the old law. An outline of the elements of Roman law called the Institutes of Justinian (or simply Institutiones) was published at about the same time.

Between 534 and his death in 565, Justinian himself issued a great number of ordinances that dealt with many subjects and seriously altered the law on many points. These ordinances are called, by way of distinction, new constitutions (Novellae Constitutiones Post Codicem); in English they are referred to as the Novels.

Corpus Juris Civilis: the civil-law code of Justinian

All of these books—the revised Codex Constitutionum (the original work was revised four and a half years later), the Digest, the Institutes, and the Novels—are collectively known as the Corpus Juris Civilis. This Corpus Juris of Justinian, with a few additions from the ordinances of succeeding emperors, continued to be the chief lawbook in what remained of the Roman world. In the 9th century a new system known as the Basilica was prepared by the emperor Leo VI the Wise. It was written in Greek and consisted of parts of the Codex and parts of the Digest, joined and often altered in expression, together with some material from the Novels and imperial ordinances subsequent to those of Justinian. In the western provinces, the law as settled by Justinian held its ground.

Categories of Roman law. *The law of persons.* "The main distinction in the law of persons," said the 2nd-century jurist Gaius, "is that all men are either free or slaves." The slave was, in principle, a human chattel who could be owned and dealt with like any other piece of property. As such, he was not only at the mercy of his owner but rightless and (apart from criminal law) dutiless. Even though the slave was in law a thing, he was in fact a man, and this modified the principle. A slave could not be a party to a contract nor own property, but he could be given a de facto patrimony, which could be retained if he were freed; if he made a "commitment," it could ultimately be enforced against his master. A manumitted slave became, in most instances, not only free but also a citizen.

The definition of citizenship was important for the purposes of private law because certain parts applied only to citizens (*jus civile*). Noncitizens could be either Latini, inhabitants of Roman settlements that had the rights of members of the original Latin League, or *peregrini,* who were members of foreign communities or of those territories governed but not absorbed by Rome. The great extension of the citizenship by the emperor Caracalla in AD 212 reduced the importance of this part of the law.

Patria potestas: the power of the paternal head of the Roman family

Family. The chief characteristic of the Roman family was the *patria potestas* (paternal power in the form of absolute authority), which the elder father exercised over his children and over his more remote descendants in the male line, whatever their age might be, as well as over those who were brought into the family by adoption—a common practice at Rome. Originally this meant not only that he had control over his children, even to the right of inflicting capital punishment, but that he alone had any rights in private law. Thus, any acquisitions made by a child under *potestas* became the property of the father. The father might indeed allow a child (as he might a slave) certain property to treat as his own, but in the eye of the law it continued to belong to the father.

By the 1st century AD there were already modifications of the system: the father's power of life and death had shrunk to that of light chastisement, and the son could bind his father by contract with a third party within the same strict limits that applied to slaves and their masters. Sons also could keep as their own what they earned as soldiers and even make wills of it. In Justinian's day, the position regarding property had changed considerably. What the father gave to the son still remained, in law, the father's property, but the rules concerning the son's own earnings had been extended to many sorts of professional earnings; and in other acquisitions (such as property inherited from the mother), the father's rights were reduced to a life interest (usufruct). Normally, *patria potestas* ceased only with the death of the father; but the father might voluntarily free the child by emancipation, and a daughter ceased to be under her father's *potestas* if she came under the *manus* of her husband.

Marriage with and without *manus,* or husbandly authority

There were two types of marriage known to the law, one with *manus* and one without, but the *manus* type of marriage was rare even in the late republic and had disappeared long before Justinian's day. *Manus* was the autocratic power of the husband over the wife, corresponding to *patria potestas* over the sons.

Marriage without *manus* was by far the more common in all properly attested periods. It was formed (provided the parties were above the age of puberty and, if under *potestas,* had their father's consent) simply by beginning conjugal life with the intention of being married, normally evidenced by the bringing of the bride to the bridegroom's house. The wife remained under her father's *potestas* if he were still alive; if he were dead, she continued (as long as guardianship of women continued) to have the same guardian as before marriage. Both spouses had to be citizens, or if one was not, he or she must have *conubium* (the right, sometimes given to non-Romans, of contracting a Roman marriage). In marriage without *manus,* the property of the spouses remained distinct, and even gifts between husband and wife were invalid.

Divorce was always possible at the instance of the husband in cases of marriage with *manus;* in marriage without *manus,* either party was free to put an end to the relationship at will. A formal letter was usually given to the spouse, but any manifestation of intention to end the relationship—an intention made clear to the other party and accompanied by actual parting—was all that was legally necessary. The Christian emperors imposed penalties on those who divorced without good reason, but the power of the parties to end the marriage by their own act was not taken away.

Concubinage was recognized in the empire as a "marriage" without a dowry, with a lower status for the woman, and with provisions that the children were not legally the father's heirs. A man could not have both a wife and a concubine. In the 4th century the emperor Constantine first enacted a law enabling the children of such unions to be legitimated by the subsequent marriage of their parents. Medieval civil law extended this rule to all illegitimate children.

Guardianship

Persons under the age of puberty (14 for males, 12 for females) needed *tutores* if they were not under *patria potestas.* Such tutors could be appointed under the will of the father or male head of the household. Failing such an appointment, the guardianship went to certain prescribed relatives; if there were no qualified relations, the magistrates appointed a tutor. Originally, children were considered adults at the age of puberty; but, after a long development, it became usual for those between the ages of puberty and 25 to have guardians who were always magisterially appointed. Originally, all women not under *patria potestas* or *manus* also needed *tutores,* appointed in the same way as those for children. By the early empire, this provision was little more than a burdensome technicality, and it disappeared from Justinian's law.

Corporations. The Romans did not develop a generalized concept of juristic personality in the sense of an entity that had rights and duties. They had no terms for a corporation or a legal person. But they did endow certain aggregations of persons with particular powers and capacities, and the underlying legal notion hovered between corporate powers, as understood in modern law, and powers enjoyed collectively by a group of individuals. The source of such collective powers, however, was always an act of state.

Four types of corporation were distinguished:

1. *Municipia* (the citizen body, originally composed of the conquered cities and later of other local communities) possessed a corporateness that was recognized in such matters as having the power to acquire things and to contract. In imperial times, they were accorded the power to manumit slaves, take legacies, and finally—though this

became general only in postclassical law—to be instituted as an heir.

2. The *populus Romanus,* or the "people of Rome," collectively could acquire property, make contracts, and be appointed heir. Public property included the property of the treasury.

3. *Collegia*—numerous private associations with specialized functions, such as craft or trade guilds, burial societies, and societies dedicated to special religious worship—seem to have carried on their affairs and to have held property corporately in republican times. The emperors, viewing the *collegia* with some suspicion, enacted from the beginning that no *collegium* could be founded without state authority and that their rights of manumitting slaves and taking legacies be closely regulated.

4. Charitable funds became a concern of postclassical law. Property might be donated or willed—normally, but not necessarily, to a church—for some charitable use, and the church would then (or so it appears from the evidence) have the duty of supervising the fund. Imperial legislation controlled the disposition of such funds so that they could not be used illegally. In such cases ownership is thought to have been temporarily vested in the administrators.

The Roman concept of absolute ownership

The law of property and possession. In Roman law (today as well as in Roman times), both land and movable property could be owned absolutely by individuals. This conception of absolute ownership (*dominium*) is characteristically Roman, as opposed to the relative idea of ownership as the better right to possession that underlies the Germanic systems and English law.

Mancipatio, or formal transfer of property, involved a ceremonial conveyance needing for its accomplishment the presence of the transferor and transferee, five witnesses (adult Roman citizens), a pair of scales, a man to hold them, and an ingot of copper. The transferee grasped the object being transferred and said, "I assert that this thing is mine by Quiritarian [Roman] law; and be it bought to me with this piece of copper and these copper scales." He then struck the scales with the ingot, which he handed to the transferor "by way of price." Clearly, this was a symbolical sale and the relic of a real sale.

In jure cessio was a conveyance in the form of a lawsuit. The transferee claimed before the magistrate that the thing was his, and the transferor, who was the defendant, admitted the claim. The magistrate then adjudged the thing to the transferee. (The sham-lawsuit theory, however, is not acceptable to all modern scholars, principally because the judgment of ownership was valid against any possible private claimant, not merely against the defendant, as in a true lawsuit.)

Usucapio referred to ownership acquired by length of possession. In early Roman law, two years of continuous possession established title in the case of land, one year in the case of movables. In the developed law, possession must have begun justifiably in good faith, and the thing must not have been stolen (even though the possessor himself may have been innocent of the theft) or acquired by violence.

In terms of *occupatio,* ownerless things that were susceptible to private ownership (excluding such things as temples) became the property of the first person to take possession of them. This applied to things such as wild animals and islands arising in the sea. In some views, it also applied to abandoned articles.

Accessio worked in this manner: if an accessory thing belonging to A was joined to a principal one belonging to B, the ownership in the whole went to B. For example, if A's purple were used to dye B's cloth, the dyed cloth belonged wholly to B. By far the most important application of this rule asserted that whatever is built on land becomes part of the land and cannot be separately owned.

Specificatio was somewhat different. If A made a thing out of material belonging to B, one school of thought held that ownership went to A, and another held that it remained with B. Justinian adopted a "middle opinion": B retained ownership if reconversion to the original condition was possible (a bronze vase could be melted down); A obtained ownership if it was not (wine cannot be reconverted into grapes).

According to *thesauri inventio,* or treasure trove, the final rule was that if something was found by a man on his own land, it went to him; if it was found on the land of another, half went to the finder, half to the landowner.

Traditio was the simple delivery of possession with the intention of passing ownership and was the method of conveyance of the *jus gentium.* If A sold and merely delivered a slave to B, under the *jus civile,* A remained the owner of the slave until a specified length of time had elapsed. The praetors, however, devised procedural methods of protecting B's possession in such a way that A's title became valueless, and B was said to own the thing *in bonis.* This was a remarkable triumph for informality in the granting of title. From the phrase *in bonis,* later writers coined the expression "bonitary ownership." Justinian abolished the theoretical distinction between civil and bonitary ownership.

The ordinary leaseholder had no protection beyond a contractual right against a landlord and could not assign tenancy. But there were certain kinds of tenure that did provide the tenant protection and that were assignable: agricultural and building leases granted for a long term or in perpetuity often enabled leaseholders to enjoy rights hardly distinguishable from ownership.

Servitudes: rights to specific, limited use or enjoyment of another's property

There were also servitudes, in which one person enjoyed certain rights in property owned by another. Rights of way and water rights were rustic servitudes; rights to light or to view were urban servitudes. *Ususfructus* was the right to use and take the fruits (such as crops) of a thing and corresponded to the modern notion of life interest. A more restricted right, likewise not extending beyond the life of the holder, *usus* permitted merely the use of a thing; thus, a person could live in a house but could not let it, as that would be equivalent to "taking the fruits."

Since ownership was absolute, it was sharply distinguished from possession, which the civil law did not protect as such. Any owner wishing to interfere with an existing possessor, however, had to bring legal action to prove his title. If he interfered on his own authority, the praetor would see that the original state of affairs was restored before adjudicating the title.

Delict and contract. Obligations were classified by classical jurists into two main categories, according to whether they arose from delict or contract. Justinian's law recognized two further classes of obligation, termed quasi-delict and quasi-contract.

As early as the 6th and 5th centuries BC, Roman law was experiencing a transition from a system of private vengeance to one in which the state insisted that the person wronged accept compensation instead of vengeance. Thus, in the case of assault (*injuria*), if one man broke another's limb, *talio* was still permitted (that is, the person wronged could inflict the same injury as he had received); but in other cases, fixed monetary penalties were set. Theft involved a penalty of twice the value of the thing stolen, unless the thief was caught in the act, in which case he was flogged and "adjudged" to the person wronged.

By the early empire, reforms had substituted a fourfold penalty in the case of a thief who was caught in the act, and the court assessed all penalties for *injuria* (which by then included defamation and insulting behaviour). The law of damage to property was regulated by statute (the Lex Aquilia), which in turn was much extended by interpretation. Additionally, there were situations in which a person could be held liable for damages even though he was not personally responsible.

In the early republic, a law of contract hardly existed. There was, however, an institution called *nexum,* of which little can be said with certainty except that it was a kind of loan so oppressive in character that it could result in the debtor's complete subjection to the creditor. It was obsolete long before imperial times. The contracts of classical law were divided into four classes: literal, verbal, real, and consensual. The literal contract was a type of fictitious loan formed by an entry in the creditor's account book; it was comparatively unimportant and was obsolete by Justinian's day. The verbal contract, or *stipulatio,* was of great importance, for it established a form in which any agreement (provided it was lawful and possible) could be made binding by the simple method of reducing it to question and answer: "Do you promise to pay me 10,000 sesterces?" "I promise." Originally it was absolutely necessary that the words be spoken, but by Justinian's day a written memorandum of such a contract would be binding, even though, in fact, nothing at all had been spoken.

Enforceable contracts

If an agreement was not clothed in the form of a stipulation, it must, to be valid, fall under one of the types of real or consensual contracts. A real contract was one requiring that something should be transferred from one party to the other and that the obligation arising should be for the return of that thing. Real contracts included loans of money, loans of goods, deposits, and pledges. Consensual contracts needed nothing except verbal or written agreement between the parties, and though there were only four such contracts known to the law, they were the most important in ordinary life—sale, hire of things or services, partnership, and mandate (acting upon instructions). In Justinian's day there was a further principle that in any case of reciprocal agreement, such as an agreement for exchange (but not sale), if one party had performed, he could bring an action to enforce performance by the other. In addition to the foregoing contracts, a few other specific agreements were recognized as enforceable, but the general recognition of all serious agreements as binding was never achieved by the Romans.

Quasi-delict covered four types of harm, grouped together by no clearly ascertainable principle. They included the action against an occupier for harm done by things thrown or poured from his house into a public place and the action against a shipowner, innkeeper, or stablekeeper for loss caused to customers on the premises through theft or damage by persons in his service.

Quasi-contract embraced obligations that had no common feature save that they did not properly fall under contract, because there was no agreement, or under delict, because there was no wrongful act. The most noticeable examples were, first, *negotiorum gestio,* which enabled one who intervened without authority in another's affairs for the latter's benefit to claim reimbursement and indemnity, and second, the group of cases in which an action (*condictio*) was allowed for the recovery by A from B of what would otherwise be an unjustified enrichment of B at A's expense, such as when A had mistakenly paid B something that was not due (*condictio indebiti*). This notion of unjust enrichment as a source of legal obligation was one of the most pregnant contributions made by Roman law to legal thought.

The law of succession. The first requirement of any Roman will of historical times was the appointment of one or more heirs. An heir, in the Roman sense of the term, was a universal successor; that is, he took over the rights and duties of the deceased (insofar as they were transmissible at all) as a whole. On acceptance, the heir became owner if the deceased was owner, creditor if he was creditor, and debtor if he was debtor, even though the assets were insufficient to pay the debts. It was thus possible for an inheritance to involve the heir in a loss. Until Justinian's day this consequence could be avoided only by not accepting the inheritance, but Justinian made one of his most famous reforms by providing that an heir who made an inventory of the deceased's assets need not pay out more than he had received. Freedom of testation, furthermore, was not complete: a man was obliged to leave a certain proportion of his property to his children and in some cases to ascendants and brothers and sisters.

With regard to intestate succession, or succession without a will, those first entitled in early times were the deceased's own heirs—that is, those who were in his *potestas* or *manus* when he died and who were freed from that power at his death. Failing these heirs, the nearest agnatic relations (relations in the male line of descent) succeeded, and, if there were no agnates, the members of the gens, or clan, of the deceased succeeded. Later reforms placed children emancipated from *potestas* on an equal basis with those under *potestas* and gradually gave the surviving spouse (in marriage without *manus*) greater rights of succession. By Justinian's day the system had evolved as follows: descendants had the first claim, and failing these heirs, came a composite class consisting of ascendants, brothers and sisters of full blood, and children of deceased brothers and sisters. Next came brothers and sisters of the half blood and, finally, the nearest cognates (relations in the female line). Husband and wife were not mentioned, but their old rights were kept alive in the absence of any of the preceding categories. Justinian also gave a "poor" widow a right to one-quarter of her husband's estate unless there were more than three children, in which case she shared equally with them. If, however, the heirs were her own children by the deceased, she received only a *ususfructus* (life interest) in what she took.

The law of procedure. The earliest forms of procedure had two stages: a preliminary one before the jurisdictional magistrate, in which the issue was developed; and then the actual trial before the *judex,* or judge. The system required that set forms of words be spoken by the parties and, sometimes, by the magistrate. The parties making an assertion of ownership, for instance, would grasp the thing in dispute and lay a wand on it, after which the magistrate would intervene and say, "Let go, both of you." So formal was the procedure that a plaintiff who made the slightest mistake lost his case. Under the system the plaintiff was also responsible for physically producing the defendant in court and, often, for carrying out the sentence of the court.

Instructions to the judge

Under new procedures developed in the 2nd and 1st centuries BC, the issue at the magisterial stage was formulated in written instructions to the *judex,* couched in the form of an alternative: "If it appears that the defendant owes the plaintiff 10,000 sesterces, the *judex* is to condemn the defendant to pay the plaintiff 10,000 sesterces; if it does not so appear, he is to absolve him." A draft of these written instructions was probably prepared for the plaintiff before he came into court, but there could be no trial until it was accepted by the defendant, for there was always a contractual element about a lawsuit under both the new and the old systems. Pressure, however, could be exercised by the magistrate on a defendant who refused to accept instructions that the magistrate had approved, just as a plaintiff could be forced to alter instructions that the magistrate had disapproved, by the magistrate's refusal to otherwise give the order to the *judex* to decide the case.

In late republican times, still another system developed, first in the provinces, then in Rome. Under the new system the magistrate used his administrative powers, which were always considerable, for the purpose of settling disputes. He could command: thus if one person brought a complaint against another before him, he could investigate the matter and give the order he thought fit. As imperially appointed officers superseded republican magistrates, this administrative process became more common. The result was that the old contractual element in procedure disappeared as did the old two-stage division. Justice was now imposed from above by the state—not, as originally, left to a kind of voluntary arbitration supervised by the state.

(H.F.J./R.Po./M.A.M./M.A.Gl.)

GERMANIC LAW

Germanic law is a designation that covers the laws of the various peoples of Germanic stock from the time that the earliest "barbarian" tribes came into contact with the Romans until their tribal laws developed into national territorial laws—a development that occurred at different times with different peoples. Thus some of the characteristics of Scandinavian legal collections of the 12th century are similar to those in the Visigothic laws of the 6th century.

Knowledge of the early Germanic period is derived mainly from the observations of tribal life contained in Julius Caesar's *Gallic War* and Tacitus' *Germania.* The first written collections of Germanic law are the so-called Leges Barbarorum, which date from the 5th century until the 9th century. They are written in Latin and show Roman influence by their use of the technical terms of Roman law. The Anglo-Saxon laws and the laws of the North Germanic group, on the other hand, are in the vernacular and owe their written form largely to the advent of Christianity.

For all of the Germanic peoples, law (West German, *reht* and *êwa;* High German, *wizzôd;* North German, *lagh,* from which the English word *law* is derived) was basically not something laid down by a central authority, such as the king, but rather the custom of a particular nation (tribe). It was essentially unwritten, being derived from popular practices, and was not sharply distinguished from morality; it was personal in the sense that it applied only to those who belonged to the nation. Thus each man followed his own law, a notion appropriate to a nomadic people who originally did not live in a clearly defined territory. When, after the fall of the Roman Empire in the West, Germanic tribes took over former Roman provinces, they did not attempt to apply their laws to their Roman subjects, for whom Roman law remained applicable.

Visigothic laws

Thus the earliest Germanic code, that of Euric, king of the Visigoths in Spain and southwestern Gaul in the late 5th century, applied exclusively to Visigoths. The Lex Romana Visigothorum, or Breviary of Alaric, was issued in AD 506 for their Roman subjects. It was a compilation of "vulgar law"—Roman law adapted to fit the social and economic conditions of the late Roman Empire—and was later the main source of Roman law in the Frankish kingdom. Only in the 7th century was Visigothic law applied to Visigoths and Romans alike, the two peoples by then having substantially fused. The Lex Burgundiorum and the Lex Romana Burgundiorum of the same period had similar functions, while the Edictum Rothari (643) applied to Lombards only.

The Leges Barbarorum, then, were not legislation in the modern sense but rather the records of customs that were first collected and then declared as law. The prologue to the Salic Law (the law of the West, or Salic, Franks) recounted how four chosen men collected the original practices in particular cases, having first discussed them with the presidents of the local popular assemblies. The Leges Barbarorum did not seek to set out all of the main rules of law as modern codes do. They were not concerned with what everyone took for granted but concentrated on matters that, perhaps as a result of migration or conquest, had become doubtful and needed authoritative exposition. They dealt with specific situations rather than general rules and focused particularly on court procedure, monetary compensation for acts of violence, and succession on death.

The initiative for declaring law usually came from the king, but the resulting laws normally required approval by the popular assemblies. Because of this collaboration between king and people, a compilation was sometimes referred to as an "agreement," or *pactus.* The Visigothic laws were an exception; they always appear to have been formulated by the king and chief landowners without popular participation. Gradually, first the Lombard and then the Frankish kings overcame their people's aversion to central government and began to legislate unilaterally. The Lombards, who invaded Italy in 568, had no single code of custom, but their kings issued edicts from the mid-7th century onward. In the Frankish kingdom the Merovingian kings called their legislation *edicta* or *praecepta,* but the succeeding Carolingians characterized them as *capitularia; i.e.,* royal ordinances divided into articles (*capitula*). These included modifications of the *leges* of the Franks or other nations in the Frankish kingdom, administrative orders to officials, and independent legislation. Like the Roman emperors before them, Charlemagne and his successors claimed the power to make laws for all their subjects, irrespective of nation, and without the consent of any assembly. The validity of the law depended solely on the oral act of the king who promulgated it.

Tribal Germanic institutions. Germanic law recognized a distinction between free and unfree persons. Only the former had legal capacity, and they were subdivided into nobles and ordinary freemen. The nobles enjoyed a larger share in land distribution, were preferentially chosen for public office, and were protected by a larger monetary compensation if they were injured. Certain west Germanic tribes recognized an intermediate status of half-free persons, who could enter into legal transactions and marry but had no political rights.

Basically a Germanic tribe was a league of clans. Its main institutions of government were the king, his council, and the tribal assembly (*mallus, witan, mot, ding,* or *thing*). The king was military leader, chief priest, and president of the assembly, and he was assisted in the routine business of government by his council of elders and higher nobles. The assembly was composed of all free members of the tribe grouped into clans. It elected kings, declared war, outlawed freemen, and generally controlled the membership of the tribe by its supervision of the manumission of slaves, the emancipation of minors, and the adoption of strangers.

Role of the king

The dominant social institution was the "sib" (*sippe*), a term that meant both a clan—the extended family composed of all those related by blood, however remotely, and subject to a clan chief—and also a household or narrow family, whose members were under the *mund* (guardianship) of the family head. A boy remained in his father's *mund* until he was emancipated on attaining physical maturity; a girl remained until she married, when she passed into the *mund* of her husband. Marriage commonly took the form of the sale of the bride to her groom for a price, which developed into a fund held by the husband for the wife's benefit. A husband could divorce his wife at will but risked being penalized financially.

Laws of property

The main notion in the law of property was *gewere,* or the power exercised by the owner, which did not clearly distinguish between legal title and physical control. Various forms of limited ownership were recognized. Land was treated differently from movables; originally it had belonged to each family collectively. Family ownership gradually developed into the private ownership of the family head, but for a long time he could alienate land only with the consent of the nearest heirs. Land transfer required much formality, and among the west Germanic peoples a glove or spear was handed over as a symbol of *gewere.*

At the death of the family head, his property passed to his descendants in the nearest degree of proximity, with a preference for males. (The declaration in the Salic Law that daughters could not inherit land was used by 16th-century French lawyers as additional support for the long-standing practice of excluding women or their descendants from succeeding to the crown.) In the absence of descendants, several *leges* provided that property deriving from the father's side should return to that side and property from the mother's side to her side. The order of succession could not be altered by will.

When trade was still on a cash or barter basis, there was little need for formal contract law. A family could obligate itself to another either by pledging a thing as security (*wadium, gage*) or by surrendering a hostage (*gijzel, born*).

Later, a debt was guaranteed by a formal oath accompanied by the surrendering of a staff to the creditor (*effestucatio*). Contractual obligation was then constituted either by oath (enforced by an action for perjury) or by delivery of a thing (enforced by an action for theft).

Laws to control crime

Offenses against the community, such as treason, secret killing, and secret theft, were punished by outlawry, which was pronounced by the tribal assembly. The convicted person could then be killed by anyone. Offenses against individuals, including open killing and open robbery, became the subject of a blood feud if the criminal and victim belonged to different family groups. Peace could be bought by the payment of compensation, known as *wergild* in homicide cases and *bot* in others. Payment was voluntary at first; only later did it become obligatory. Even in the 7th century, Visigothic law still allowed retaliation in kind for all injuries except those to the head. The *leges* contained elaborate tariffs of compensation for different kinds of injury, the amount varying according to the social status of the victim. Private feuds were eventually restricted by the growth of royal authority in the Frankish period and the notion of the king's peace, the breach of which was punishable by the king's court.

When parties appeared before a court and stated their cases, the court decided on an acceptable method of proof, which could be by oath of the parties, supported by *compurgatores* (literally "oath-helpers"), the number required depending on the gravity of the case, by ordeal, or by battle. A successful claimant had to enforce judgment himself on the person or property of the defendant.

Rise of feudal and monarchial states. With the disintegration of the Frankish kingdom in the late 9th century, government became highly decentralized. Already the pattern of landholding, which determined the more important legal relationships, had begun to take on the characteristics of feudalism. Before the end of the Roman Empire much of the land had been concentrated in the hands of magnates, secular and ecclesiastical. But, unlike their predecessors under the Romans, the holders of secular land in the Germanic states became largely independent of the central government. By the 9th century, many lords had become strong enough to challenge the power of the Carolingian kings of the Frankish Empire and to make the inhabitants of their own areas their vassals. These vassals held their land from the lords as tenants of a so-called feud, or fee. Each feudal lord held a court for his tenants in which he applied the same law to all of the tenants, irrespective of their racial or national origin. Thus the old Germanic personal principle was abandoned in favour of the territorial principle, or the application of the custom of the region. This type of feudal law usually was based partly on Germanic law and partly on the Roman law of the Lex Romana Visigothorum, adapted in the interests of the feudal lords.

Influence of canon law

During the same period the Roman Catholic Church became the main unifying force in western Europe and began to claim jurisdiction over many matters that earlier had been considered secular rather than ecclesiastical. Church courts had existed since the Roman Empire, and their power in matters of faith was recognized by the secular authorities. The personal law of the church as an institution was always Roman, and indeed the law of the Ripuarian Franks on the Rhine expressly declared that "the church lives by Roman law." The canon law applied in the church courts was largely influenced by Roman law and contained very few Germanic elements. As the power of the church grew, the church courts applied this law to matters that had previously been dealt with by the secular courts, such as marriage, adultery, wills, and succession. In many countries these matters remained withdrawn from Germanic law and subject to church law even after the Reformation.

Merchants also found that the old Germanic customary law was inadequate to cope with the problems created by the rapid growth of commerce that had occurred by the 12th century. A special commercial law, based mainly on Roman law as developed by the Mediterranean seaborne traders, was developed to settle disputes between merchants, without regard to their nationality or place of residence.

These developments reduced the range of cases that were subject to the jurisdiction of the local county courts. In Germany some of the earlier codifications of customary law were forgotten, partly because the local judges were unable to understand the Latin in which they were written and partly because the rules that they contained were unsuited to the new social and economic conditions. The local courts applied an unwritten customary law based on the dominant tribal law of the area, and it was this that formed the basis of such codifications as the *Sachsenspiegel* ("Mirror of the Saxons") in the 13th century.

Regional variations

In France the legal development in the north differed from that in the south. The regional customs in the north were made up of Germanic and Roman law, the Carolingian capitularies, and canon law, but Germanic elements predominated. In the south, the so-called *pays de droit écrit* ("land of written law"), where Gallo-Romans had been far more numerous than Franks, the custom of each district was based mainly on the vulgar law of the Lex Romana Visigothorum. In Italy this law existed side by side with Lombard law. In the 7th and 8th centuries that law was subjected to a relatively sophisticated codification, whose form showed Roman influence.

In England the Norman conquerors continued the movement toward legal unity begun by the Anglo-Saxons by imposing on the country a centralized form of government more powerful than any on the Continent. In the 12th century Henry II made the king's court a permanent court of professional judges with jurisdiction over many matters that earlier had been dealt with by other courts. The common law developed by this court was largely Germanic law. (P.G.S./M.A.Gl.)

Civil-law systems

THE HISTORICAL RISE OF CIVIL LAW

In the 5th and 6th centuries western and central Europe were dominated by Germanic peoples, especially those who had overrun the Roman Empire. Among them were the Anglo-Saxons of England, the Franks of western Germany and northern France, the Burgundians, the Visigoths of southern France and Spain, and the Lombards of Italy. Although Roman law traditions lingered on for some time, the Germanic customs came to prevail in most regions. In the Middle Ages these customs underwent vigorous growth in an effort to satisfy the complex needs of a society that encompassed changing feudalism, chivalry, growing cities, Eastern colonization, increasing trade, and a constantly refined culture. Among the many strands that went into the weaving of the complex pattern of medieval law, the customs of the merchants and the canon law of the Roman Catholic Church were of special significance. It was through the canon law that the concepts and ideas of ancient Rome continued to make their presence felt, even when, as a whole, Roman law had been forgotten. In the late 11th century Roman law was rediscovered and made the subject matter of learned study and teaching by scholars in northern Italy, especially at Bologna. With the increasing demand for trained judges and administrators, first by the Italian city-republics, then by princes in other localities, students flocked to Bologna from all over Europe, until the study and teaching of law were gradually taken over by local universities. As a result of this process, Roman law penetrated into the administration of justice north of the Alps, especially in Germany and the Netherlands, where the Roman-law influence became particularly strong.

Revival of Roman law

In the Holy Roman Empire of the German Nation the reception of Roman law was facilitated because its emperors cherished the idea of being the direct successors of the Roman Caesars; Roman law, collected in the Corpus Juris Civilis by the emperor Justinian I between 527 and 565, could be regarded as still being in effect simply because it was the imperial law. Decisive for the reception, however, was the superiority of the specialized training of Roman-law jurists over the empiricist methods of lay judges and practitioners of the local laws. Equally decisive was the superiority of the Roman-canonical type of procedure, with its rational rules of evidence, over the local forms of procedure involving proof by ordeal, battle, and other irrational methods. Nowhere, however, did Roman law completely supplant the local laws, and, as far as the content of the law was concerned, various amalgams developed. Roman law strongly influenced the law of contracts and torts; canon law achieved supremacy in the field of marriage; and combinations of Germanic, feudal, and Roman traditions developed in matters of property and succession. The conceptual formulations in which the norms and principles of the law were expressed, as well as the procedural forms in which justice was administered, were also strongly Roman. The system that thus emerged was called the *jus commune.* In actual practice it varied from place to place, but it was nevertheless a unit that was held together by a common tradition and a common stock of learning. Although the law of the Corpus Juris Civilis (especially its main part, the Digest, the writings of the jurists) was, as such, in effect nowhere, it constituted the basis of study, training, and discourse everywhere. In spite of all local variety, the civil-law world experienced a sense of unity that corresponded to the strongly felt unity of European civilization.

National codifications

This unity was undermined by the religious split of the Reformation and Counter-Reformation and by the rise of nationalism that accompanied the unification and stabilization of the European nations and their struggle for hegemony. In the field of law the split found expression in the national codifications, through which the law was unified within each nation but simultaneously was set apart from that of all others. In Denmark codification occurred in 1683, in Norway in 1687, in Sweden-Finland in 1734, and in Prussia in 1794. Because of the personality of their promoter and the novel technique applied, great fame and influence were achieved by the Napoleonic codifications of the private and criminal law of France, especially their central piece, the Civil, or Napoleonic, Code (Code Civil or Code Napoléon) of 1804.

Codification continued after the Napoleonic era. In Belgium and Luxembourg, which had been incorporated into France under Napoleon, his codes were simply left in effect. The Netherlands, Italy, Spain, Portugal, and numerous countries of Latin America followed the French model not only by undertaking national codification but also by using the same techniques and arrangements. Naturally, their courts and legal scholars were, at least in the early 19th century, inclined to pay great attention to French legal learning.

In Germany national codification came considerably later than in France. Only a commercial code had been uniformly created by the independent German states shortly after the revolution of 1848. The unification of the criminal law took place almost simultaneously with the political unification of the country, which occurred in 1871. Codification of the organization of the courts and of civil and criminal procedure came in 1879. But the Civil Code (Bürgerliches Gesetzbuch für das deutsche Reich) was not completed until 1896, and it did not take effect until Jan. 1, 1900.

Throughout the 19th century the vigorous German science of law exercised much influence in Austria (which as early as 1811 had codified its law in a technique different from that of France), in Switzerland, in the Nordic countries, and, later, in most of eastern Europe. When Swiss law was codified in 1907–12, it became the model for the Turkish codification of 1926 and strongly influenced that of China, which is still in effect in Taiwan.

Due to the different dates of codification and the different style and attitude of legal learning, the civil-law family of laws is thus divided into the French, or Romanist, branch and the German, or Germanic, branch. Their main features are determined by those of their prototypes. The legal system of Japan essentially belongs to the German branch, but it presents important features of its own.

THE FRENCH SYSTEM

Changes induced by the French Revolution

In France the revolutionary period was one of extensive legislative activity, and long-desired changes were enthusiastically introduced. A new conception of law appeared in France: statute was deemed the basic source of law. Customs remained only if they could not be replaced by statutes. The Parlements, the major courts of the nation, were dismantled and replaced by a unified system of courts that were merely supposed to apply the law and never to lay down general rules.

The main ideas embodied in the revolutionary legislation were to be found in the motto of the Revolution (which is still that of France), "*Liberté, égalité, fraternité.*" The passionate desire for liberty and equality aroused by the 18th-century philosophers inspired the changes that took place.

The system that had come to be called feudal, although it had little to do with the feudalism of the High Middle Ages, was hated by the peasants and the bourgeoisie for its unbalanced distribution of privileges—especially those exempting the nobles and clergy from taxation. These privileges were abolished early in the Revolution. The revolutionaries detested organized groups of any kind, for it was thought that only one authority should exist over the citizens—that of the state. As a result the guilds, which demanded compulsory membership and regulated every profession, were suppressed, and freedom of commerce was established. The old-style universities were dissolved; in the same spirit the property of the Roman Catholic Church was secularized, and the priests and bishops were made state employees, a situation that most of them did not accept.

Family relations were deeply transformed according to the principles of liberty and equality. Marriage was organized merely as a civil act, divorce was permitted, paternal authority was limited, and parents' consent was not required for marriages of children over 21 years of age. A short experiment was made with "family courts" that were permitted to overrule paternal decisions, and the wife was declared equal to her husband. In matters of succession, equal parts were given to all children, and the testator's right to dispose of property by will was limited in order to prevent the reestablishment of inequalities by this device.

Throughout the revolutionary period, successive governments were committed to consolidating the legal changes in a set of codes. Drafts were made, but time and authority were lacking, and none was enacted until society was restabilized under Napoleon.

The concept of codification. From a practical point of view, the Civil Code achieved the unification of French civil law. This was not, however, the only concern of its drafters. They shared with most of their contemporaries and with most modern French lawyers the belief that the law should be written in clear language so that it would be accessible to every citizen. This view implied that the new code must be complete in its field, setting forth general rules and arranging them logically. Finally, it must not unnecessarily break with tradition.

The Civil Code was organized as a series of short articles because it was assumed, first, that legislators could not foresee all circumstances that might arise in life and, second, that only conciseness could make the code flexible enough to adapt old principles to new circumstances. The general rules contained in the code have since been applied to concrete circumstances without much difficulty. When an interpretation has been required, the courts have had the responsibility to give it, taking into consideration the "spirit" of the code in an effort to apply to each case the solution that would have been desired by the legislator.

Logic and experience in the codification

The drafters of the code strove toward inner consistency in their work, so that reliance on logic might ensure satisfactory application of it. They saw no contradiction between logic and experience. Since the 17th-century beginnings of the Age of Reason, abstract reasoning had characterized the French approach to law and to life in general. For this reason articles of the code were not regarded as narrow rulings. If no one article was found to apply exactly to a given situation, it was proper to consider several articles and to draw from them a more general rule that could either itself be applied to the case or be combined with others to reach a solution.

Although the code was a work of logic, it relied mainly on experience. Its drafters were exceptionally well qualified in this respect: they had lived the first half of their lives under the laws of ancient France and had also known the Revolution. Their purpose was not so much to create new laws as to restate existing laws, subject to choice when revolutionary enactments varied from previous ones and when previous laws differed from one another. They were ready to adopt any rules that seemed best suited to the French people on the basis of experience; they recognized that laws could not be inflexible "but must be adapted to the character, the habits, and the situation of the people for whom they are drafted."

Later changes and adaptations. No important changes were made in the Civil Code from 1804 to 1880, except the repeal of divorce in 1816, when a Catholic monarchy was restored. The political and legislative power was held by the bourgeoisie, and they were entirely satisfied with the basic principles of the code, which favoured individualism and free will. In fact, from 1804 until the enactment of the constitution of the Third Republic in 1875, the Civil Code remained the law of France despite several changes in political regimes. Jurisprudence was centred upon it; in both teaching and writing, scholars discussed it article by article. The courts fulfilled the role that the drafters had stressed for them; imbued with the spirit of the code, they applied its general rules to particular cases.

Changes during the Third Republic

The social atmosphere changed during the Third Republic when universal suffrage gave the labouring class an influence on legislation. Faith in liberalism was shaken, and the idea grew that the state should intervene to protect the weak. Statutes increased in number. This movement was accentuated by World Wars I and II, when a mass of emergency regulations had to be passed, and the power of the state to encroach on private interests for the sake of the community was increased.

Subsequent amendments to the code revealed two trends: first, greater individualism in family law; second, qualification of individual rights for the sake of social interests—what has been called "socialization" of the law.

Adaptation of the law to new social needs was not made by statute alone: the courts, to a certain extent, adjusted the law to modern circumstances. They did this, however, while maintaining a consciousness of their subordinate position. They recognized that, as a general rule, basic changes were the province of the legislature and not of the judge, though this did not prevent them from gradually adapting the law to the modern conditions of life.

Legal learning also had a role. A number of important statutes were drafted by commissions that included judges, professors, and lawyers; and authors often suggested to the courts new developments in the application of rules of law. Although most of the statutes passed during the 19th and 20th centuries were left outside the code, they continued to be published with the new editions of the code.

By the middle of the 20th century, it had become apparent that the code should be revised. This task was entrusted to a commission, which produced several important drafts. The effort to replace the old code with a completely new one was halted when Charles de Gaulle came to power in 1958. Revision has since occurred only on a sporadic, piecemeal basis, except for the sections concerning family law, which have been thoroughly reformulated. The continuing development of multinational institutions such as the European Communities may require readjustments on a broad scale. This seems particularly likely in the field of commercial law, where company law has become the first target for harmonization, and perhaps even in contract law.

The main categories of French private law. The French Civil Code uses many of the categories that were developed in ancient Rome, but its law is that of its own time.

Marriage viewed as basic to civilized society

Marriage and family. The drafters of the French Civil Code regarded marriage as the basic institution of a civilized society. Taking into account the variety of religious attitudes in France, they decided that only marriage ceremonies celebrated before secular officials should be legally valid. This did not deprive ministers of the various faiths of the right to celebrate religious marriage ceremonies, but these were devoid of any legal effect and had to take place after the secular ceremony in order to avoid any risk of confusion. Parental control over children's marriages was partially restored; consent was required for sons under 25 and daughters under 21. After 1900 the formalities of marriage were lessened and parental control over it curtailed. Twentieth-century statutes gradually reestablished the revolutionary rule that the consent of the parents was not necessary when the parties were over 21. In 1974 the age of majority for this and other purposes was reduced to 18.

In France under the ancien régime the family had been centred upon the husband, whose strong authority and powers were inherited from the Roman paterfamilias (head of family). Although the Revolution proclaimed women to be equal in rights with men, it did little to implement this view in law. The drafters of the code saw no reason to modify the traditional situation, and Napoleon himself favoured subordination of the wife to the husband. The code expressly stated that she owed him obedience. With very few exceptions, she had no legal capacity to act. Without the written consent of her husband, the wife could not sell, give, mortgage, buy, or even receive property through donation or succession. Statutes in the 20th century, however, severely diminished the authority of the husband over his wife and endowed her with full legal capacity. In 1970 the old language stating that "the husband is the head of the family" was abandoned in favour of a new principle of joint family decision-making power, which did not, however, extend to the management of community property.

Rights of husband and wife

In recent years matrimonial property regimes have been revised in numerous countries, the tendency being toward a partnership in property acquired after the marriage, with each party retaining control over the property he or she had before the marriage. Although the Napoleonic Code provided for a statutory regime (if no particular marriage contract had been made), under which all chattels and earnings of the spouses would be community property to be shared equally between them or their heirs at the dissolution of the marriage, the husband was vested with all active powers, even over his wife's property. In 1965 movables owned by either spouse before marriage were excluded from the community fund, which now, in the absence of an agreement to the contrary, consists only of the fruits of the spouses' work or frugality during marriage. With the acquisition of legal capacity in the early part of the century, a French wife was free to manage and dispose of her own earnings and property, but it was not until 1985 that the long predominance of the husband in the management of the couple's common property was replaced by a system of equal comanagement.

Divorce. Divorce was first introduced into France after the Revolution. It was made very easy and was even allowed by mutual agreement.

The drafters of the code decided that since many persons were not prevented by religious conviction from seeking divorce, it was not for the legislator to prevent unhappy spouses from terminating their marriages and from entering new legal unions. Divorce, therefore, was allowed, but only within strict limits, so that "the most sacred of contracts should not become the toy of caprice." The only grounds for divorce were adultery, sentences for the most serious crimes, excesses such as gambling habits and expenditures, cruel treatment, or serious insult. Mutual agreement was added under the personal pressure of Napoleon, already intent on divorcing his first wife, by whom he had no child. But the procedure of divorce by mutual agreement was extremely long, complicated, and costly, and no second marriage could take place within six years thereafter.

Divorce was repealed in 1816 after Napoleon's fall and the restoration of the monarchy, and it was not reintroduced until 1884. From 1884 to 1975, divorce was permitted only on the grounds of adultery, conviction of a serious crime, and cruelty. Divorce by mutual agreement was not reinstated until 1975, when a comprehensive reform of the divorce law permitted a marriage to be terminated by consent or by petition of one spouse unopposed by the other, or when the marriage had broken down after six years of separation or after six years of mental illness of one spouse, in addition to the traditional grounds of fault.

Succession and gifts. The Napoleonic Code adopted many of the ideas of the Revolution concerning succession. But its formulators tempered them with exceptions and combined them with ideas from the ancien régime.

The revolutionary law on intestate succession (succession without a valid will) relied upon two basic principles: (1) that no distinctions be made within the estate of the deceased, land and chattels being treated in the same way and no account being taken of the origin of landed property; and (2) that equal parts be given to all heirs of the same degree of kindred, the advantages accruing through some customs to the firstborn or to male children being abolished. Using these two principles, the code provided that an estate should devolve first of all upon the children and other descendants. If heirs of one degree died before others of the same degree and left children, representation (the principle that the children of a deceased heir inherit his share) applied. In other cases distribution was made per capita, with equal shares going to those heirs of equal degree. Illegitimate children could inherit from their parents, but they received less than legitimate children and could not cut out either the deceased's own parents or his brothers and sisters. Through reforms in the 1970s, the rights of illegitimate children to succeed to their parents have largely been assimilated to those of legitimate offspring.

Rights of heirs and surviving spouse

According to the code, the spouse could succeed only if there were no persons who were related to the deceased up to a degree specified by law. A surviving wife was, thus, in a poor position if no gift or legacy had been made to her, though under the statutory matrimonial regime she received half of the community property into which all chattels of both spouses fell. The rights of the surviving spouse have been increased at various times during the 20th century. Today the surviving spouse is entitled to at least the usufruct (similar to a life interest) of one-quarter of the property left by the deceased. The survivor inherits half of the estate if there are no children and if there are surviving ascendants on only one side of the deceased's family. If the decedent leaves no blood relatives within a certain degree of kindred, the surviving spouse receives the entire estate.

Wills may be formal or informal. Unwitnessed wills are valid, provided that they are written throughout, and dated and signed, by the testator's own hand. Wills are effective upon the death of the testator and do not need to be probated. Freedom to dispose of property by will or by gift is limited in order to protect children and other descendants as well as parents and grandparents, who have to be allowed a certain proportion.

Property. The intricate system of obligations and rights inherited by the ancien régime from feudalism was rejected by the Revolution, which restored a system patterned on that of Roman law.

Concept of movables and immovables

The only classification of goods is the basic one of immovables (which are defined as having a fixed place in space) and movables (which include all goods that are not immovables). In contrast to the "feudalist" complexities in common law, the normal relationship between persons and things is ownership, which is defined as a complete, absolute, free, and simple right. But, as in the law of other modern nations, the use of property is subject to many kinds of restrictions imposed in the public interest. Usufructs, or servitudes, are possible, but rights in an estate never require the person in whom they are vested to do anything. The code states that a servitude "is a charge laid on an estate for the use and utility of another estate belonging to another owner," and it emphasizes that "servitudes do not establish any pre-eminence of one estate over another." Title in land may be acquired within 10 or 20 years if the possessor believed, in good faith, that he was the real owner. Furthermore, the bona fide purchaser of movable property immediately becomes its owner, and nobody can prove a better title against him unless the property has been lost or stolen.

The section on mortgages in the Civil Code was weak. An excellent statute of the revolutionary period was developed in 1798 to set up a system of registration for all transfers of land titles and real estate mortgages. It enabled a buyer of land to ascertain whether he was buying from a regular owner and whether the land was mortgaged; if it was, the buyer could clear his title by offering the price to the mortgagee.

The drafters of the code maintained this system of compulsory registration, but only for gifts and for contractual mortgages. Sales of real estate and a number of legal mortgages were not subject to registration. This gap left prospective creditors or buyers with insufficient information. It was only after reforms were made in 1855, 1935, 1955, and 1967 that there was a comprehensive, but still not fully reliable, system of publicity for mortgages and conveyances of immovable property.

Contracts and torts. The French Revolution brought no changes into the law in this relatively nonpolitical field. The drafters of the code merely restated the law that had developed during the course of centuries and that authors already had analyzed.

Informality and freedom of contract

The basic principles of contract law are informality and freedom; the latter is limited, however, when demanded by public policy. The code states that "agreements legally entered into have the effect of laws on those who make them." The entire matter of torts is dealt with in only five short articles. The general basis for liability is the following: "Any act of a person that causes injury to another obligates the person through whose fault the injury occurred to give redress." The subsequent articles in the code regulate liability for damages caused by things, animals, children, and employees. It was left to the courts to work out a complete system of tort law based on these few articles.

THE GERMAN SYSTEM

Roman law, as embodied in the Corpus Juris Civilis, was "received" in Germany from the 15th century onward, and with this reception came a legal profession and a system of law developed by professionals (*Juristenrecht*). Roman law provided the theoretical basis for legal progress that culminated in the work of the scholars of the 19th century. Under this tradition, the legal process has been viewed in Germany as the application of more or less generally formulated rules to individual cases. German courts traditionally have not been as dominant in developing the law as have their counterparts in the common-law countries. Roman law provided tools to strengthen sovereignty, as well as the correlative ideas that the legislative function is a state monopoly and that the responsibility for the development of law rests with a legally trained, state-controlled bureaucracy rather than—as in 18th- and 19th-century England—with a combination of gentry and leaders of the bar. German judges traditionally have been university-trained experts under the authority of the state and the anonymity of the court. In the post-World War II period, however, West German judges have steadily assumed a more active role, especially in constitutional law.

Contrasts between the German and French codes

The German Civil Code. Because the German Civil Code of 1896 came almost 100 years later than the code of France, its draftsmen profited from the intensive efforts of German scholars who had systematized, clarified, and modernized the law during the 19th century. As a result, the German code is markedly different from its French predecessor: its arrangement is more orderly, its language more precise, and its use more exacting.

The appeal of the German code is from lawyers to lawyers; the matter-of-fact, neutral tone contrasts with the livelier mood in which the French Civil Code was written. It does not try to teach men in a broad sense, but it emphasizes ethical imperatives. Good faith and fair dealing are to be observed in all affairs. Breaches of good morals, abuses of rights, and underhanded legal transactions are deprived of legal effect. The code was meant to fit the society of the turn of the century, but through the use of general clauses that leave the elaboration of specific norms to the judges, it has demonstrated an adaptability to new economic, cultural, and sociopolitical postulates.

Modern law in West Germany assumes that the proper form of society is that of a social democracy, which not only confers individual rights but also holds the state responsible for social welfare and individuals responsible for behaving in a socially responsible way. The former concern of German law with abstract concepts has given place to a more pragmatic approach that attempts to apply the scale of values of a pluralistic society, containing strong elements of a welfare state, while emphasizing civil liberties.

The main categories of German private law. The German Civil Code begins with the proposition that at birth every person acquires the capacity to exercise rights and to fulfill duties. A minor's interests are protected by a representative who acts in his name, and although certain legal transactions may be entered into at age seven, full legal capacity is not acquired until age 18 (formerly 21). Every person possesses the right, protected by an action in court, to freedom from personal injury and from attacks on individual dignity.

Marriage and family. Since 1875 marriage has required civil celebration by a registrar, who cannot be a priest. Celebration in church may follow the civil ceremony. Marriage can be declared null and void on application by one of the spouses or by the public prosecutor on various grounds, such as lack of form or affinity, but the consequences of such nullity approximate those of divorce: the children are not necessarily illegitimate. Since 1976, the sole ground for divorce has been the breakdown of the marriage, which is presumed if the spouses have lived apart for a year and are in agreement on the divorce, or if the spouses have lived apart for three years.

Rights of husband and wife

The provisions of the German Civil Code concerning the rights of women in marriage were less restrictive than those of the French Civil Code. After World War II nearly all rules contravening the principle of equality of men and women were repealed. The ordinary statutory marital-property regime, with the husband administering and using the wife's estate, was replaced in 1957 by a system of separate management and equal sharing in the value of acquisitions made during the marriage. Upon the death of one spouse the surviving spouse is entitled to a generous share in the estate. Care for the person and property of the children belongs to both spouses.

Succession. In contrast to Anglo-American law, the assets of the decedent pass directly to the heirs, who are determined by the rules of intestacy or by testamentary disposition. As a general rule, the estate does not pass through a stage of administration by an administrator or executor. The heirs are liable for the debts of the decedent with their own property but by taking appropriate steps may limit their liability to the assets of the estate. A testator may appoint an executor to perform certain functions in the settlement. A will may be unwitnessed, but then it must be entirely in the testator's handwriting. Public wills are either made orally before a public official, who records them, or set down in a document that the testator hands to the official with a declaration that it is his last will. Descendants and other close relatives, including the surviving spouse, cannot be deprived of more than one-half of their intestate shares.

Property. Property is declared to entail obligations of the owner to the community. This is particularly important in terms of farmland, which can be pooled and redistributed to make better use of machinery and to increase production. Every creation, transfer, encumbrance, or cancellation of a right in immovable property requires, in addition to the agreement of the parties, registration with the district court. A person who, in good faith, acquires an interest in land from the person registered is protected. In order to obtain title to a chattel from a person who does not own it, the transferor must have had possession, the transferee must have been in good faith, and the owner must not have lost possession involuntarily. But neither in the case of land nor in that of chattels is it required that the transfer to the transferee be for value. Even if the transferee acquires a title, he may be required to surrender the asset or to pay its value if the acquisition appears to be a legally unjustified enrichment.

Contract and delict. Parties are free to regulate their relations by contract, within limits set by express statutory prohibitions and by good morals. Strict limits are set to eliminate fraudulent practices by one of the contracting parties. In the case of a valid contract, the parties must observe the requirements of good faith, with ordinary usage taken into consideration. The determination of "ordinary usage" is left to the courts. This has been particularly advantageous given the rapidly changing conditions of the 20th century.

Compensations for breach of contract and injury

Unless the promisor can prove that a breach of contract has been caused in a way entirely outside his sphere of risk, he is liable for damages. But if the promisee chooses to do so, he may have the promisor ordered to complete the contract as long as it is not shown that this is impossible. The principle that "anyone who through an act performed by another or in any other way acquires something at the expense of that other without legal justification is bound to return it to him" is stated in broad terms, but it is cautiously applied by the courts.

In terms of delict, the German Civil Code provides that any person who intentionally or negligently injures unlawfully the life, body, health, property, or any other absolute right of another person is bound to compensate him for any damage arising therefrom. Damages also are due for harm caused by the violation of a statute meant to protect others and for harm caused intentionally and immorally. If a public officer violates his statutory duty, court remedies are readily available against the government.

OTHER SIGNIFICANT CODIFICATIONS

Swiss and Italian codes

Swiss law. Shortly after German law was codified, Switzerland followed suit. Its Civil Code of 1907, together with a separate Code of Obligations, went into effect in 1912. These new federal codes superseded the earlier codes of the separate cantons (which had generally been patterned after the Austrian or the French model). The Swiss draftsmen took advantage of earlier experiences with codification technique—drawing especially upon both the Code ,Napoléon and the German code. The Swiss Civil Code, which exists in German, French, and Italian versions of equal authority, is a masterly attempt at summarizing and systematizing civil law and has influenced codification in countries as diverse as Brazil and Turkey.

Italian law. The French code was introduced into parts of Italy during the Napoleonic conquests. Even after the collapse of Napoleon's empire, when French law was abrogated, the Code Napoléon still served as the model for the new codes of several Italian states. The new Civil Code for the Kingdom of Italy was enacted in 1865 while the peninsula was being united politically. Its structure and content were reproductions of the French Civil Code. Unlike France and Germany, which have occasionally tried to draft new codes but still have not replaced their original ones, Italy succeeded in introducing a reformed code in 1942, during the Fascist era. This code remains in force, with its only amendments due mainly to changes in political regimes. In comparison with the Civil Code of 1865, Italy's code of 1942 appears inspired by less individualist views—for example, in property law and in labour legislation the social aspects of the law are stressed.

The Japanese code

Japanese law. After the Meiji Restoration of 1868, which abolished feudal privileges and restored titular power to the emperor, the leaders of the new government sought to construct an economic, political, and legal structure capable of commanding respect internationally. The introduction of Western law was one element of a wholesale importation of things Western. In legal matters, the Japanese took for models the systems of continental Europe, especially the German. The drafters of the Japanese Civil Code of 1898 surveyed many legal systems, including the French, Swiss, and common-law, and they took something from each. Their final product was, however, best characterized as following the first draft of the German Civil Code. In subsequent developments, the Japanese legal system remained true to these sources. In 1947 revisions of code provisions dealing with family law and succession, which had reflected traditional Japanese attitudes, completed the transition of Japanese civil law to the continental European family of laws.

In several respects, however, Japanese law is closer to that of the United States than to European models, especially in matters of public and constitutional law. This is largely a result of the post-World War II occupation and of subsequent contacts with U.S. legal thinking and education. From the perspective of the rules and institutions of private law, the Japanese legal system remains closer to the civil law of Europe than to the common law of the United States. In many ways, nevertheless, the Japanese legal order differs markedly from all Western legal orders.

The fact that Japanese law is not the product of organic evolution suggests that the role of law in modern Japanese society differs markedly from its role in Western societies. In Japan, law plays a far less pervasive role in the resolution of disputes and in the creation and adjustment of rules regulating conduct. The size of the Japanese bar is small, and extralegal methods of resolving disputes continue in large measure.

For many purposes the Japanese family transcends husband, wife, and dependent children. The notion that a business is analogous to a family unit also persists and colours all labour relations, especially in small and middle-sized firms. In the relatively homogeneous Japanese society, social status carries heavy obligations, and community pressure is extremely powerful.

Thus, although Japan early adopted a version of the German Civil Code, it did not adopt the Germans' strong consciousness of legal rights. In many areas of Japanese life it is still difficult to predict whether a dispute will be settled under legal standards, and it is often impossible to know whether a person will enforce those rights that are legally available to him. The concepts pervasive in Western law—that the legal consequences of a particular conduct should be predictable before the conduct has occurred, that in any dispute the courts should give full effect to claims (a plaintiff receiving all or nothing), and that individual disputes should be resolved without considering the parties' social and economic background—have not penetrated deeply into Japanese law. In contrast, facilities for conciliation are used to promote adjustment in terms of nonlegal considerations: local police stations provide conciliation rooms; elders act as go-betweens. Compromise based on legally irrelevant considerations is encouraged, and disputes are often resolved by techniques that fall outside formal law.

Alternatives to legal action

Despite deliberate governmental efforts to limit the number of legal professionals, the continuing westernization of Japan may force Japanese law to play a role fully comparable to the role of law in the West; the sociological supports essential to the continued vitality of the traditional Japanese conception of law are clearly being undercut by Japan's shift to a highly urban, mechanized society.

(M.Rh./M.A.Gl.)

Common-law systems

THE HISTORICAL RISE OF COMMON LAW

English common law, or the body of customary law embodied in reports of decided cases, originated in the early Middle Ages in the decisions of local courts, which applied what Blackstone called "the custom of the realm from time immemorial" and practical reason to everyday disputes with the aid of but few formal enactments. Until the late 19th century, English common law continued to be developed primarily by judges rather than legislators.

The common law of England is in fact largely a Norman creation. The Anglo-Saxons, especially after the accession of Alfred the Great (871), developed a body of rules resembling those being used by the Teutonic peoples of northern Europe. Local customs governed most matters, while the church played a large part in government. The concept of crimes originated in this era, but they were treated as wrongs for which compensation was made to the victim.

The Norman Conquest of 1066 brought a practical end to the Saxon laws, except for some local customs. All of the land was allocated to Norman feudal vassals of the king. Serious wrongs were regarded mainly as public crimes rather than as personal matters, and the perpetrators were punished by death and forfeitures of property. Government was centralized, a bureaucracy built up, and written records maintained. Royal officials roamed the country, inquiring into the administration of justice. Church and state were separate and had their own law and court systems. This led to centuries of rivalry over jurisdiction, especially since appeals from church courts, before the Reformation, could be taken to Rome. Some elements of Saxon practice lingered, including trial by ordeal (by burning the hand, for example), which was retained until 1215. Outlawry, a Saxon procedure whereby a fugitive was placed outside the protection of the law, was retained for centuries to deal with people who fled from justice. Gradually, however, new procedures took the place of these crude devices.

Remnants of Saxon law

The Normans spoke French and had developed a customary law in Normandy. They had no professional lawyers or judges; instead, they used "clerks," or literate clergymen, to act as administrators. Some of the clergy were familiar with Roman law and the canon law of the Christian Church. Canon law was adopted by the English church, but the Normans resisted any attempt to introduce Roman law, which was applied only to certain claims under wills in the church courts, to marine disputes in the admiralty courts from the 14th century, and to military law. Norman custom was not simply transplanted to England, and a new body of rules, based on local conditions, grew up.

The feudal land law. At the critical formative period of common law, the English economy depended largely on agriculture. Wages and profits were important only in commercial centres such as London, Norwich, and Bristol. Political power was rural and based on landownership. Landowners voted at elections as Parliament evolved, and they acted as sheriffs and magistrates and sat on juries.

Land was held under a chain of feudal relations. Under the king came the aristocratic "tenants in chief," then strata of "mesne," or intermediate tenants, and finally the tenant "in demesne," who actually managed the property. Each piece of land was held under a particular condition of tenure; that is, in return for a certain service or payment. An armed knight, for example, might have to be provided to serve in the king's armies for a certain period each year. Nonmilitary service, such as making deliveries of grain, was often substituted for the uncertain obligations of knight service. Periodic services tended to be commuted into fixed annual payments, which ceased, under the impact of inflation, to have much value. The "incidents," or contingency rights, however, such as the right of the feudal lord to take the land if the tenant died without heirs and his rights regarding wardship and marriage of the tenant's infant heirs (that is, his rights to compensation for exercising wardship or granting permission to marry) were assessed at current land values and remained important.

Conditions of tenure

Succession to tenancies was regulated by a system of different "estates," or rights in land, which determined the duration of the tenant's interest. Land held in "fee simple," for example, meant that any heirs could inherit (that is, succeed to the tenancy), whereas land held in "fee tail" could pass only to direct descendants. Life estates (tenancies that lasted only for one person's lifetime) could also be created. Title to land was transferred by a formal ritual rather than by deed because the population was largely illiterate. Few elaborate rules regulating the terms by which land was held could be agreed upon in such circumstances, so statutes were passed to regulate matters of detail. The life tenant, for example, was forbidden in the 13th century to use the property in such a way as to damage it or to cause it to deteriorate unless the grant specifically allowed it, and the tenant "in tail" was forbidden to ignore the system of descent laid down for his property. The common-law judges devoted themselves to working out the proper rules to apply to all of these estates and tenures.

Primogeniture—*i.e.*, the right of succession of the eldest son—became characteristic of the common law. It was designed only for knight-service tenures but was inappropriately extended to all land. This contrasted with the widespread practice on the Continent, whereby all children inherited equal shares.

Development of a centralized judiciary. The unity and consistency of the common law were promoted by the early dominant position acquired by the royal courts. A single royal court, the King's Court (Curia Regis), was set up for most of the country at Westminster, near London. Whereas the earlier Saxon *witan*, or king's council, dealt only with great affairs of state, the new Norman court assumed wide judicial powers. Its judges (clergy and statesmen) "declared" the common law.

By straining the interpretation of a statute, royal judges greatly reduced the jurisdictions of local courts. With their increased powers, royal judges went out to provincial towns "on circuit" and took the law of Westminster everywhere with them, both in civil and in criminal cases. Local customs received lip service, but the royal courts controlled them and often rejected them as unreasonable or unproved: common law was presumed to apply everywhere until a local custom could be proved. This situation contrasted strikingly with that in France, where a monarch ruled a number of duchies and counties, each with its own customary law, or of the situation in lands such as Germany and Italy, which were divided into independent kingdoms and principalities with their own laws.

Dissemination of common law by itinerant judges

This early centralization also removed the need for England to import a single advanced foreign system of law, a need that led to the reception of Roman law in Europe after the decline of feudalism. The expression common law, devised to distinguish the general law from local or group customs and privileges, came to suggest to citizens a universal law, founded on reason and superior in type.

In the 13th century the common central court split into three courts—Exchequer, Common Pleas, and King's Bench. Although the same law was applied in each, they vied in offering better remedies to litigants in order to increase their fees.

The court machinery for civil cases was built around the writ system. Each writ was a written order in the king's name issued at the instance of the complainant and ordering the defendant to appear in the King's Court or ordering some inferior court to see justice done. It was based on a form of action (*i.e.,* on a particular type of complaint, such as trespass), and the right writ had to be selected to suit that form. Royal writs had to be used for all actions concerning title to land.

Bracton and the influence of Roman law. Under Henry III, who reigned from 1216 to 1272, an assize judge (*i.e.,* an itinerant judge of the periodical local assize courts), Henry de Bracton (originally Bratton), prepared an ambitious treatise known as "Bracton." It was modeled on the order of the 6th-century Roman legal classic, the Institutes of Justinian, and shows some knowledge of Roman law. Its English character derived from the space it devoted to actions and procedure, to the reliance on judicial decisions as declaring the law, and to the statements limiting absolute royal power. Bracton abstracted several thousand cases from court records (plea rolls) as the raw material for his book. The plea rolls formed an almost unbroken series from 1189 and included the writ, pleadings, verdict, and judgment of each civil action.

Early statute law. Edward I has been called the English Justinian because his enactments had such an important influence on the law of the Middle Ages. Edward's civil legislation, which amended the unwritten common law, remained for centuries as the basic statute law. It was supplemented by masses of specialized statutes that were passed to meet temporary problems.

Major statutory enactments of Edward I

Four of Edward's statutes deserve particular mention. The first Statute of Westminster (1275) made jury trial compulsory in criminal cases and introduced many changes in the land law. The Statute of Gloucester (1278) limited the jurisdiction of local courts and extended the scope of actions for damages. The second Statute of Westminster (1285), a very long enactment, confirmed the estate tail in land, which had often been linked with the maintenance of titles of honour; made land an asset for purposes of paying judgment debts (debts judged to exist by a court); liberalized appeals to high circuit courts; improved the law of administration of assets on death; and created a new form of action, action on the case, that gave broad approval to the creation of new remedies for new types of contract and tort cases. The Statute of 1290 (Quia Emptores) barred the granting of new feudal rights, except by the crown, and made all land held in fee simple freely transferable by denying interference by relatives or feudal lords.

In modern times the statutes issued prior to 1285 are sometimes treated as common law rather than statute law. This is because these laws tended to restate existing law or give it a more detailed expression. They explained what the law was, but they did not make an entirely new law; some authorities, in fact, doubted whether governments had the right to change ancient customs at all. In addition, judges did not always adhere closely to the words of the statute but tried to interpret it as part of the general law on the subject. Prior to the rise of the House of Commons in the 13th century, it also was difficult to distinguish acts of Parliament from the less binding decisions or resolutions of the royal council, the executive authority. Some statutes were passed but never were put into force, while others seem to have been quietly ignored.

The second Statute of Westminster, however, clearly made new law and allowed time for citizens to study its provisions before it came into force. Even so, this statute was freely interpreted by the courts, who read into it things that were not in the text.

Growth of Chancery and equity. Since legal rules cannot be formulated to deal adequately with every possible contingency, their mechanical application can sometimes result in injustice. In order to remedy such injustices, the law of equity (or, earlier, of "conscience") was developed. The principle of equity was as old as the strict common law, but it was hardly needed until the 14th century, since the law was still relatively fluid and informal. As the law became firmly established, however, its strict rules of proof began to cause hardship. Visible factors of proof, such as the open possession of land and the use of wax seals on documents, were stressed, and secret trusts and informal contracts were not recognized.

Power to grant relief in situations involving potential injustices lay with the king and was first exercised by the entire royal council. Within the council, the lord chancellor, a leading bishop, led the meetings and, by 1474, dealt personally with petitions for relief. Eventually the chancellor's jurisdiction developed into the Court of Chancery, whose function was to administer equity. Much of the work concerned procedural delays and irregularities in local courts, but gradually the power to modify the operation of the rules of common law was asserted.

Early grounds for equitable relief

The chancellor decided each case on its merits and had the right to grant or refuse relief without giving reasons. Common grounds for relief, however, came to be recognized. They included fraud, breach of confidence, attempts to obtain payment twice, and unjust retention of property.

Proceedings began with bills being presented by the plaintiff in the vernacular language, not Latin; the defendant was then summoned by a writ of subpoena to appear for personal questioning by the chancellor or one of his subordinates. Refusal to appear or to satisfy a decree was punished by imprisonment. Because the defendant could file an answer, a system of written pleadings developed.

Inns of Court and the Year Books. During Edward I's reign (1272–1307) the office of judge was transformed from a clerical position into a full-time career. Admission to the bar (*i.e.,* the right to practice as a barrister before a court) was made conditional on the legal knowledge of the applicant so that law began to emerge as a profession which required permanent institutions and some kind of organized education.

As the legal profession grew, the more experienced barristers were admitted to the dignity of serjeant-at-law and later banded together with the judges, who were appointed from their ranks, at Serjeants' Inns, in London. There, burning legal problems were informally discussed, and guidance was given to all concerning the decisions of actual or likely cases. The four Inns of Court (Gray's Inn, Lincoln's Inn, Inner Temple, and Middle Temple) evolved from the residential halls of junior barristers to become the bodies officially recognized as having the right to admit persons to the bar. Education consisted of attending court, participating in simulated legal disputes (moots), and attending lectures (readings) given by senior lawyers.

Bracton's work was adapted for purposes of study for a time, but it soon became outdated. Bar students therefore had to make notes in court of actual legal arguments in order to keep abreast of current law practices. These notes varied widely in quality, depending on the ability of the notetaker and the regularity of his attendance, and from about 1290 they seem to have been copied and circulated. In the 16th century they began to be printed and arranged by regnal year, coming to be referred to as the Year Books.

The Year Book reports were written in highly abbreviated law French. They did not always distinguish between the judges and barristers and often simply referred to them by name. The actual judgment also was often omitted. Previous decisions were not generally binding, but great attention was paid to them, and it appears that the judges and barristers referred to earlier Year Books in preparing their cases. Thus, case law became the typical form of English common law.

Effects of feudal conflict on the law

The dynastic Wars of the Roses in the latter part of the 15th century led to a practical breakdown of the legal order. Powerful hereditary aristocrats in the country, backed by private armies, and dominant commercial families in the towns were beyond the effective reach of the royal writ. When legal proceedings were possible, they were often manipulated or frustrated by the crown's "overmighty subjects," who intimidated and corrupted justices, sheriffs, juries, and witnesses.

Thus the years preceding the Tudor period were a time of insecurity and stagnation, a "Gothic age" in which lawyers tried to consolidate the law but made no new advances. Parliamentary authority also was weakened, and the royal council was called on more and more to rule the country and try to maintain order.

The rise of the prerogative courts. The accession of Henry VII in 1485 was followed by the creation of a number of courts that stood outside the common-law system that Henry II and his successors had instituted. These newer courts were described as prerogative courts because they were identified with the royal executive power, although some of them had a statutory origin. Thus, the Council of the North at York was set up by statute in 1537, and the Council of Wales and the Marches at Ludlow was confirmed by statute in 1543, though both had been preceded by older prerogative courts in those "frontier" regions. The Court of Requests (see below) was given regular status by an administrative action in 1493. The Court of Star Chamber, once thought to have been given its authority by a statute of 1487, is now believed to have evolved from the royal council, which began acting as a judicial committee in the early 16th century. All these courts rested on the comparative authority and efficiency of the council in times when regular courts were unable to operate properly.

The Court of Requests

In the Court of Requests, which had counterparts in France, the costs of procedure were lower than in common-law proceedings; it was designed to accommodate small civil claims by the poor. The judges of the court were styled masters of requests, and they had many other duties, which often caused delays. The court flourished in the 17th century until the Civil War (1642–51), when the procedure by which it operated was abolished. Its example of offering a simple, cheap procedure was imitated by several statutory courts that were set up in towns in later times, also known as courts of requests.

Whereas the common-law courts punished "hanging crimes," such as murder and robbery, the Star Chamber dealt with more sophisticated offenses, such as forgery, perjury, and conspiracy. Fines and sentences of imprisonment were the usual punishment. Common-law judges, lay peers, and bishops sat on this court, which also exercised civil jurisdiction. It lost its original popularity when the early Stuart kings used it to stifle political opposition, and its name eventually became synonymous with repression. It was abolished in 1641, and most of its jurisdiction was absorbed by the common-law courts in 1660.

The rather specialized High Court of Admiralty developed under royal prerogative in the 14th century; a statute of 1391 prohibited it from meddling in cases not arising at sea. In Tudor and early Stuart times, however, it exercised a wide commercial jurisdiction. After the Civil War it was confined exclusively to trying maritime disputes.

Further Roman-law influences. As described above, the common law had begun to break down in the 15th century. Abroad, law was in a state of flux. The customs of northern France were codified in 1453, and modified Roman law became a main source of imperial German law in 1495 and of Scots law in 1532. At the same time, the scope of canon and Roman law in England was increasing. Admiralty law, for example, drew on Greek, Roman, and Italian law and used documents drawn up in continental form, and the crimes of forgery and libel tried in Star Chamber were based on Roman models. Ecclesiastical courts applied canon-law rules based on Roman law, for example, to wills and marriages. The Councils of Wales and the North also used Roman law. All of these bodies competed with common-law courts for jurisdiction over the same cases and followed a written procedure modeled after that still being used on the Continent. Roman law and canon law, furthermore, were taught at Oxford and Cambridge, which gave doctorates to the practitioners in these courts.

One of the accusations reportedly made against Thomas Cardinal Wolsey, who fell from favour in 1529, was that he planned to introduce Roman law into England; Wolsey did appoint many clergy to the Council of the North and as justices of the peace. The 19th-century English legal historian F.W. Maitland discussed this legal crisis in a famous essay on English law and the Renaissance. Maitland ascribed the survival of the common law, in part, to the solid front presented by the Inns of Court, which trained lawyers practically and not theoretically. The English law tradition did not depend on abstract scholarly commentaries but on detailed judicial rulings about specific points of law arising in practice.

The influence of Roman-law ideas, however, was probably greater than generally admitted. The actions of trespass and disseisin (dispossession) had Roman analogies, and the estate tail was clearly influenced by a law made by Justinian. The equitable remedy of injunction had analogies in canon law, and the law of redemption of mortgages may have been related to the usury laws, which forbade making excessive profits from loans. The law of trusts and deceit resembled the breach of faith of the church courts. Continental mercantile law, which contained Roman-law elements, was absorbed into English law as it stood. Continental law also contributed to some of the rules of contract, such as the effect of mistake, and the Roman concept of fault played a part in the law of negligence. Many old European legal ideas, in fact, survived longer in England, where they escaped being eliminated in codifications, than in Europe.

Coke as a champion of the common law

An account of the development of common law in the Tudor-Stuart period would be incomplete without mention of Sir Edward Coke. Coke, who combined a distinguished career as a barrister and a judge, produced a wealth of legal writings. In 1606 he risked removal from the office of chief justice by challenging the exaggerated claims of the royalist party to prerogative powers outside of the common law. He disapproved of legislation by proclamation, of dispensation from the law in individual cases, and of the mushrooming jurisdictions of the prerogative courts. He helped draft the Petition of Right in 1628.

Coke's 11 volumes of *Reports* appeared between 1600 and 1615, and two posthumous volumes followed. Coke commented, rather than reported, but he was careful to supply a copy of the court record of each case. As the only formal series of collected law cases available at the time, his reports formed the main source for the citation of cases for many years. His four volumes of *Institutes of the Lawes of England,* published between 1628 and 1644, dealt with the law of real property (*Coke on Littleton*), the medieval statutes, the criminal law (pleas of the crown), and the jurisdiction of the courts.

Coke was no objective historian but an open advocate of the common law. Though he was old-fashioned and at times in error, his greatest works restated the common law in acceptable form and did much to save it.

Further growth of statute law. The Tudors made use of proclamations by the king to invoke emergency measures, to establish detailed regulations, especially on economic matters, and to grant royal charters to trading companies. Parliament passed laws of a political character, such as those enforcing the king's supremacy over the new established church. Statutes also regulated imports and exports, farming, and unfair competition. A law of 1562–63 regulated apprenticeships and provided for annual wage fixing by magistrates in accordance with the cost of living.

Among other statutory innovations were the Statute of Monopolies of 1623, which confirmed that monopolies were contrary to common law but which made exceptions for patentable inventions; a statute of 1601 that became the basis of the privileges enjoyed by charitable trusts; and the series of Poor Laws, which were enacted in the late 16th century to remedy the neglect of the poor caused by the dissolution of the monasteries.

In 1540 legal actions to recover land were subjected to time limits, and in 1623–24 the principle of limitation of actions by lapse of time was introduced into the law of contract and tort.

During the Commonwealth (1649–60) many reform projects were drafted; although they anticipated 19th-century reforms, none of them was carried out. These reforms included supplying counsel to prisoners, modernizing the land and law procedure, and permitting civil marriages.

The outstanding enactment of the later Stuart period was the Statute of Frauds of 1677. As a response to the growth of literacy and the prevalence of perjury and fraud, wills and contracts for sale of land or goods (of more than a certain amount) were required to be in writing. Though drafted by eminent judges, the statute was to require endless interpretation.

Further development of equity. Although one eminent contemporary observer, the legal historian John Selden, regarded the fate of a lawsuit in Chancery as varying with the chancellor's personality, the types of suits that would be granted relief had eventually become fairly clear. Precedent was being followed, and law reports of equity decisions and books on equity began to be published.

In 1615 the King declared that the Chancery was to retain its traditional superiority over the common-law courts but only in areas in which its authority was well recognized. If the applicability of equity was in doubt, the common law was to be followed.

Developments in the law of trusts

The main development in this period was in the law of trusts (see PROPERTY LAW). In medieval England, from the 14th century, most land was held "to uses"; *i.e.,* by nominees for the true owners. This situation may have been partly due to devices used to evade taxation, but it also enabled wills of land to be made. "Death duties" were payable if a man died while he was the legal proprietor; by transferring the land to another person, these could be avoided. Wills of land were not allowed before 1540, but the use of land could be transferred to another person while the owner was still alive, as long as the transferee observed the owner's wishes regarding the land while the owner lived. The beneficiary of such a trust usually stayed on the land as apparent owner, though the trustee held the legal title. A statute of Richard III, however, allowed the beneficiary to transfer the property, and in 1535–36 the Statute of Uses eliminated the middleman and revested the legal title in the beneficiary. The device of the use was exploited to create new and complicated legal interests in land. The old use was revived as the modern trust in Chancery, first for trusts involving money and leases and finally for trusts of land itself. The spur was the desire to separate the legal and beneficial titles, especially when the beneficiary was young or inexperienced. But the trust was adapted to many other ends, such as giving property to clubs and other unincorporated bodies and to churches.

THE MODERNIZATION OF COMMON LAW IN GREAT BRITAIN

Influence of Blackstone. Of extraordinary influence in the development of common law and in its dissemination to other parts of the world was the most famous of English jurists, Sir William Blackstone. Born in 1723, he entered the bar in 1746 and in 1758 became the first person to lecture on English law at an English university.

His most influential work, the *Commentaries on the Laws of England,* was published between 1765 and 1769 and consisted of four books: "Persons" dealt with family and public law; "Things" gave a brilliant outline of real-property law; "Private Wrongs" covered civil liability, courts, and procedure; and "Public Wrongs" was an excellent study of criminal law.

Blackstone was far from being a scientific jurist and was criticized for his superficiality and lack of historical sense. The shortcomings of the *Commentaries* in these respects, however, were offset by its style and intelligibility, and lawyers and laymen alike came to regard it as an authoritative exposition of the law. In the following century the fame of Blackstone was even greater in the United States than in his native land. After the Declaration of Independence the *Commentaries* became the chief source of knowledge of English law in the New World.

Reform in the 19th and 20th centuries. *Bentham.* Following the social turmoil of the French Revolution and the economic upheaval of the Industrial Revolution, there were many demands for reforms to modernize the law. The most significant figure in the reform movement was the English Utilitarian philosopher Jeremy Bentham, who was prepared to reform the whole law along radical lines. A brilliant student, Bentham disliked the picture of the law that was presented in Blackstone's lectures. In 1769 he entered the bar, but since he was living on an inheritance, he never found it necessary to enter practice. He worked to make law less technical and more accessible to the people, but he was slow to complete or publish his writings, and not until 1789 did his basic work, *An Introduction to the Principles of Morals and Legislation,* appear.

Bentham attacked legal fictions and other historical anomalies. He advocated two basic changes in the legal system: in order to achieve the greatest happiness for the greatest number, legislators, rather than courts, should make the law; and the aims of law should vary with time and place.

The fame of the *Principles* spread widely and rapidly. Bentham was made a French citizen in 1792, and his advice was respectfully received in most of the states of Europe and America. Although he wanted most of all to be allowed to draw up a legal code for his own or some foreign country, his practical influence was far more indirect and derived largely from the diffusion of Utilitarian ideas during the 19th century.

Increased protection of the rights of the accused

Changes in procedure and criminal law. In England the restrictive framework of the separate forms of action in civil cases was replaced in 1852 by a new system of uniform writs of summons, and liberal amendment of pleadings was permitted. Fixed dates were established for trials. In 1933 jury trial was ended in civil cases, except in libel and a few other actions. Evidence acts of 1938, 1968, and 1972 simplified civil proof. A major trend in criminal procedure since the early 19th century has been better protection of the rights of the accused. Since 1836 the accused has been entitled to counsel and since 1898 has been allowed to testify on his own behalf. In 1903 provision for the state to pay for defense was made and since expanded, and in 1907 the right of appeal against criminal convictions was created. In 1967 verdicts by a majority of the jury were made possible, and restrictions were imposed on press coverage of preliminary hearings.

The 19th century saw the enactment of a series of statutes that codified the part of criminal law dealing with individual crimes, apart from homicide. Basic ideas have changed little, other than the fact that some modern statutes have imposed responsibility without fault and that corporations can now be held responsible for the acts of their management.

The rules of legal insanity were laid down in the 19th century and supplemented in 1957 by the limited defense of "diminished responsibility." Capital punishment was gradually ended for most felonies and was finally eliminated for murder by the Homicide Acts of 1957–65. In 1968 a new Theft Act replaced the rather crude medieval idea of larceny by a broader concept that resembles the Roman delict (offense) of theft. Experimentation has led to new remedies, one of these being the suspended sentence, which only has to be served if a further crime is committed.

Reorganization of the judiciary. The lay jurisdiction of the church courts ended in 1857, when the divorce and probate courts were set up. These merged into the High Court of Justice in 1875 as a result of the Judicature Acts of 1873–75, which reformed the civil courts. The Judicature Acts were much more than a regrouping and renaming of courts; they attempted to fuse law and equity by making available legal and equitable remedies in all divisions of the High Court and by providing that the equitable rule should prevail when conflicts arose. Common law and equity nevertheless preserved their separate identities, partly because of the different subject matter with which they often dealt and partly because lawyers persisted in maintaining the distinction.

In the late 19th century the three central courts of common law were amalgamated as the Queen's Bench Division, which to this day continues to try suits for damages. Since 1875 cases have been tried by a single judge (before 1933 with a jury), not by a full bench of judges.

After it became a division of the High Court in 1875, the Chancery not only dealt with equity suits but also administered the voluminous legislation on property, bankruptcy, succession, copyrights, patents, and taxation. Contested probate cases were transferred to the Chancery by the Courts Act of 1971.

Before the Courts Act criminal cases were tried two or three times a year at assizes or four times a year at quarter sessions in the provinces. As of January 1972 a system of provincial crown courts replaced these. Civil assizes were replaced by allowing the High Court to sit at certain cities.

Small civil cases, tried at statutory county courts since 1846, are now regulated by an act introduced in 1984.

A remarkable feature of English criminal justice, as compared with most European systems, has been the continuing role of lay justices of the peace, who remain important despite the appointment of paid, legally trained magistrates in London and some of the larger cities, of barristers as recorders at borough quarter sessions, and of legally qualified chairmen at county quarter sessions. An important aspect of the magistrates' work has been their jurisdiction over young offenders, for whom special juvenile courts were first set up in 1908. The report of a royal commission on justices of the peace in 1948 strongly defended the position of lay justice against public criticism; its cautious recommendations as to the appointment of justices and as to the organization of their courts were largely put into effect by the Justices of the Peace Act (1949) and the Magistrates' Courts Act (1980). The Criminal Justice Administration Act (1962) extended the power of justices of the peace to try indictable offenses summarily. A series of statutes in 1972, 1973, 1977, 1981, and 1982 rendered the procedure more flexible, made detailed provision for penalties and their execution, and added a number of new offenses. In 1964 elementary judicial training for lay justices was introduced. These developments since 1948 show both the persistence in English law of ancient institutions and a preference for reforming rather than totally abolishing them.

Routes of appeal

A modern appellate court for civil cases in the High Court was set up in 1830 but was replaced in 1875 by a Court of Appeal consisting of special appellate judges. In 1907 a Court of Criminal Appeal was established, but it was merged into the Court of Appeal in 1966. A divisional court hears appeals from magistrates on points of law. A final appeal, subject to conditions, can be made to the House of Lords from all lower courts.

Reform in private law. Property law has been changed often. Wills are regulated mainly by a statute of 1837 (amended in 1982), and the freedom to disinherit has been curtailed by family provision acts of 1938, 1952, 1966, and 1975. Title to land is subject to a system of registration that has been gradually introduced under an act of 1925. Succession on intestacy (*i.e.,* in the absence of a valid will) for all kinds of property was unified in the same year. The law of leases has been modified by social legislation such as the numerous Rent (control) Acts, which protect residential tenants. The terms of trusts can be modified by the Chancery (since 1958), and a wider range of trustee investments has been allowed since 1961.

Grounds for divorce have been enlarged by a number of 20th-century statutes, culminating in the broad "breakdown of marriage" approach of the Divorce Reform Act of 1969, now the Matrimonial Causes Act of 1973 (as amended in 1984). Under this legislation a marriage may be terminated not only on traditional fault grounds but also when the parties have lived apart for at least two years and consent to divorce or when the parties have been separated for at least five years.

After several piecemeal laws addressed trade (labour) unions, a more comprehensive, but controversial, Industrial Relations Act was passed in 1971, requiring registration of unions and arbitration of disputes. This statute was repealed in 1974, but aspects of it were revived with considerable modification in 1980 and 1982.

In the field of tort, manufacturers' liability to consumers was established by case law in 1932 and later strengthened by legislation. Liability in libel has been cut down by many statutes. A law of 1945 introduced the Roman principle of apportioning damages when both parties are at fault.

Commercial law, with the Bills of Exchange Act (1882), Sale of Goods Act (1893 and 1979), the Unfair Contract Terms Act (1977), and consumer protection statutes in 1965 and 1974, has become primarily the domain of legislation. Arbitration, too, is regulated by statute.

THE DEVELOPMENT OF COMMON LAW IN THE UNITED STATES

The first English settlers on the Atlantic Seaboard of North America brought with them only elementary notions of law. Colonial charters conferred on them the traditional legal privileges of Englishmen, such as habeas corpus and the right to trial before a jury of one's peers, but there were few judges, lawyers, or lawbooks, and English court decisions were slow to reach them. Each colony passed its own statutes, and governors or legislative bodies acted as courts. Civil and criminal cases were tried in the same courts, and lay juries enjoyed wide powers. English laws passed after the date of settlement did not automatically apply in the colonies, and even presettlement legislation was liable to adaptation. English cases were not binding precedents. Several of the American colonies introduced substantial legal codes, such as those of Massachusetts in 1648 and of Pennsylvania in 1682.

The reception of English law

By the late 17th century, lawyers were practicing in the colonies, using English lawbooks and following English procedures and forms of action. In 1701 Rhode Island legislated to receive English law in full, subject to local legislation, and the same happened in the Carolinas in 1712 and 1715. Other colonies, in practice, also applied the common law with local variations.

Many legal battles in the period leading up to the War of Independence were fought on common-law principles, and half of the signatories of the Declaration of Independence were lawyers. The U.S. Constitution itself uses traditional English legal terms.

After 1776 anti-British feeling led some Americans to advocate a fresh legal system, but European laws were diverse, couched in foreign languages having unfamiliar turns of thought, and unavailable in textbook form. Blackstone's *Commentaries,* reprinted in America in 1771, was widely used, even though new English statutes and decisions were officially ignored.

In the 1830s two great judges, James Kent of New York and Joseph Story of Massachusetts, produced important commentaries on common law and equity, emphasizing the need for legal certainty and for security of title to property. These works followed the common-law tradition, which has never been fundamentally altered in the United States, except in Louisiana, where French civil law has survived.

American innovations. The American states saw law as a cementing force, and they used it to facilitate cooperation in the face of the hazards of nature and other difficulties arising in the development of the new continent. Special laws were developed to deal with timber, water, and mineral rights. Simple procedures were followed. Dogma was rejected in favour of personal experience and experiment, and old decisions soon became outdated. The pioneer spirit favoured freedom and initiative and distrusted central authority and a paternal government. Homespun local justice was preferred, as was the common sense of the local jury. For a time some of the colonies even tried to base their law on the Bible. But, even when English law reasserted itself, many of its institutions were rejected. On death intestate, for example, all of the children inherited land in America and not just the eldest son, as in England. Freehold title was the rule, not long leases under landlords. Church courts did not exist.

Growth of statute law and codes. After the War of Independence a drive to replace judge-made law by popular legislation was revived. In 1811 Jeremy Bentham proposed a national civil code to Pres. James Madison, but his proposal was premature. In the mid-19th century, the legal reformer David Dudley Field presided over the drafting of several codes and campaigned vigorously for the systematic, rational codification of U.S. law. Except for a code of civil procedure, which was widely copied, Field's codes found little acceptance in state legislatures. Field's civil code was adopted by five states, including California and New York, but the common-law tradition was so strong in these jurisdictions that the civil code became just another statute; it was read against the background of, and supplemented by, existing case law, rather than seen as a complete set of authoritative starting points for legal reasoning as were the continental civil codes. Louisiana, whose legal system is a hybrid of civil- and common-law elements, is the only American state that has a code in the civil-law sense. Despite the failure of the codification

Codes of David Dudley Field

movement, U.S. law became increasingly statutory, so that by the late 20th century legislation predominated over judge-made law.

U.S. statutes are not construed so narrowly as those in England, and there is less reluctance to change the older law. Statutes are also regularly revised; for example, New York state has had a Law Revision Commission since 1934.

Equity and probate. In 18th-century England the Court of Chancery administered equity and the church courts handled the probate of wills. In the American colonies the governor and his council acted as a court of equity. For a time after independence, equity was suspect as a remnant of royal prerogative, but it has come to be generally applied by the same court as the regular law. Although U.S. common law is more flexible than English law, and the need for equity is less, important remedies have nevertheless been developed within the system. Probate, with a few exceptions, is usually a matter for the regular courts.

Federal and state judicial systems. State courts try 90 percent of all civil and criminal cases. Local magistrates may sit on county or district courts. One appeal is always given, and two levels of appeals exist in many states. The highest court is usually called the supreme court of the state, but this varies. In New York state, for example, the Supreme Court is a trial court, and the highest court is the Court of Appeals.

The Constitution of 1789 set up a federal Supreme Court, and the 1789 Judiciary Act provided for federal district courts and circuit courts. The plan for inferior courts has undergone changes from time to time, notably in 1891, when circuit courts of appeal were established, and in 1911, when the old circuit courts were abolished.

Most federal law is statutory and enforced by federal courts. Laws concerning tax, labour, securities regulations, admiralty, interstate commerce, antitrust, patent, and copyright matters fall into this category. By a decision of 1803, the Supreme Court became the ultimate authority for determining the conformity of all legislation with the federal Constitution, which guarantees many fundamental rights.

To ensure the fair treatment of out-of-state citizens or of corporations incorporated elsewhere, federal courts can try cases involving a diversity of citizenship. In such cases they act as if they were state courts, however, being bound by state statutes since 1842 and by state interpretations of common law and equity since 1938. Federal procedure is followed, but state rules on vital matters, such as statutes of limitations, are enforced.

Federal courts also try claims by and against the United States, such as cases undertaken to protect federal assets. In the absence of statutory provisions for such cases, a "federal common law" is applied.

Personal and property rights. The guarantees of due process of law given in Magna Carta in 1215 and the English Bill of Rights of 1689 are reflected in the first ten amendments to the federal Constitution, which were passed in 1791 and are known as the Bill of Rights. Since the passage of the Fourteenth Amendment in 1868, the rights of life, liberty, and property have been protected from deprivation by both the states and the federal government without due process of law; this has tended to shield private property from government regulation and private contracts from government interference. The use of property, however, is increasingly restricted by zoning laws and health and safety measures, and the acquisition of property for public purposes may be justified under the doctrine of eminent domain (power of the government to take private property for public use without the owner's consent upon payment of compensation).

The 1929 Depression was followed by the rejection by the Supreme Court of many welfare measures. Since 1937, however, the power of the Congress to regulate the economy under its authority to oversee interstate commerce has generally been upheld by the Supreme Court. State legislation is, as a rule, also held to be constitutional in this area. Minimum-wage laws and the right to collective bargaining in industry are recognized as well.

Human rights in the Supreme Court

Since the 1950s the emphasis in constitutionality cases has shifted to human rights. The requirement of "equal protection of the laws" and the Civil Rights Act of 1866 led to the ruling in 1954 that public schools must be racially integrated and to later rulings against using public funds for segregated private schools. The Federal Civil Rights Act of 1964 applies not only to official laws and actions but also to the conduct of private citizens. Thus, no discrimination on the basis of race, sex, religion, or national origin is allowed in places of public entertainment or resort or in employment practices by larger firms.

Since 1962 the Supreme Court has insisted on a regular redrawing of electoral districts to give each vote roughly the same value (seat reapportionment). It has also interpreted the constitutional prohibition of the establishment of a state religion to render school prayers and religious instruction illegal. In 1971 freedom of the press was held to justify *The New York Times* in publishing confidential political material.

COMPARISONS OF ENGLISH, AMERICAN, AND COMMONWEALTH LAW

The legal systems rooted in the English common law have diverged from their parent system so greatly over time that in many areas the legal approaches of common-law countries differ as much among themselves as they do with the civil-law countries. Indeed, England and the United States have so many legal differences that they are sometimes described as "two countries separated by a common law." The most striking differences are found in the area of public law: England has no written constitution and no judicial review, whereas every court in the United States possesses the power to pass judgment on the conformity of legislation and on other official actions to constitutional norms. Throughout the 20th century, many areas of U.S. law have been "constitutionalized" by the increasing exercise of judicial power. Other factors that account for much of the distinctiveness of public law in the United States are its complex federal system and its presidential, as distinct from parliamentary, form of government. In the area of private law, however, family resemblances among the common-law systems are much greater. Yet even there, despite broad basic similarities, the common-law countries have developed distinctive variations over time.

Personal law. The law of personal status (nationality, capacity, domicile, and so on) has been transformed by the advancement of the principle of equality of the sexes. In the area of divorce law, the intense legislative activity of the 1960s and 1970s left most common-law countries with systems of "mixed grounds" for divorce: one can obtain a divorce either for the fault of the other spouse or upon some no-fault ground such as separation or breakdown of the marriage. A minority of U.S. states have eliminated fault grounds entirely. The major differences among common-law systems appear in the legal treatment of the economic consequences of divorce: most common-law countries follow the English model that permits judges to use their own discretion in reallocating the property and income of the spouses in the way that seems fair; a minority of U.S. states adhere to the principle of equal rather than discretionary division of assets.

Property and succession. The basic principles of property and succession are much the same everywhere, but the newer countries have special laws on forests, mines, and water rights. In Australia, for example, the crown reserves all mineral rights to itself. The transfer of land in England is governed by a system of title registration. In Canada and the United States the separate deeds are recorded, and title insurance is widely used to protect the purchaser.

Succession on intestacy is broadly similar throughout common-law countries but varies everywhere in detail. The widow, for example, may get more in one country and the children more in another. All children of both sexes generally take equal shares. In regard to testate succession, nearly all U.S. states protect the surviving spouse against disinheritance by securing to him or her a fixed indefeasible share of the decedent's estate. In England, and most of the former Commonwealth countries, however, not only the spouse but also children and certain other dependents of the deceased are permitted to petition the court for discretionary financial provision out of an estate if, in the judgment of the court, the testator did not make reasonable provision for them.

In most U.S. states and some Canadian provinces there are homestead laws, which protect the family house or a certain minimum sum of money from the claims of creditors.

Extensions of tortious liability

Tort law. Tort law (*i.e.,* the law relating to private civil wrongs) is largely common law in England, Canada, and the United States. Several major reforms have been introduced along the same lines in different countries. Allowing claims by dependents of persons tortiously killed and removing the immunity of the crown or government or charitable institutions from tort claims provide examples. The liability of manufacturers to the ultimate consumer was first laid down by U.S. and then by English judges.

In the field of libel, U.S. practice is less strict than the English, and in the United States a public figure cannot sue for honest but unfair and untrue criticisms of his activities, whereas in England published facts must be true and comment fair. In some Australian states truth is not necessarily a defense to an action.

A notable U.S. tort is interference with privacy, examples being a stranger using one's photograph for advertising without permission, using "bugging" (*i.e.,* electronic eavesdropping devices) in one's home or searching it, or taking photographs of persons in embarrassing situations.

Contracts. Contract law is basically similar in the common-law countries. The most interesting difference relates to the question of enforcement of contracts by third parties who are not actually parties to the contract but who are persons for whose benefit the contract was made. English law excludes such rights, except in an occasional statute. The Indian Contract Code of 1872 generally allows it, as does U.S. state law.

English law still requires the use of a seal on a gratuitous contract (such as one agreeing to make a gift) but has largely repealed the laws requiring written evidence of ordinary contracts. Written evidence is often called for in the United States.

The various areas of special contracts, such as those applying to employment, sale of land, and agency, are broadly similar everywhere but are regulated by local legislation and by a wealth of labour legislation.

Criminal law and procedure. As regards criminal law and procedure, the substance of the law is much the same throughout the common-law countries. More important differences appear in the rules of criminal procedure. This rests in England on modern legislation, whereas the old procedure bore heavily on the accused. Accused persons may now testify at the trial or not, as they wish; they are entitled to legal counsel; and they are assisted out of public funds when they are accused of serious crimes and are unable to afford to pay the costs themselves.

Canada has a Dominion Criminal Code, which covers major crimes. It also has a Canadian Bill of Rights and provincial laws such as the Ontario Human Rights Code. India has an overriding Bill of Rights.

Developments in the United States are the most interesting. Criminal procedure has become a constitutional matter, with a kind of federal common law of criminal procedure overriding state law in many instances. Thus, "due process of law" under the Fourteenth Amendment to the federal Constitution and the Federal Rules of Criminal Procedure confer wide protection on accused persons—too wide, some think, for public safety.

Different approaches to the admissibility of evidence

English courts are reluctant to admit tape recordings unless supported by direct evidence of persons present, and this is generally the position taken in the United States, although, with the permission of a court, emergency wiretapping is permitted. English and U.S. law exclude confessions unless they are made freely and spontaneously. If evidence is found by unlawful means, such as by searching a house without a warrant, English law permits such evidence to be used, but U.S. law does not.

The main difference between English and U.S. safeguards is that English protections rest on statute or case law and may be changed by ordinary statute, whereas U.S. safeguards are constitutional and cannot be relaxed unless the Supreme Court later reverses its interpretation or the Constitution is amended.

The future of the common law. In the past the law performed the function of a referee in a free economy and was called in to apply generally accepted ideas of right and wrong to individual disputes. Today, law often forms an instrument of governmental policy or results from social pressures on the government. Law, therefore, is increasingly administrative.

Another tendency, and one that is likely to be reinforced, is an increasing reliance on statute law and codification as instruments of legal development. At one time the English Law Commission considered drafting a contract code, and the law of tort has been the subject of several statutes. When Britain entered the European Economic Community it was thought that there might be pressures to make English law more accessible by codifying it along the lines of the continental model. Harmonization of the laws of the member states, however, has not thus far required this. In the United States the legal sovereignty of the states impedes such a radical change, but uniform state laws are becoming more common.

In view of the general tendency in modern society of shielding the individual as fully as possible from the consequences of chance accidents, the judge-made law of tort may in time be replaced, as it has been in New Zealand, by a comprehensive system of official or private insurance, similar to the present compulsory third-party risk insurance available for motor vehicles. Public law is also gaining on private law in other fields—in real-property development, for example, public zoning or town-planning rules are already more important than the traditional restrictions imposed by individual neighbouring landowners. In family law, public-welfare laws on child care and adoption, pensions, and social security are often more important than the older private law based on the rights of spouses and children. (A.R.Ki./M.A.Gl.)

Comparison of civil law and common law

Between the 11th and 15th centuries the law of England was strongly influenced by Roman-law learning, and, in the 16th century, experts trained in Roman law were welcomed as administrators as much by the kings of England as by continental rulers. In contrast to the countries of the Continent, however, where justice was administered locally, the English judicial system had, as a result of the Norman Conquest, been centralized. There had grown up at the king's courts of Westminster a profession of practitioners expert in the law and procedure of the centralized court system, strongly organized and unwilling to yield its position, power, and income to a new group of specialists of Romanist learning. In its resistance to royal innovation, the organized bar allied itself with the parliamentary party in the great constitutional struggle of the 17th century. Thus, a reception of Roman law, such as had developed on the Continent, was prevented in England. At the same time a strong connection was established between the principles of constitutionalism and individual freedom on the one side and the common law on the other, which over time fostered the image of the common law as the legal system of freedom, in contrast to the civil law, in which the state is exalted over the individual. This image gained additional support from the fact that free political institutions were developed earlier and have been maintained more firmly in countries of the common law than in countries of the civil law.

Legal procedure in free societies

Political institutions are one thing, however, and techniques of dealing with civil litigation and criminal prosecution are another. Intimate connections exist, of course, between the two; a society is not free if civil cases are not handled impartially and if persons accused of crimes are not safeguarded against injustice. In both of these respects, however, neither of the two great legal systems lags behind the other. The ways of argumentation and procedure differ, but each of the two systems has developed its own guarantees and safeguards, and neither can be shown to be superior to the other. The view frequently found in

England and the United States, that in civil-law criminal procedure the accused is presumed guilty until he has proved his innocence, is as unfounded as the view widely held on the Continent that trial by jury is tantamount to a lawless appeal to passion and emotion.

If one compares countries having firmly established institutions of constitutional government, such as the United Kingdom and the United States, to civil-law countries, such as Belgium, The Netherlands, Switzerland, the Federal Republic of Germany, or France, it appears that the protection of the individual against illegal actions by executive agencies is generally about the same. In some respects individual protection is even more elaborate in civil-law countries than in the United States and the United Kingdom. In general, it also may be said that it is less expensive for citizens on the Continent to seek legal protection of their private rights than it is in the common-law countries.

It is difficult to define what constitutes the real difference between common law and civil law. It would be erroneous simply to identify civil law with codified or even statutory law and common law with judge-made or case law. For one thing, the contrasts between the two systems existed long before the civil-law countries began to enact their codes. In addition, large parts of Anglo-American law are also contained in statutes or even codes, while in France, West Germany, and other civil-law countries parts of the law have never been reduced to statute at all but have been developed by the courts. Also, many of the statutes and code provisions have come to be overlaid by so many judicial opinions and interpretations that, in effect, they are dominated by judge-made law.

Role of judicial precedent

No essential difference can be found, either, in the role of judicial precedent. In theory, common-law courts are bound by precedent in the sense that once a legal question has been decided a certain way by a court, it must be decided in the same way by courts of inferior rank in the same jurisdiction until a higher court or the legislature sees fit to change the rule. Though higher courts are not "bound" by their previous decisions, they habitually follow them in the absence of a strong justification for overruling. In civil-law countries, on the other hand, courts are, in official theory, free to consider anew any legal question irrespective of how often it may have been determined before by other courts. In practice, however, common-law courts, especially those in the United States, have developed techniques for distinguishing new cases from older ones so that observable adherence to precedent is less a matter of strict obligation and more an acknowledgement of the importance of maintaining reasonable predictability in the law and of supporting the principle that like cases ought to be decided alike. Civil-law courts, for their part, have tended to follow precedent, not only for the sake of continuity and evenhandedness, but also in accord with the inclination of courts everywhere to save time and effort and to avoid reversal by a higher court.

The main difference between the systems consists of the ways in which the norms of the law are articulated and in which new rules are derived from older ones in novel cases. Though law cannot remain static, adaptation to new circumstances should be orderly and gradual so as not to unduly undermine predictability and stability in human affairs. In the common law, the role of adapting the law to changing conditions has traditionally belonged to judges. Because the administration of justice has been centralized, English judges have been well situated to assume this responsibility. In civil-law countries, where the multiplicity of courts has prevented this task from being performed by the judges, it has been professors who have taken the leading role.

Role of the professors

This difference in the identity of the influential lawmakers has had far-reaching consequences in the development of the law. Since courts must proceed from case to case, and cases arise in isolation and without prearrangement, a judge's opportunities for systematic theorizing are limited. Professors, by contrast, often deal with hypothetical cases as well as actual ones. They can develop comprehensive ideas and principles, and they are impelled for pedagogical reasons toward systematization and conceptualization. The civil law, as a professorial product, has thus tended to become more theoretical and consistent in its propositions and terminology than the judge-made common law, which has tended, for its part, to be closer to life and perhaps more detailed.

These traditional differences may diminish, however, for in the civil-law countries judicial power has increased with the national centralization of the administration of justice and with the assumption of a more active role by the judiciary, while in most of the common-law parts of the world the centralized English courts long ago lost their supremacy to a multiplicity of supreme courts in the United States and in the Commonwealth. The role of maintaining the internal coherence of the law is thus increasingly shared by the courts with that group of professionals—the professors—who traditionally have performed this function in the civil-law world. Gradual assimilation of the legal methods of the two great systems of law may thus well be expected.

Role of the adversary system

What is likely to continue, however, is the marked difference between the two great families of legal systems that exists in the field of procedure and, to some extent, in the personnel by whom justice is administered. All civil-law countries have adopted the adversary type of procedure that for centuries was peculiar to the common law, and they have abandoned the canonical procedure in which the evidence was presented to the judge in the form of a written record made up by a public officer, mostly in the absence of the parties. In modern adversary procedure, lawyers address their arguments directly to the court in both the civil-law and the common-law systems. Certain differences, however, are striking. In common-law systems the parties and their lawyers gather and present factual evidence in each case; in civil-law systems greater responsibility for investigation of facts is placed upon the judge, and it is generally the judge who plays the leading role in examining witnesses and who summons experts when needed. In general, the civil-law judge plays a much more active role in directing the course of a lawsuit than does his common-law counterpart.

Many differences between civil-law and common-law procedures and rules of evidence seem attributable to the absence of a jury in civil cases in the continental systems. During the 19th century, several civil-law countries experimented with the use of a jury for criminal cases. During the 20th century, however, the criminal jury has been abandoned in favour of a system having a mixed bench, on which professional, legally trained judges sit with a jury of laymen or lay judges to decide not only questions of fact (as in the common law) but also those of law.

(M.Rh./M.A.Gl.)

Variant and hybrid legal systems

Some Western legal systems cannot clearly be classified as belonging to the civil-law, common-law, or Socialist law traditions. Some developed in relative isolation from other systems, such as those in the Nordic countries; more often, variant legal systems arose from a confluence of influences, such as those that occurred when one colonial power succeeded another. Notable among these variants and hybrids are the Scandinavian, Roman-Dutch, and Scottish systems. Scandinavian law, though Germanic in origin and influenced by a revival of Roman law, is generally classified as sui generis and is set apart from the civil law by several distinctive features. Roman-Dutch law, an amalgam of Roman *jus commune* and the law of early modern Holland, survives in several former Dutch colonies in an altered form and in combination with several other legal influences. In Scotland, Roman law and common law have combined to form a hybrid system.

(M.A.Gl.)

SCANDINAVIAN LAW

Scandinavian law in medieval times constituted a separate and independent branch of early Germanic law and in modern times, in the form of codifications, became the basis of the legal systems of Norway, Denmark, Sweden, Iceland, and Finland.

Historical development of Scandinavian law. Before the Scandinavian states emerged as unified kingdoms in the 9th century, the several districts and provinces were virtually independent administratively and legally. Although social organization in the main was the same, and legal developments followed similar lines, there came into existence a number of separate legal systems, or "laws." Originally there were no written laws; the legal system consisted of customary law that was conserved, developed, and vindicated by the people themselves at the so-called *things,* or popular meetings of all free men. Between the 11th and 13th centuries the provincial customary laws were recorded in writing (invariably in the vernacular). These writings were most often private compilations but were occasionally instructions from the king. The best known laws of this period are the Gulathing's law (written in the 11th century, Norwegian); the law of Jutland (1241, Danish); and the laws of Uppland (1296) and Götaland (early 13th century), both Swedish. Other Scandinavian communities and states followed suit.

The early laws or codes did not have the character of civil codes as they are understood today. In addition to the subjects of private law (matrimony, inheritance, property, and contract), they contained constitutional and administrative law, criminal law, and laws of procedure. Ecclesiastical law was usually excluded and treated separately. In the main, the codes represented collections of customary law; influences from abroad were negligible except for some traces of canon law. Whereas the provincial laws, in common with other early Germanic laws, had tolerated and regulated blood feuds (setting up detailed tariffs for manslaughter and offenses against the body), the codes are, in several respects, more progressive. Thus, King Magnus' Swedish code (1350) abolished private vengeance, declaring that the king's officials should initiate criminal proceedings and provide for the punishment of wrongdoers. Furthermore, presumably under the influence of Christianity, legal provisions were introduced to assist paupers and the helpless. Rules concerning landed property (*e.g.,* the right of redemption belonging to the family) were markedly original.

In 1380 Norway and Denmark were united under a common king (Olaf IV), but the two countries retained their separate laws. During the next 300 years, before the acquisition of absolute royal power by Frederick III (1660), supplementary laws were issued by the king in conjunction with an assembly of nobles. Finally, during the reign of Christian V, a comprehensive work of codification was accomplished, and the earlier and often obsolete law was replaced by Christian V's Danish Law (1683) and Norwegian Law (1687). The new codes were mainly based on the existing national laws of the two countries, and the influences of German, Roman, and canon laws were comparatively slight. Like the early codes, the newer codes consisted of public as well as private law and purported to treat exhaustively all more or less permanent legal rules and institutions. They were excellent codes for their times, drafted in a plain and popular style and inspired by respect for individual rights and the idea of equality before the law. The provisions of criminal law were relatively humane when compared with legislation in other European countries.

Codification under Christian V

In Sweden a revised edition of the original code, issued by King Christopher (1442), was expressly confirmed by Charles IX (1608). The need for more modern legislation, however, made itself increasingly felt, and following the Danish-Norwegian example a royal commission was entrusted with the task of drafting a new code. The result, commonly called "the Law of 1734," was promulgated by Frederick I.

Finland, annexed by Sweden in the 13th century and made subject to Swedish law, came under the Swedish code of 1734, which was translated into Finnish as "Law of the Realm of Finland."

Modern Scandinavian law. The old codes have been all but completely displaced by modern parliamentary statutes. In Sweden the law of 1734 has been conserved as a formal framework. Elsewhere, plans for new and all-embracing codes are no longer entertained, but an extensive codification of important parts of the public and private law has taken place.

An interesting feature of Scandinavian law is the organized legislative cooperation that was begun in 1872 and has steadily increased in importance. In this way the Nordic states, including Iceland and Finland, have to a considerable degree obtained uniform legislation, especially regarding contracts and commerce, as well as in such fields of law as those concerned with family, the person, nationality, and extradition.

While conserving their national character, the Scandinavian legal systems have adopted certain conceptions of civil law (mainly German and French), chiefly through the influence of the law schools; commercial law and the laws of shipping and of companies, for example, conform more or less to common European patterns. Modern social welfare legislation, which has reached a high standard, also has strong international connections. Scandinavian law is pliable and close to life, less dogmatic than other European legal systems, and relatively free of formal rules and exigencies. Great attention is paid to rules and principles that have evolved in practice, especially in the courts. Much of the law is judge-made; and because the principle of stare decisis (*i.e.,* being bound by precedent) does not obtain, the courts have been free to meet the demands of changing social conditions. The extensive participation of laymen in both civil and criminal proceedings may have contributed in some measure to the pragmatic and flexible character of modern Scandinavian law. (F.Hi.)

ROMAN-DUTCH LAW

Roman-Dutch law is the system of law produced by the fusion of early modern Dutch law, chiefly of Germanic origin, and Roman, or civil, law. It existed in the Netherlands province of Holland from the 15th to the early 19th century and was carried by Dutch colonists to the Cape of Good Hope, where it became the foundation of modern South African law. It also influenced the legal systems of other countries that had once been Dutch colonies, such as Sri Lanka (formerly Ceylon) and Guyana.

Today Roman-Dutch law is in force throughout the Republic of South Africa and South West Africa/Namibia, and in Lesotho, Swaziland, Botswana, and Zimbabwe. In Sri Lanka it is present to a lesser degree, and in Guyana was from 1917 largely superseded by the common law of England. Reservation is made in favour of indigenous law and custom, so far as these are recognized; moreover the general law of these countries has in many respects departed from its original type.

Development of Roman-Dutch law in the Netherlands. In the 15th and 16th centuries the Roman law was "received" in the province of Holland (as it was sooner or later in the Netherlands generally), although general and local customs held their ground. These were based ultimately on Germanic tribal law—Frankish, Frisian, Saxon—supplemented by privileges and by-laws (*keuren*) and were themselves affected by an earlier infiltration of Roman law. The resulting mixed system, for which Simon van Leeuwen in 1652 invented the term "Roman-Dutch law," remained in force in the Netherlands until it was superseded in 1809 by the Code Napoléon, which in its turn in 1838 gave place to the Dutch civil code. The old law was also abrogated in the Dutch colonies. The Dutch civil code of 1838 has since been extensively revised.

There is, however, a third element in the Roman-Dutch system, namely the legislative acts of the Burgundian and Spanish periods, the most important of which were passed during the 16th century. Although a large quantity of legislation was later passed in the 17th and 18th centuries, it had little effect on the general character of the legal system. Roman-Dutch law can also be studied in collections of decided cases and of opinions (commonly termed *consultatien* or *advijsen*) and in the rich juristic literature of the system. The first attempt to reduce the Roman-Dutch civil law to a system was made by Hugo Grotius in his *Introduction to the Jurisprudence of Holland,* written while he was in prison in 1619–20 and published in 1631; this short treatise, a masterpiece of condensed exposition, remains a legal classic. Grotius' commentaries

The commentaries of Grotius

were followed by those of Johannes Voet and Simon van Groenewegen van der Made. Toward the end of the 18th century Dionysius Godefridus van der Keessel, professor at Leiden, lectured on the *jus hodiernum* ("law of today"), of which he published a summary in *Select Theses on the Laws of Holland and Zeeland* . . . (1800). The lectures, commonly known as the *Dictata,* still circulate as manuscript copies and have been cited in judgments by South African courts.

Survival and growth abroad of Roman-Dutch law. The law of the province of Holland was followed in the colonial empire, supplemented by local ordinances of the governors in council and, in the East Indies, by laws of the governors-general established at Batavia in Java (now Jakarta, Indon.). The ultimate legislative authority in the colonies was vested in the states general.

After the colonies passed to the British crown the old law underwent profound modifications, owing partly to changed social and economic conditions and partly to the incursion of rules and institutions derived from English common law.

The influence of English law (which was operative even during the period of the republics of the Transvaal and Orange Free State) has been most marked in criminal law and procedure, civil procedure, evidence, constitutional law, and, particularly, the commercial field of companies, bills of exchange, maritime law, and insurance. The law of tort or delict has also been considerably affected by English doctrines. On the other hand, the laws relating to property, persons, succession, and, to a lesser extent, contract still preserve their predominantly Roman-Dutch character. It is, for example, settled in both South Africa and Sri Lanka that "consideration" is not necessary for the validity of a contract.

The South Africa Act (1909) provided for the continuance of all laws in force in the several colonies at the establishment of the union until repealed by the Union Parliament or by the provincial councils within the sphere assigned to them. But thereafter, the Union Parliament and the appellate division of the Supreme Court of South Africa were active in consolidating, amending, and explaining the law and in making it more uniform. Many rules of the old law were pronounced obsolete by reason of disuse.

South African law

Modern South African law is a mixture of Roman-Dutch and English law. Constitutional law and administrative law have developed along English lines. The law of procedure and evidence is almost wholly English, as is most law relating to business associations and such areas as patents, trademarks, copyright, insurance, and maritime operations. On the other hand, criminal law is a combination of elements from Roman-Dutch and English common-law sources. In the law of succession, the rules governing the making of wills are English, whereas the substantive law of testamentary and intestate succession is largely Roman-Dutch. The law of persons and the law of property are almost purely Roman-Dutch, and the principles of the law of contract and of the law of delict are Roman-Dutch, only mildly influenced by common law.

(R.W.L./D.V.Cn.)

SCOTS LAW

At the union of the parliaments of England and Scotland in 1707, the legal systems of the two countries were very dissimilar. Scotland, mainly in the preceding century, had adopted as a guide much of the Roman law that had been developed by the jurists of Holland and France. But it is a fallacy to suppose that the law of Scotland is founded on the law of Rome: the Scots only turned to Roman, or civil, law when there was a gap in their own common or customary law. There is, however, a considerable infusion of civil law, not least in legal nomenclature and in the emphasis on principle rather than precedent. Perhaps the most important distinction is that Scotland, unlike England, did not separate the administration of equity and law. The Scottish conception of equity differs from the English system, which is parallel to the common law. The Scottish conception instead consists of a few fairly simple rules aimed at supplementing the law in order to prevent hardship. It also relegates certain remedies to the class of equitable remedies, of which the court has a large discretion to grant or withhold. The word equity in the law of Scotland has always retained its original meaning. The Scottish outlook upon this whole topic places Scots law clearly alongside the continental civil law and not the English system.

Scots notion of equity

Historical development of Scots law. The period following the union has been characterized by the merging of Scots and English law. One main cause of the merger is that much of the existing law of Scotland depends on statutes applicable to both countries. The House of Lords, consisting in its legal aspect until 1876 exclusively of English lawyers acting as the supreme court of appeal from Scotland, had a tendency to apply English law in Scottish appeals, and, in some cases, it ignored the distinction between its legislative and judicial functions. Another reason for the merging of systems is the influence of Scottish legal text writers, some of whom have tended to treat English law as though it were the law of their own country. The citation of English authorities in court has also had considerable effect. Not surprisingly the most complete merger of the systems has occurred in the field of mercantile law. In other fields the systems are still widely separated.

Courts of law. The system of Scottish courts is completely different from that of the English and again is closer to the continental pattern. The supreme Scottish court (the House of Lords not being a native court) is the Court of Session, instituted by King James V in 1532, probably upon a French model. The court has two main functions. It has original jurisdiction in a very wide range of cases, which is exclusive in a few matters; in its appellate capacity it hears appeals (by reclaiming petition) from the nine Court of Session courts of first instance (called compendiously the Outer House), each presided over by a lord ordinary, and also from the sheriff courts. The appellate court (Inner House) sits in two divisions, the first and second, presided over, respectively, by the lord president of the Court of Session and the lord justice clerk. All the judges have the courtesy title of "lord" but are not on that account peers.

While the judges of the Court of Session are traditionally judges of both fact and law, in the early 19th century the civil jury was introduced, less because it was wanted in Scotland than because the House of Lords was weary of the great number of appeals it had to hear. Because the decision of a jury cannot in the ordinary sense be appealed, the House of Lords determined that caseloads would be drastically reduced by the change. From the Inner House appeal lies in many cases to the House of Lords by right and not, as in England, by leave. The right of audience in the Court of Session is possessed exclusively by members of the Faculty of Advocates (the Scottish Bar).

The lower civil courts are the justice of the peace courts and the sheriff courts, which are ancient courts distributed by counties, having administrative as well as judicial functions. In civil matters the courts consist of an appellate judge only, who hears appeals from the sheriff of his jurisdiction. The position of sheriff is of great importance, and, except in a few matters, his jurisdiction almost exactly coincides with that of the Court of Session. As a local and comparatively inexpensive court the sheriff court is more popular than the Court of Session. The sheriff court, which has also a wide criminal jurisdiction, cannot be compared with the English County Court, with its very limited civil jurisdiction and complete lack of criminal jurisdiction. The justice of the peace court has a limited jurisdiction in small debt actions. The dean of Guild Court has a quaestorial jurisdiction in questions of building in the towns.

The Court of Session has absorbed the functions of certain ancient courts—the Court of Exchequer, the Admiralty Court, the Teind (or Tithe) Court, and the Commissary Court—which dealt with questions of marriage law and executry, while the judges have by statute been given separate duties in a Lands Valuation Appeal Court, a Registration Appeal Court, and an Election Petition Court.

The Scottish Land Court, established in 1911, has jurisdiction in a wide range of matters relating to agriculture.

Disputes between landlords and tenants of agricultural holdings may be brought before it by judicial process or, by agreement of the parties, in lieu of arbitration. It also deals with questions referred to it by the secretary of state for Scotland. (A.D.G./J.I.S.)

Socialist law systems

Socialist law derives from the system of public order devised by early Soviet leaders after the Russian Revolution in 1917. The Soviet model has subsequently been emulated in those parts of the world where Communist parties have become dominant, ranging from eastern Europe to Central and Southeast Asia and from the Caribbean to some parts of Africa. Socialist law has taken somewhat different forms in these countries because of differences in their prerevolutionary legal systems, which in the newer Socialist states were typically an amalgam of religious, customary, and received or imposed civil law. The immediate legal background of Socialist law in the Soviet Union and the eastern European Socialist countries was the Romano-Germanic system. For this reason and because Soviet law itself was much influenced by the civil-law tradition, some legal scholars classify Socialist law as a subdivision of civil law. They point out that while the rules of substantive law in Socialist countries have been influenced in varying degrees by the principles of Marxism and Leninism, the rules of civil and criminal procedure, the conceptual apparatus of the law, and the legal methodology are still essentially civilian. Guyana and Tanzania are notable exceptions because they are the only Socialist countries to have had a common-law background.

Other scholars, including Socialist jurists, claim an independent status for Socialist law. In their view, Socialist systems are distinguished from other legal systems by the influence of state ownership of the principal means of production, the special role that the Communist Party plays in the legal system, the close relationship between the legal system and national economic planning, the denial of any distinction between public and private law, and, perhaps above all, by a conception of the role of law as an instrument for restructuring society, shaping the new order, and dismantling the old. The claim that Socialist law is a separate legal family was recognized in practice after World War II when judges representing the Soviet legal system were elected to the International Court of Justice in accordance with requirements of the International Court's statute (article 9) that the court must include judges who are representative of the main forms of civilization and of the principal legal systems.

THE HISTORICAL DEVELOPMENT OF SOVIET LAW

Since Karl Marx and Friedrich Engels never prescribed a particular system of law for a state organized to achieve the aims set forth in their *Communist Manifesto* of 1848, it was left for V.I. Lenin and his colleagues in the Russian Communist Party after the Bolshevik Revolution of 1917 to improvise a legal system. General guidelines provided by Marx and Engels stated that law should be regarded as an instrument of the state, not as a limitation upon those who make policy, and that it should enunciate rules of public order to facilitate the transition to Socialism and ultimately to Communism. It was to have two major tasks: (1) elimination of the political power of the bourgeoisie (the property-owning middle class) by depriving them of their ownership of productive resources; and (2) education of citizens in the disciplined pattern of life claimed to be requisite for the achievement of the social order they desired. Completion of these two tasks was expected to assure both the abundant production necessary to realize the aim of distribution according to need and also the self-discipline of citizens necessary to eliminate coercion as the method of preserving order. The police, the army, and courts would become unnecessary, and law would, in Marxist terminology, "wither away." Citizens would perform social obligations, expressed in morals and unsanctioned administrative regulations, because they believed them to be desirable, not because they feared punishment for violating them.

The new principles were embodied in the first constitution of the Russian Soviet Federated Socialist Republic (R.S.F.S.R.), promulgated on July 10, 1918 (article 9), as follows:

> The basic task placed during the present transitional moment on the constitution of the R.S.F.S.R. is the establishment of the dictatorship of the city and village proletariat and of the poorest peasantry in the form of a powerful all-Russian Soviet authority with the objective of complete suppression of the *bourgeoisie,* the exploitation of man by man and the installation of socialism, under which there will be neither division into classes nor a State authority.

Early implementation. Lacking any precise pattern for a legal system designed to achieve the purposes stated in the first constitution, the new government issued only a few decrees designed to establish a framework for the new society and then set up a primitive institutional structure to enforce them. The decrees deprived private individuals of the ownership of land, banks, insurance companies, merchant fleets, and large-scale industry as an implementation of the policy of expropriation of the bourgeoisie. They also created restrictions on the employers of labour and secularized marriage and divorce. The enforcement instruments, called "people's courts," operated without benefit of professional prosecutors or a bar.

"Revolutionary conscience" as a guide

To permit regulation of the multitudinous social relationships for which no new law was prescribed, judges were directed at the start to apply Russian imperial laws, but only to the extent that they had not been revoked by the Revolution and were not contrary to the revolutionary conscience of the judges. Perhaps because Lenin and his colleagues were too busy to do otherwise, or perhaps because they preferred to let develop in practice a new legal order, no all-inclusive systematized body of law was prescribed until 1922. Local judges were guided during the first years, apart from the basic decrees indicated, only by suggestions issuing from the people's commissar of justice in the form of instructions, procedural manuals, and journal articles praising some court decisions as the proper application of revolutionary conscience and denouncing others as improper.

Political enemies of the Communists were not brought before the courts but were condemned by political bodies created at the same time as the people's courts and called revolutionary tribunals, or they were arrested and imprisoned without public hearings by the political police (Cheka). Lenin and his jurists professed that the treatment of opponents outside the regular court system was only a temporary measure. Nevertheless, the conduct of revolutionary tribunals and of the Cheka greatly influenced the course of Soviet law long after their formal abolition in 1922, for they stimulated a lack of respect for formality and strict adherence to law even among those charged with enforcing and administering the law.

After 1924 new agencies emerged within the People's Commissariat of Internal Affairs (NKVD), now the Committee for State Security (KGB), which enabled Lenin's successor, Joseph Stalin, to rid himself of opponents. Following Stalin's death in March 1953, the Communist Party's first secretary, Nikita S. Khrushchev, disclosed that many innocent persons had been convicted and sentenced to long prison terms and even to death by special boards of the secret police. The Soviet method of preserving order was revealed officially to have been dual, with courts on one side dealing publicly with nonpolitical offenses and social disputes, while, at the same time, administrative boards and the security police were secretly punishing Stalin's enemies.

Dual system

Codification. The New Economic Policy of 1921 reintroduced a strictly controlled private enterprise system in limited areas of the economy to speed reconstruction and to overcome the devastating effects of world and civil war. This reintroduction necessitated, in the Communist view, a stabilization of law to induce capitalists to invest in the country. Lenin's commissar of justice, Dmitry I. Kurskii, considered these measures retrogressive but necessary, while Nikolay V. Krylenko, later to become federal commissar of justice, called them the natural evolution of five years of Soviet law. Stabilization efforts resulted in the first systematic codification of Socialist law and led to the construction of a complex institutional framework of courts.

On Oct. 31, 1922, a judiciary act established within the Russian republic a three-tiered system of general courts with civil and criminal jurisdiction:

1. The name "people's court" was retained for the lowest level, but the institution was less primitive than that of 1917. Its bench comprised a full-time judge appointed annually and two lay judges selected from a panel of intelligent and politically trustworthy citizens, each serving for a few days. Lay judges shared responsibility with the professional judge in deciding issues of law and fact.

2. Appeals from this court were handled by a new provincial court that evolved from the Congress of People's Judges, which, under the prior rules, had gathered periodically in each province to survey the work of the local courts. The provincial courts also had jurisdiction to try offenses against the security of the regime and other serious civil and criminal cases. As an appellate court the provincial court consisted of a bench of three professional judges. As a trial court it resembled the people's court, except that the lay judges were selected from a more experienced panel of politically sound citizens.

3. The Supreme Court of the republic was created to coordinate policy in all provinces. Evolving from a control department established earlier in the Commissariat of Justice, this court was authorized not only to hear appeals from cases tried in provincial courts but also to discipline lower courts, to issue rulings interpreting the codes, and to try cases of an unusually important nature.

To provide specialized treatment for crimes relating to military matters and to the disruption of transport, then deemed critical to the success of the regime, special courts that had evolved earlier were continued by the judiciary act and subordinated to the Supreme Court. Following experiments that used non-professionals to conduct prosecution and defense, an office of public prosecutor and a college of defenders, who performed the functions of a bar, were established to aid the judges.

The substantive law and procedure applied by the new courts were established by codes enacted in 1922 and 1923; included were criminal, civil, family, land, and labour codes as well as codes of criminal and civil procedure. The drafters of these codes essentially followed patterns similar to those of the Romanist states of the European continent. The codes survived, with amendments, until the 1960s and '70s.

Soviet federal system

Patterns of legal institutions and the substantive and procedural law enacted by the R.S.F.S.R. were copied with little variation by the Soviet-type republics that emerged in peripheral regions of the old Russian Empire. The only change caused by the federation of the Soviet republics into the Union of Soviet Socialist Republics on Dec. 30, 1922, was the establishment, as a coordinator of all practice, of a Supreme Court of the U.S.S.R. The first federal constitution, adopted provisionally on July 6, 1923, and permanently on Jan. 31, 1924, extended to the federal government the authority to enact general principles of law that would be followed by the republics in maintaining their codes. While this development might have caused a change in the Soviet legal system, it did not; the first federal judiciary act of Oct. 29, 1924, confirmed the system of courts that already existed in each of the republics, and no change was made in substantive or procedural law. The sole innovation was the placing of the military and transport courts under the control of the Supreme Court; thus was created a self-contained system of federal courts, with lower branches functioning throughout the republics but beyond the reach of republic officials. A sharp line was thus drawn between courts that dealt with matters vital to the security of the regime and those concerned with social disputes.

The pattern established for Soviet courts and codes of law in 1922 and 1923 continued until Stalin's death, although the second federal constitution of Dec. 5, 1936, transferred authority to the federal level to enact codes of law to replace those of the republics. Although a second federal judiciary act was issued on Aug. 16, 1938, new codes were not enacted. On Feb. 11, 1957, the constitution was amended to restore the original relationship between individual republics and the federation, namely that they enact their own codes but conform to the general principles established federally. The first of the new general principles were enacted for criminal law and procedure on Dec. 25, 1958, and for civil law and procedure on Dec. 8, 1961. Others later were enacted at intervals, and codes for the republics were subsequently revised. The major innovation of the 1958 enactments was the requirement that punishment be ordered only by a court in accordance with the rules of the procedural code. This provision was heralded by Soviet jurists as a means of preventing the return to extralegal procedures such as those exercised by the secret police during Stalin's dictatorship.

Applications to other Communist states. The application of the pattern of law that had evolved within the Soviet Union was first exported to other areas when Outer Mongolia declared the establishment of a "people's government" on July 11, 1921. Its provisions for courts and prosecutors were like those of the Soviet Union, and the subsequently adopted law on land use of Feb. 6, 1942, provided the same formula of state ownership of land allocated for use gratis and in perpetuity by the nomadic herdsmen for their privately owned cattle. A labour law of Feb. 14, 1941, followed the Soviet pattern, as did a social insurance law of June 22, 1942.

The extension of Soviet law

The major opportunity to install the Soviet legal system outside the Soviet Union came only after World War II, when Soviet-type governments were created in the European states that were occupied by the Soviet army. The system was also established in China, North Korea, and North Vietnam when military victories placed the Communists in power. In each country there was some variation, related in part to the degree of national economic development, from the Soviet pattern evolved in the Soviet Union.

Throughout all of the Communist states, law was given the functions previously established in the Soviet Union. Thus, the first constitution of the Hungarian People's Republic of Aug. 20, 1949, declared (article 41):

> Courts of the Hungarian People's Republic punish enemies of the working people, defend and secure the state, economic and social structure of the people's democracy, its offices and the rights of the toilers, educate the toilers in the spirit of observance of the rules of socialist intercourse.

Regulations governing the organization of the people's courts in the People's Republic of China, promulgated on Sept. 3, 1951, specified that courts were "to consolidate the people's democratic dictatorship, uphold the new democratic social order and safeguard the fruit of the people's revolution." The People's Republic of Bulgaria was even more outspoken than the Soviet Union in clarifying the function of law; in its first constitution, enacted on Dec. 4, 1947, it granted the right to citizens to create associations, but (article 87) only "if they are not directed against state and public order established by the present constitution," and then it stated:

> It is forbidden and will be punished by law to form organizations having as their purpose taking from the Bulgarian people the rights and freedoms won by the people's uprising of September 9, 1944, and guaranteed by the present constitution, or to limit these rights and freedoms, to place under threat the national independence and state sovereignty of the country or to propagate open or concealed fascist or anti-democratic ideology, or to facilitate imperialist aggression, and also to participate in these organizations.

The first constitution of the People's Republic of China, promulgated on Sept. 20, 1954, nearly five years after proclamation of the People's Republic on Oct. 1, 1949, fixed a system of courts that emulated the judicial system in the Soviet Union. There were to be lay assessors, full-time professional judges, Soviet-type prosecutors, and three levels of courts; judges were "independent and subject only to the law," as they had been declared to be in the U.S.S.R. constitution.

Judicial power in Chinese law

The Chinese government, among its first actions, had abrogated all legal codes of the previous government. A commission to draft new codes was established in 1950, but little major legislation was enacted in China until the 1970s. In the interim, the courts were guided by governmental decrees, a few basic statutes, and the program of the Communist Party.

The Socialist states in eastern Europe moved more rapidly toward codification than had China, and, in contrast, confirmed their prewar civil codes, although with modifications. Poland adopted a General Statute on Civil Law with 119 articles on July 18, 1950. Before the expulsion of the Yugoslav Communist Party from the Cominform on June 28, 1948, Yugoslavia, under the constitution of Jan. 31, 1946, had followed the Soviet legal pattern. Under Tito, Yugoslavia accepted the concept of a monopoly of power in the Communist Party and the necessity for state ownership of productive resources. Tito, however, refused to accept the dictates of the Soviet Union on every detail, especially on the centralization of industrial control and the speed of implementing collectivization. The Yugoslav government experimented with bringing workers into the managerial process of public corporations. The government also broadened the autonomy of local government councils. Yugoslav legal philosophers rejected the view that the state could not begin to wither away until full achievement of Communism; they wanted to start relaxing coercion immediately. Stalin's attempts to unseat Tito failed, and after Stalin's death his successors attempted a reconciliation with Tito by indicating an acceptance of the idea that there could be many roads to Socialism.

Changes after the death of Stalin

Post-Stalin developments. Stalin's death in 1953 and his subsequent vilification by the Communist Party of the Soviet Union at its 20th congress three years later opened a new era in the law of Communist states. The people's democracies sought to be relieved of Soviet tutelage. Uprisings in Hungary and Poland forced Soviet leaders to accept the evolution of divergent Socialist legal systems. Polish jurists, including Communists, demanded humanistic attitudes toward legal procedure, greater flexibility within the system, and an increased regard for the individual; other eastern European states followed their lead. In 1958 Soviet legal philosophers formally endorsed a new humanism and declared that it was always fundamental to Lenin's thought.

Only Chinese Communists resisted the trend toward strengthened legal procedures, denouncing Soviet policies of humanism and legality in an open letter published in 1963. After 1957 Chinese leaders abandoned their policy of copying Soviet legal patterns and ceased their efforts to draft codes of law, declaring that they wanted flexibility in court. They evolved a legal system seemingly inspired by relics of traditional Chinese attitudes as modified by their own guerrilla experience before 1949 and by their study of Soviet practices during the period before 1921.

During the Cultural Revolution, which lasted from 1966 to 1976, the Chinese political leadership adopted the attitude that law was dispensable and an obstacle to political progress. In the post-Mao period, a major reevaluation of the role of law in a Socialist state took place, and in 1978 a long-range plan was adopted to enact legislation, reestablish the legal profession, improve court procedures, expand legal education, and make legal consciousness a part of the popular culture. In accordance with this program, China adopted the Constitution of 1982 as well as several important statutes, including election laws, a criminal code, codes of criminal and civil procedure, a marriage law, tax laws, and several laws on the organization of government, the courts, and the procuracy.

Concurrently with revival of concern for humanistic values, most Communist states revised their constitutions and law codes to give formal support to their claim that they had passed beyond capitalism and arrived at Socialism. New civil codes designed for Socialist economies were introduced in Poland and Czechoslovakia in 1964, the latter incorporating terminology that departed sharply from Romanist tradition.

Several features, however, remained unchanged in the law of these states, even after the introduction of variations. Most notable was the continuation of Lenin's principle that after the Russian Revolution no basis remained for the Romanist division of law into public and private spheres. To Lenin the concern of the Soviet state with every detail of social intercourse required the state to maintain the right and opportunity of intervention in any matter at any time. Consequently, he held that in the Soviet system all law must be public; that is, it must reflect the state's vital concern with the social relationships governed by the law. This principle, sometimes called *dirigisme* by French scholars, remains a characteristic of Socialist law.

CHARACTER OF SOVIET LAW

Montesquieu's concept of the separation of powers as embodied in the Constitution of the United States was rejected by the founders of the Soviet legal system. The legislature was made supreme over the executive and the judiciary. In principle only the enactments of the legislature were a source of law.

In reality, however, the executive became a source of law because the Council of Ministers often acted without convening the legislature. The second federal constitution of the Soviet Union, adopted in 1936, attempted to stop this practice, which violated the constitution of 1924, but the executive again soon usurped legislative authority. Post-Stalin reformers again sought to restore the monopoly of lawmaking power to the legislature. In practice, however, it was not the full assembly of the legislature but its derivative body, the presidium, elected from its membership to legislate during intervals between full sessions, that created the day-to-day changes in law. Although subsequent ratification from the full body was sought, it was extremely difficult in practice for the legislature to revoke presidium actions that had already gone into effect. The presidium thus became the most important source of law, except for proclamations of state economic plans and budgets and for the enactment of general principles for republic codes.

Legislative domination

Communist Party orders, except for certain decrees issued jointly by the party and a state agency, are not in the technical sense sources binding on courts. Still, the party provides the initiative for legislative action. Some vital laws, such as those reforming agriculture or industry, are given preeminent status if they are signed by the party secretary as well as by the chairman of the presidium when they are published as law.

Having embraced the principle of legislative supremacy and rejected the notion that judges can make law, the Soviet Union accepts no concept of judicial precedent like that of the Anglo-American common law. Nevertheless, the Supreme Court of the U.S.S.R., during the exercise of its authority to enunciate principles of interpretation, has proclaimed such far-reaching principles as to make of them de facto sources of law. Even individual decisions of the Supreme Court, although lacking the binding force of precedent, are given careful attention by lower courts, since they are published as guides. The Soviet approach to precedent is, in practice, generally similar to that of Romano-Germanic systems.

The Supreme Court of the U.S.S.R., being subordinate to the legislature, has no right to declare a law unconstitutional or to create any limitation upon the legislature or the executive on the ground that some higher or general principles of law have been violated. The legislature controls the judiciary at all levels above the people's court by exercising appointive and recall power through the soviet operating at the level of the court. Only people's judges are elected by the people. Elections follow the pattern established by Stalin for all representative bodies in the Soviet Union: the ballot consists of one candidate who is chosen by local professional groups with the guidance of the Communist Party members in the groups. In practice, judicial decisions have to generally agree with Communist Party policy or the judge risks recall.

The constitutional provision that "judges are independent and subject only to the law" (article 154) has to be understood as a means of denying local officials the right to intervene in the formation of an individual decision on personal grounds, but not as a means of prohibiting Communist Party intervention when such intervention is properly formulated in criticism of a line of decisions out of keeping with party wishes.

Judges are not entirely deprived of initiative by the rule of legislative supremacy or by the leading role played by the Communist Party. In both the civil and criminal codes adopted in 1922 there was authorization to depart from the rigid provisions of the code under special circumstances. The Civil Code opened with article 1, which reads, "Civil rights are protected by law except in cases where they are exercised contrary to their social and economic purpose." The exception permitted judges to apply their concept of state policy, when it was not specifically stated, to achieve an appropriate result in an individual situation, but the decision created no rule of law for future cases. The article was used frequently to justify judicial refusal to protect the rights of capitalists during the period of reconstruction (1922–28), but after 1930 it fell into disuse. Soviet jurists argued that it should be deleted from the code as being unrelated to a stable society from which all capitalist elements already had been eradicated, but it was retained in the civil-law general principles of 1961 (article 5).

Civil rights and state policy

The Criminal Code of 1922, revised in 1926, permitted the judge to exercise discretion in withholding a conviction despite the fact that a crime had clearly been committed, provided that the case posed no social danger. The judge also had the power to punish a citizen who had committed an "apparent" crime even though it was not defined as such by the code; the formula for such judicial initiative was that the act be found socially dangerous in a manner analogous to another act specifically defined as a crime. This provision, which led to unpredictability in the application of the law, was eliminated from the criminal law by the general principles adopted in December 1958.

SUBSTANTIVE LEGAL PROVISIONS

Soviet jurists and their colleagues in eastern Europe and other Communist states claim uniqueness for their system because of its constitutionally defined function of promoting Socialism. The end, in their view, transforms the means. While the novelty of the system is, therefore, primarily its philosophical base, the substantive law has been revised to introduce some novel institutions.

Property. The *Communist Manifesto*'s interpretation of property ownership as the source of political power was accepted by all Communist parties as the guiding principle to be followed in their revisions of the law. Consequently, private ownership of productive wealth was abolished everywhere, to some degree, as soon as state officials felt competent to administer production efficiently. The federal constitutions of the Soviet Union (1936 and 1977) set forth the ideal, denying to private individuals the right to employ labour in productive enterprise and reaffirming Lenin's decree of Feb. 19, 1918, which nationalized all land.

The ideal achievement of the goal was not realized everywhere. Polish and Yugoslav Communists continue to permit peasants to own farmlands and also allow private citizens to conduct service trades, although with limited numbers of employees. China alone permits the private ownership of factories to continue, but these are subject to the limitations that owners retain no financial link to foreign capitalists and that they share their ownership with the state as a partner.

The various legal codes reflect the preponderant importance given to state-owned productive property. For example, severer penalties are prescribed by the criminal code for the theft of state property than of private property. In the Soviet Union labour codes are concerned solely with state employment. The provisions of civil codes related to sales, contracts, property damage, and inheritance govern only consumer goods in traditional Romanist legal fashion, while new forms have been developed to govern transactions involving state property. Title to state property is subject to restrictions: no sales or leases are permitted, with limited exceptions for land allocated to peasant households that have been temporarily disrupted by military service or by relocation for industrial employment.

The ownership of land and personal property

Land has been declared state property only in the Soviet Union and the Mongolian People's Republic (constitution of 1924). In all Marxian Socialist states except Poland and Yugoslavia, land has increasingly been brought under the administration of cooperative-type associations structured on the model of the Soviet collective farm. Peasants have been induced, often under strong pressure, to assign to these associations the use of plots originally owned by or assigned to them. A variation introduced in China (1958) is called a "commune," and it unites both local government and land administration into one administrative unit that supervises "brigades" (groups of villages) and their subordinate "teams" (individual villages). By law, land ownership in China is not placed in the state as it is in the Soviet Union but in the "team" as a means of stimulating peasant initiative in land development.

The Soviet model collective farm places the ownership of machinery, tools, commonly used farm buildings, and crops in the cooperative; but member families are declared the owners of their family homestead, which also includes hand tools and barnyard animals. Small plots of state-owned land are assigned to each family for private use as a household garden. Farm government is carried out by a membership assembly that elects managers who are usually introduced to the members by central government authorities. Remuneration is in the form of a share of the produce that is graduated according to the value of work performed, computed by an accounting unit called a "labour day."

Agricultural cooperatives, generally regarded as transitional structures designed to guide peasants toward the superior type of organization patterned on industry, enable peasants to become farm employees and receive wages and social insurance benefits. Only China rejects this ultimate goal, preferring instead to perfect the "commune" as its model for rural productive units.

Nonproductive property and the tools of artisans are left to private ownership and are labeled "personal property" to distinguish them from income-producing "private property." Early efforts to equalize the ownership of consumer goods were abandoned by Stalin in 1930, and egalitarianism was denounced as a petty bourgeois utopian idea. The shortage of goods forced this change in attitude, and wages were adjusted to individual work performance as a means of stimulating production. This differentiation in wages led progressively to the weakening of restraints on inheritance and the ownership of luxury items.

Restrictions on personal property

The federal constitutions of the Soviet Union have entrenched the systems of property incentives by guaranteeing inheritance and the private ownership of savings accounts. To preclude reversion to capitalist practices, various restrictions are placed on the use of personal property. Thus, dwellings are limited in size to satisfy only family needs; sales of dwellings and apartments can occur only after long intervals; members of a family sharing a joint household may own only one dwelling and may lease excess space only temporarily and at rentals not exceeding those charged for occupancy of state-owned premises. Property purchased with income gained illegally can be expropriated. Neglected property can be expropriated if repairs are not made following a warning.

The determination to eliminate income resulting from the investment of private funds has been weakened by various paramount state pressures. Thus, interest is paid on state savings bank accounts to stimulate deposits as a means of accumulating funds available for state investment and to reduce the threat of an inflation of currency. Investment in state bonds also is encouraged, but returns are not paid in conventional interest but rather as lottery winnings. Inheritance from abroad is paid to heirs within the Soviet Union.

Marxian Socialist states outside the Soviet Union have established comparable restrictions, and Czechoslovakia has gone as far as requiring that the source of personal ownership be "honest"; *i.e.,* derived mainly from work for the benefit of society. China, under Mao Zedong, took the strict position, opposing extensive use of property incentives among individuals and espousing asceticism.

Communal ownership is reserved in Marxian Socialist states, except in China, for the distant future when abundance has been achieved, money is no longer used as a medium of exchange, and the state has "withered away," in accordance with the prophecy of Marx and Engels in the *Communist Manifesto.* Until that time, the state represents the community and therefore seeks to enlarge its segment of productive property.

Yugoslav Communists have rejected the Soviet model of state ownership of industry. Since 1950 they have used a system based on the "social ownership" of factories, under which the title to industrial property does not reside with the state but instead with the employees of each factory, who are organized into a "workers' council." Management is appointed not by a central ministry, as in other Marxian Socialist states, but by the workers' council, which acts in concert with the local government and an industry-wide chamber that represents all of the workers' councils in the republic in the same branch of industry. The primary aim of the Yugoslav system is to overcome the evils thought to be inherent in a centralized bureaucracy.

Economic planning. Production is planned in all Marxian Socialist states, although to varying degrees and through different types of institutions. Stalin favoured detailed planning and the centralized decisions of the Council of Ministers, who were advised by the State Planning Commission (Gosplan). The Yugoslav opposition to a centralized bureaucracy places an emphasis on planning by local government and workers' councils; these groups, however, are subject to influences exerted by the central government, which allocates major capital investment and establishes national goals.

After 1957 Stalin's successors experimented with forms of decentralized decision making; regional economic councils (Sovnarkhoz) were established for industries not directly related to national defense in the hope of stimulating managerial initiative. Most centralized industrial ministries were abolished from 1957 to 1965, but Gosplan and specialized financial and technical institutions retained control over the allocation of key resources. With the abolition of the Sovnarkhoz and the reconstitution of industrial ministries, in 1965 the Soviet Union returned to centralization—but in a form differing from Stalin's model. Under the new organization the responsibilities of the authorities of the republics were heightened, and plant managers were directed to search out local sources of supplies and markets.

The encouragement of local initiative went further in some Marxian Socialist states: Czechoslovak specialists even proposed to dramatize the change in emphasis by declaring that the public corporations administering state-owned property were the "owners" of the buildings and tools rather than only the administrators. This extreme experimentation was terminated by the Soviet intervention in 1968, although the Yugoslav Communists continue to rely on their system of local ownership and direction subject to indicative plans having no force of law.

Contracts as tools for economic planning

An interesting legal feature connected with production in all Marxian Socialist states is the system of contracts that are executed by plant managers for implementing production plans. These contracts prescribe quantity, design, technical specifications, price, and delivery date, and they use plan requirements as a basis for negotiation. After 1957 elements of consumer choice were also considered. These contracts are often used by permanent state tribunals, called State Arbitration, to settle disputes. The tribunals also may require a contract to be executed to implement the plan if the parties cannot agree on terms. China alone provides a variation by eschewing a permanent tribunal and opting for mediation through ad hoc committees established by local economic commissions that are linked to the planning apparatus.

Labour. While production and distribution in the Soviet Union are subject to compulsory orders, labour remains free to move as it wishes, although exceptions occurred during the emergency conditions of World War II. Although the Soviet labour supply is regulated by monetary and other inducements, it continues to be subject to the centralized determination of job classifications, hours, and wages.

The role of collective labour agreements has changed throughout the years as the central authorities increasingly have prescribed details of employment. The agreements fell into disuse in the mid-1930s but were restored after World War II to mobilize the labour force for production. Since job classifications, wages, and hours were established by law, the agreements concentrate on bettering local working conditions and on indicating the labour union's preferences regarding the use of bonuses that are collectively earned as incentives for exceeding planned production.

Strikes and grievances

Since production is planned and, therefore, is by definition rational, strikes are given no place in Marxian Socialist states. Remedies sought by management or labour unions against apparent injustices are supposed to take the form of petitions to higher authority for legislative or administrative rectification. Collective agreements are enforced primarily through moral suasion, but they are strengthened by provisions of the penal code that apply to state managers who violate the code, the collective agreement, or legal activity of labour unions, and to labour union officials who disrupt production so severely as to commit the crime of "wrecking." Work stoppages do occur in Socialist countries, however. Major strikes took place in Poland and Hungary in 1956 when Communist influence weakened, and reforms had to be instituted to enable the Communist Party to maintain power. In the 1980s, strikes in Poland and Yugoslavia became a significant de facto element of industrial relations.

The "right to work" has been established by all constitutions of Marxian Socialist states, but no system has implemented it by court orders that force an employment office to provide a job. The concept has come to require only the elimination of discrimination on the basis of race, sex, age, or domicile in filling such jobs as exist.

Individual employment contracts reflect the statutory provisions that govern employment and its termination. Since there is no individual bargaining in establishing the terms of employment, written contracts are often omitted in practice. A grievance procedure under labour union supervision is the established means by which aggrieved parties appeal to the courts for review of managerial decisions concerning payment and dismissal. Labour that becomes redundant with the advance of automation is to be retrained for other employment by state centres.

The "workers' self-government" of Yugoslavia presents a variation in the theory of employment relationships since workers can be said to employ themselves, being both owners and workers. Their activity is regarded as a social duty rather than a labour relationship, but in practice standards of employment are established to which management is required to adhere, these being generally the same as in other Marxian Socialist states.

Social insurance. Social insurance has been introduced in all Marxian Socialist states to protect employees who are injured on the job or elsewhere. No coverage was offered to collective farmers until 1966, or to housewives and children; the only means of recovery for these groups were suits based on the Civil Code's provisions on obligations. These rules conform to traditional Romanist legal models with some variation. Thus, the first Russian Civil Code (1922) required no fault for recovery, but this provision became conventional through court practice and later legislation. Fault is now necessary for liability, except in cases in which the instrument causing injury is classified as particularly hazardous. The list of such instruments includes automobiles.

The social utility of negligence suits

Social insurance benefits originally were inadequate because of the impoverishment of the state treasury. Consequently, victims of accidents were encouraged to sue for the difference between social insurance payments and total losses. Soviet policy makers later concluded that it was desirable to continue this practice even when social insurance funds were increased, the argument being that state managers would be prompted to exercise care because of the threat of suits, which would likely reduce profit margins and their eventual chances for promotion. Social insurance benefits thus continue to be set at levels lower than damages suffered.

Damages are generally limited to the loss of wages or anticipated wages for unemployed victims. Poland and Czechoslovakia also continue the ancient practice of allowing recovery for pain and suffering. No capitalization of damages is permitted, in part to avoid the transfer of large sums to irresponsible persons and in part to discourage an accumulation of wealth that might provoke capitalist speculation.

No personal liability insurance is made available to individuals in the Soviet Union, except to foreigners, in the belief that its availability would reduce the deterrent effect created by civil suits for accidents. Czechoslovak and Hungarian civil codes, however, retain pre-World War II provisions that created the opportunity of purchasing liability insurance. State insurance provides the state enterprises in all Marxian Socialist states with protection from suits by employees injured by hazardous equipment, unless charges of fault are made. The burden of proving absence of fault is on the defendant.

Liability for injury is attached to all state activities, even those not of an economic character, such as with ministries and local government, although until 1961 in the Soviet Union sovereign immunity for governmental functions had to be waived explicitly by statute to permit a suit. Thereafter, liability was extended to all types of state administration unless immunity was claimed by specific law. The Czechoslovak Civil Code (1964) established an approach to state liability by drawing no distinction between types of state activities in its rule of liability to suit.

The Chinese penchant for mediation over litigation developed after 1957 to eliminate possibilities of suits for injury in accidents, but in cases of reprehensible conduct the police were authorized to impose administrative sanctions for minor infringements of citizens' personal or property rights, or to require payment of medical expenses and property damage under the Security Administration Punishment Act of 1957.

A duty to protect state property (article 131) and to observe rules of Socialist conduct (article 130) was created by the second Soviet federal constitution (1936). It initiated judicial practice establishing an obligation of a state enterprise whose property had been protected to pay damages to a victim injured while attempting to perform his duty. This principle was written into the Fundamentals of Civil Law in 1961. No implementation was provided, however, for failure to rescue another in performance of the obligation to observe the rules of social conduct. Soviet authors concluded that this duty was moral and not legal, and violation could create no civil liability in favour of a citizen or his heirs if no rescue was attempted.

Artistic creation and invention. The legal rights that protect authors, artists, and inventors present problems in socialized economies unless such persons are employed by the state. Practice indicates that authors, artists, and inventors are most imaginative when they are unimpeded by state supervision. To provide the requisite freedom coupled with the stimulation of initiative, Soviet codifiers have created a status for innovators that stands midway between Western concepts and those of the Socialist ideal, in which wages are paid only for production. This status is referred to as "a law of personal relationships rather than of property relationships."

Inventors in the Soviet Union are offered two alternatives: an "author's certificate," or a patent. The first is granted after proof of novelty is submitted to the state patent office. It qualifies the inventor to receive payments computed on the savings that will accrue to the state from the use of the invention, with the computation being based on the best of the first five years of use. Additional benefits include tax exemption on a portion of the receipts, priority in the allocation of scarce living space, and acceptance into a technical school. The patent form is traditional, except that restrictions on private enterprise limit its use to assignment to state enterprises. In practice only foreigners choose patents, and their patent rights are protected under an international convention that was ratified by the Soviet Union in 1965.

Although other Marxian Socialist states in eastern Europe also offered a form similar to the "author's certificate," patents remained in common use until 1963, when the trend toward the Soviet model became marked. All states, however, retain the patent form for local inventions that are likely to have appeal abroad so as not to discourage potential foreign licensees.

Protection of artists' rights

Authors and artists are provided a more conventional protection than inventors, but they also are deprived of some of the traditional rights of a copyright owner. They cannot bargain for royalties or sell their work to foreign publishers except through a state trading enterprise. Royalties are paid in accordance with a state tariff that is based on the type and size of the edition. In addition to monetary rewards, which are payable for life and to heirs for a term of years, moral protection is offered in the form of the rights to control editorial changes and the use of the manuscript. Community interests are exerted against an individual author's will if the work is regarded as a national treasure. In such a case, it may be translated or published continuously after the first disclosure by the author. Royalties are paid, except when the language of translation is other than Russian—the reasoning behind this being that minority peoples need culture, but they may not be able to afford it. Nonmonetary inducements, including medals and honorary titles, are also used to stimulate innovation.

In 1973 the Soviet Union joined the Universal Copyright Convention, of which the United States and more than 70 other countries are members. Several other eastern European Marxian Socialist states are also members of the convention, and some belong to the Berne Convention.

Family law. Under the 1977 Soviet constitution, the family is placed under the protection of the state. This protection is manifested by the express constitutional obligation of the state to establish an extensive network of child-care institutions, to organize and improve services to families, and to provide allowances and benefits to families with children. Soviet family policy is closely allied with population and labour policy, aiming simultaneously at raising the birthrate while enabling mothers to remain in the labour force.

Equality of the sexes, a major aim of the Russian Revolution, is not understood as being inconsistent with the special protection offered in Socialist countries to women, especially to those with children. The 1977 Soviet constitution provides that women shall have equal rights with men, but that the state shall ensure the exercise of these rights by protecting the health of women; by providing material support for mothers, which includes paid maternity leaves and other benefits; and by aiding unmarried mothers.

Another goal of the Revolution in the area of family relationships was the secularization of marriage. Before 1926 Soviet family law presented little variation from western European models, but the 1926 code introduced unregistered marriage and divorce. Marriage could be recognized as legally constituted if cohabitation were proved orally. Divorce could be registered at the request of one party after the termination of marital relationships had been proved.

A trend toward the discouragement of promiscuity emerged in the mid-1930s after evidence linked increasing juvenile delinquency to the breakup of homes. A law enacted in 1936 required both parties to appear for divorce if the care and maintenance of children was an issue and sought to discourage repeated divorces by establishing a system of graduated fees for each suceeding divorce. The 1926 provisions, which accepted proof of marriage or divorce other than by registration, were revoked by law in 1944. Thereafter registration was compulsory to establish a marriage, and courts were given sole jurisdiction over divorce. Judges were required, however, to attempt reconciliation and to grant divorce only when they were convinced that the restoration of the family would be impossible.

The compromise on marriage and family

Hostility to the system grew during the succeeding years until a compromise position was formulated and adopted by law in 1968. While the registration of a marriage remains a requirement, divorce may be registered without prior court decision if both parties consent and if there are no offspring. In other circumstances court proceedings are compulsory.

While other Marxian Socialist states generally follow the Soviet model, all have introduced variations to preserve national cultural patterns. None but China has adopted the early attitudes espoused in the Soviet Union concerning unrestrained divorce, and even China discourages the use of its simple procedure of granting divorce at the request of either party. All Marxian Socialist states seek to preserve the conjugal family against gratuitous decisions to separate. The formula establishes no precise grounds for divorce; instead it authorizes divorce only in the event of a complete and permanent breakdown of the marital relationship. In practice marriages are considered to be broken if cohabitation has ceased. Children born out of wedlock generally are accorded the same rights as those born legitimate.

The marital property regime in the Soviet Union was established after early experimentation with individually owned property. The system that was finally chosen in 1926 incorporates the concept of community of acquests; that is, property owned individually by the spouses before marriage remains their separate property, but property acquired during marriage is co-owned, unless it was received by one spouse as a gift from outside or by inheritance. The Soviet model is that which is favoured by civil-law systems in general and has been adopted elsewhere among the Marxian Socialist states.

Criminal law. Novel definitions of crime have emerged in Marxian Socialist states to reflect Socialist principles. To the traditional crimes against the person and property, the Soviet Union has added "economic crimes." This development occurred in 1932 and coincided with the termination of the private enterprise system in favour of a system of state monopoly of large-scale production and state planning. Under these altered priorities, merchandising became a crime.

The constitutional prohibition of all private employment of labour for purposes of production is reflected in the Criminal Code, which penalizes employers for those practices. Even artisans employing no labour are forbidden to engage in various production activities unless their customers provide the raw materials. The leasing of unused space in small private dwellings at rates exceeding those established for tenants of state-owned apartment buildings is also a crime.

Crimes against the state

Fear of opposition to Communist Party leadership prompted the early introduction in the Soviet system of penalties for "counterrevolutionary" crimes. This phrase has been replaced in modern Socialist criminal codes by the term "crimes against the state." Although all legal systems punish attempts to subvert state authority, the definitions and practices under Stalin exceeded the norms of Western democracies in their vagueness of expression and in the severity of penalties they prescribed. After Stalin's death a decade ensued in which punishments for these crimes were lessened, but a new epoch was initiated in 1966 when authors were prosecuted for publishing allegorical manuscripts abroad that had been rejected by Soviet censors. The ground given for these prosecutions was that the authors should have anticipated the hostile use to which their writings would be put by enemies of the Soviet Union.

To implement Marxist attitudes that describe religion as an intolerable superstition to be discouraged, the conduct of religious schools and of social activities related to religious worship is prohibited in the Soviet Union. Strong measures were adopted during the 1920s when the religious hierarchies opposed Communist programs; although these were relaxed during World War II to gain the support of peasants, basic policies remained unchanged. An order in 1966 prescribed punishment for religious activities that are deemed harmful to health, and in practice the law is interpreted broadly in order to hamper worship by minority sects.

While other Marxian Socialist states have introduced less restrictive laws, their leaders have sought to discourage religion. Chinese Communists have expressed the belief that monks and priests should be reeducated. Eastern European leaders initially were militant in prosecuting clergy, but they have relaxed their vigilance as religious institutions have become more tractable. Poland is relatively liberal with respect to religion and religious practices, and Yugoslavia even went so far as to sign a concordat with the Vatican in 1966.

Procedural law. Romanist influence is strong in the Soviet codes of criminal and civil procedure, which resemble the French codes in their main features, except that they allow for more possibilities of control by the state. Soviet procedural law has in turn influenced all of the other Socialist systems.

Emphasis in criminal procedure is placed on investigation before trial, as in French practice. The examiner has to establish more than the prima facie case required of the grand jury in the traditional common-law system. The suspect has to be given an opportunity to testify and to produce witnesses and evidence in his defense. The examiner is authorized to bring the case to trial only if he becomes convinced of the suspect's guilt, so that the public trial consists primarily of verifying the examiner's work rather than hearing the defendant's case for the first time. Contrast with the French model is provided by the fact that the examiner is not a magistrate but a subordinate to the office of prosecutor, and until 1958 in the Soviet Union the defendant was not permitted to have an attorney during the preliminary investigation. Even now, only juveniles and certain special classes of defendants, such as deaf and mute persons, are given an absolute right to an attorney.

The "inquisitorial" system

Trial procedure follows the French model closely in that the judges are instructed to establish their personal conviction of guilt. To do so they may go beyond the evidence that is presented by the prosecution and defense. The court may call witnesses, seek its own experts, examine material evidence, and visit places connected with the crime. In practice, the judges do most of the interrogating in the courtroom and rely on the prosecutor and the defense only to bring out points that may have been overlooked. No rule of evidence of any kind binds the court, except for the requirements that evidence be relevant, probative, and not unduly repetitious.

Civil procedure requires the observance of rules of evidence only for certain formalities as established by the civil codes, such as for contracts or wills. Otherwise, the judges are as free in civil cases as they are in criminal matters to seek or hear what they feel they need to establish personal conviction. As in certain Romano-Germanic legal systems, the prosecutor has the authority to intervene in civil cases when intervention is deemed necessary to protect broader social interests.

Appellate courts that find violations of procedural requirements or of substantive law or that find evidence unconvincing are required to remand for a new trial. Revision of a sentence or of a decision without a retrial in a court in the original jurisdiction is permitted only when the original court had no jurisdiction, where an amnesty had been improperly applied, or where there had been no basis for trial. A penalty cannot be increased on the defendant's appeal.

Protest of decisions

A singular feature of Soviet procedure is the authority given to Supreme Court presidents and to the prosecutors of the republics and of the Soviet Union to protest a civil or criminal decision, after it has become final on appeal or in the absence of appeal. This procedure permits the reopening of convictions or even of acquittals if the highest legal authorities find grave errors in their periodic audits of inferior-court activities. Prosecutors frequently request such reopenings, even of convictions obtained by inferior prosecutors. This creates an opportunity of petition to superior prosecutors or to court presidents after all other remedies have been exhausted. At the same time it gives state authorities a second chance, not given to the convict or to the party of a civil suit, to change an otherwise final decision.

Socialist law in perspective. The industrialization that has been characteristic of the 20th century requires legal systems in all states to accept increasing state intervention in social relationships. Employers are no longer free to execute labour contracts as they wish; property owners are subject to increasing restraints on use (in some countries farms or dwellings may be lost if they are not used productively); nationalization or extensive regulation in banking, industry, and transport has become widespread; compulsory insurance against injury to third parties has become commonplace, as has social insurance against industrial accidents; and the secularization of marriage and divorce nearly everywhere prevents citizens from conducting their marital relationships solely within religious rules.

Socialist law, in reflecting the pervasive presence of the state in human affairs, appears to be an extreme extension of the trends common to all modern legal systems. Yet, with its sense of historic mission to advance society toward Socialism and ultimately to Communism, and with its educational ambition to help create the "new Socialist man" purged of selfish bourgeois propensities, Socialist law is closer to the religious legal systems of the world than it is to the civil or common law. Thus, despite its many common features with the Romano-Germanic systems, the conclusion seems justified that Socialist law differs in more than degree from other legal systems and therefore requires separate categorization. (J.N.H./M.A.Gl.)

BIBLIOGRAPHY

Roman law: W.W. BUCKLAND, *A Text-book of Roman Law from Augustus to Justinian,* 3rd ed. rev. (1975), is the standard reference in English. Other general introductions are R.W. LEAGE, *Roman Private Law Founded on the Institutes of Gaius and Justinian,* 3rd ed. edited by A.M. PRICHARD (1962); M. KASER, *Roman Private Law,* 3rd ed., trans. by ROLF DANNENBRING (1980; originally published in German, 10th rev. ed., 1977); BARRY NICHOLAS, *An Introduction to Roman Law* (1962, reprinted 1975); and J.A.C. THOMAS, *Textbook of Roman Law* (1976). For a scholarly treatment of Roman law in the classical era, see F. SCHULZ, *Classical Roman Law* (1951, reprinted 1954). For civil procedure, see M. KASER, *Das Römisches Zivilprozessrecht* (1966); and LEOPOLD WENGER, *Institutes of the Roman Law of Civil Procedure,* rev. ed. (1955; originally published in German, 1925). See also TONY HONORÉ, *Emperors and Lawyers* (1981).

For the historical development of Roman law, see the scholarly and highly readable accounts by H.F. JOLOWICZ and BARRY NICHOLAS, *Historical Introduction to the Study of Roman Law,* 3rd ed. (1972); and HANS JULIUS WOLFF, *Roman Law: An Historical Introduction* (1951, reprinted 1981). LEOPOLD WENGER, *Die Quellen des römischen Rechts* (1953), is a detailed historical study that relates Roman legal development to that of the surrounding legal systems. For the early history, see C.W. WESTRUP, *Introduction to Early Roman Law,* 5 vol. (1934–54). For a bibliographical appendix, see WOLFGANG KUNKEL, *An Introduction to Roman Legal and Constitutional History,* 2nd ed. (1973, reprinted 1975; originally published in German, 6th ed., 1971).

For the reception of Roman law in Europe, see PAUL KOSCHAKER, *Europa und das römische Recht,* 4th ed. (1966); and the briefer account by PAUL VINOGRADOFF, *Roman Law in Medieval Europe,* 2nd ed. (1929, reissued 1968). An interesting study may be found in W.W. BUCKLAND and ARNOLD D. MCNAIR, *Roman Law and Common Law: A Comparison in Outline,* 2nd ed. rev. by F.H. LAWSON (1952, reprinted 1965). For classified references to modern literature, see A. BERGER, *Encyclopedic Dictionary of Roman Law* (1953, reprinted 1980). Periodical literature from 1800 is elaborately indexed by L. CAES and R. HENRION, *Collectio Bibliographica Operum ad Ius Romanum Pertinentium* (1949–78), issued in two series with supplements. The volumes of *Iura* (irregular) contain bibliographies of current literature, including articles and reviews.

Germanic law: For sources, see *A General Survey of Events, Sources, Persons, and Movements in Continental Legal History* (1912, reprinted 1968); for an exhaustive survey, see KARL VON AMIRA, *Germanisches Recht,* 4th ed. edited by KARL AUGUST ECKHARDT vol. 1, *Rechtsdenkmäler* (1960); for discussion, see R. BUCHNER, *Die Rechtsquellen,* published as a supplement to W. WATTENBACH and W. LEVISON, *Deutschlands Geschichtsquellen im Mittelalter: Vorzeit und Karolinger,* vol. 2 (1953); and EDWARD JENKS, *Law and Politics in the Middle Ages: With a Synoptic Table of Sources,* 2nd ed. (1913, reprinted 1970). For substantive law, see HEINRICH BRUNNER, *Deutsche Rechtsgeschichte,* 2nd ed., vol. 1 (1906, reprinted 1961), the classic treatment; H. CONRAD, *Deutsche Rechtsgeschichte: Ein Lehrbuch,* vol. 1, *Frühzeit und Mittelalter,* 2nd ed. (1962, reprinted 1982); and C. VON SCHWERIN, *Grundzüge der deutschen Rechtsgeschichte,* 4th ed. prepared by HANS THIEME (1950). For very early law, see MARCO SCOVAZZI, *Le origini del diritto germanico: fonti, preistoria, diritto pubblico* (1957).

For Visigothic and Burgundian law, see E.A. THOMPSON, "The Barbarian Kingdoms in Gaul and Spain," *Nottingham Mediaeval Studies,* 7:3–33 (1963); and P.D. KING, *Law and Society in the Visigothic Kingdom* (1972). For Anglo-Saxon law, see F.W. MAITLAND, "The Laws of the Anglo-Saxons," *The Collected Papers of Frederic William Maitland,* ed. by H.A.L. FISHER, vol. 3, pp. 447–473 (1911, reprinted 1981); and H.G. RICHARDSON and G.O. SAYLES, *Law and Legislation from Aethelberht to Magna Carta* (1966). For north Germanic law, see L.B. ORFIELD, *The Growth of Scandinavian Law* (1953).

Modern comparative law: The world's legal systems are comparatively presented in KONRAD ZWEIGERT and HEIN KÖTZ, *An Introduction to Comparative Law,* 2 vol. (1977; originally published in German, 1969–71), which contains in its second volume comparative studies of selected topics of the law of contracts, torts, and restitution; and in RENÉ DAVID and JOHN E.C. BRIERLEY, *Major Legal Systems in the World Today: An Introduction to the Comparative Study of Law,* 3rd ed. rev. (1985; originally published in French, 1966). Both works contain extensive bibliographies. See also MARY ANN GLENDON, MICHAEL WALLACE GORDON, and CHRISTOPHER OSAKWE, *Comparative Legal Traditions: Text, Materials, and Cases on the Civil Law, Common Law, and Socialist Law Traditions, with Special Reference to French, West German, English, and Soviet Law* (1985). For a historical discussion explaining the origins, the essential values, and the unifying features of the Western legal heritage, see HAROLD J. BERMAN, *Law and Revolution: The Formation of the Western Legal Tradition* (1983). JOHN P. DAWSON, *The Oracles of the Law* (1968, reprinted 1978), is a penetrating analysis of the methods of legal thought in the Anglo-American, French, and German systems, as developed through the different roles played in them by judges and scholars. The *International Encyclopedia of Comparative Law* (1971–), is a compendium of articles by an international group of scholars that compares legal systems on a wide variety of topics. Current materials and articles are published in the *American Journal of Comparative Law* (quarterly); and the *International and Comparative Law Quarterly.* A continuing list of all books and articles published in English since 1790 on the subject of civil law can be found in CHARLES SZLADITS (comp.), *A Bibliography on Foreign and Comparative Law* (1953–), with supplements in the *American Journal of Comparative Law.*

Civil law: ARTHUR TAYLOR VON MEHREN and JAMES RUSSELL GORDLEY, *The Civil Law System,* 2nd ed. (1977), is a rich collection of cases and other source materials from France and Germany. A useful introduction for general readers as well as for those with legal training is JOHN HENRY MERRYMAN, *The Civil Law Tradition: An Introduction to the Legal Systems of Western Europe and Latin America,* 2nd ed. (1985). Helpful guides to the legal systems of particular countries are: F.H. LAWSON, A.E. ANTON, and L. NEVILLE BROWN (eds.), *Introduction to French Law,* 3rd ed. (1967); OTTO KAHN-FREUND, CLAUDINE LÉVY, and BERNHARD RUDDEN, *A Source Book on French Law: System, Methods, Outlines of Contract,* 2nd ed. (1979); F.H. LAWSON, *A Common Lawyer Looks at the Civil Law* (1955, reprinted 1977), primarily concerned with French law; E.J. COHN, *Manual of German Law,* 2nd rev. ed., 2 vol. (1968–71), a book written for lawyers; NORBERT HORN, HEIN KÖTZ, and HANSI G. LESER, *German Private and Commercial Law: An Introduction,* trans. from German (1982); MAURO CAPPELLETTI, JOHN HENRY MERRYMAN, and JOSEPH M. PERILLO, *The Italian Legal System: An Introduction* (1967); and ARTHUR TAYLOR VON MEHREN (ed.), *Law in Japan: The Legal Order in a Changing Society* (1963).

Common law: A.K.R. KIRALFY, *The English Legal System,* 7th ed. (1984); and PHILIP S. JAMES, *Introduction to English Law,* 11th ed. (1985), are general outlines. There are numerous general historical works, such as EDWARD JENKS, *A Short History of English Law: From the Earliest Times to the End of the Year 1939,* 6th ed. (1949); A.K.R. KIRALFY, *Potter's Historical Introduction to English Law and Its Institutions,* 4th ed. (1958); THEODORE F.T. PLUCKNETT, *A Concise History of the Common Law,* 5th ed. (1956); J.H. BAKER, *An Introduction to English Legal History,* 2nd ed. (1979); and S.F.C. MILSOM, *Historical Foundations of the Common Law,* 2nd ed. (1981), a difficult but classic text.

For the United States, see E. ALLAN FARNSWORTH, *An Introduction to the Legal System of the United States,* 2nd ed. (1983), a broad study of the American legal system. A good historical treatment is LAWRENCE M. FRIEDMAN, *A History of American Law,* 2nd ed. (1985).

General studies of other common-law nations include G.W. PATON (ed.), *The Commonwealth of Australia: The Development of Its Laws and Constitution* (1952); BORA LASKIN, *The British Tradition in Canadian Law* (1969); E. MCWHINNEY (ed.), *Canadian Jurisprudence: The Civil Law and Common Law in Canada* (1958); and M.C. SETALVAD, *The Common Law in India,* 2nd ed. (1970).

Socialist law: For a comprehensive survey of the Soviet legal system, see WILLIAM E. BUTLER, *Soviet Law* (1983). HAROLD J. BERMAN, *Justice in the U.S.S.R.: An Interpretation of Soviet Law,* rev. ed. (1963), is a study of Soviet law to the end of the Khrushchev era against the backdrop of Russian legal history and Marxist philosophy. The relationship of Soviet law to other

Socialist legal systems is treated in JOHN N. HAZARD, *Communists and Their Law: A Search for the Common Core of the Legal Systems of the Marxian Socialist States* (1969). The effect of sociopolitical change on Soviet legal developments is discussed in JOHN N. HAZARD, *Managing Change in the U.S.S.R.* (1983). JOHN N. HAZARD, WILLIAM E. BUTLER, and PETER B. MAGGS, *The Soviet Legal System: The Law in the 1980's* (1984), presents original source materials and contains an extensive bibliography. Current documentary materials are available in a looseleaf service, WILLIAM E. BUTLER, *Collected Legislation of the Union of Soviet Socialist Republics and the Constituent Union Republics* (1979–), major enactments from which are collected in WILLIAM E. BUTLER (comp.), *Basic Documents on the Soviet Legal System* (1983). The constitutions of Socialist legal systems are translated with introductions in WILLIAM B. SIMONS (ed.), *The Constitutions of the Communist World* (1980, reissued 1984).

(M.A.M./P.G.S./M.Rh./A.R.Ki./J.N.H./M.A.Gl.)

Lenin

Militant Marxist, founder of the Russian Communist Party (Bolsheviks), inspirer and leader of the Bolshevik Revolution, Lenin was the architect, builder, and first head of the Soviet state. He was the founder of the organization known as Comintern (Communist International) and posthumous source of "Leninism," the doctrine codified and conjoined with Marx's works by Lenin's successors to form Marxism-Leninism, the Communist world view. If the Bolshevik Revolution is—as some people have called it—the most significant political event of the 20th century, then Lenin must for good or ill be regarded as the century's most significant political leader. Not only in the Soviet Union but even among many non-Communist scholars, he is regarded as both the greatest revolutionary leader and revolutionary statesman in history, as well as the greatest revolutionary thinker since Marx.

Tass—Sovfoto

Lenin, 1918.

EARLY LIFE

The making of a revolutionary. It is difficult to find anything in his childhood that might have turned him onto the path of a professional revolutionary. Vladimir Ilich Ulyanov was born in Simbirsk (renamed Ulyanovsk) on April 22 (April 10, old style), 1870. (He adopted the pseudonym Lenin in 1901 during his clandestine party work after exile in Siberia.) He was the third of six children born into a close-knit, happy family of highly educated and cultured parents. His mother was the daughter of a physician, while his father, though the son of a serf, became a schoolteacher and rose to the position of inspector of schools. Lenin, intellectually gifted, physically strong, and reared in a warm, loving home, early displayed a voracious passion for learning. He was graduated from high school ranking first in his class. He distinguished himself in Latin and Greek and seemed destined for the life of a classical scholar. When he was 16, nothing in Lenin indicated a future rebel, still less a professional revolutionary—except, perhaps, his turn to atheism. But, despite an ideal upbringing, all five of the Ulyanov children who reached maturity joined the revolutionary movement. This was not uncommon in tsarist Russia, where even the highly educated and cultured intelligentsia were denied elementary civil and political rights.

In adolescence Lenin suffered two blows that unquestionably impelled him to take the path of revolution. First, his father was threatened shortly before his untimely death with premature retirement by a reactionary government grown fearful of the spread of public education. Second, in 1887 his beloved eldest brother, Aleksandr, a student at the University of St. Petersburg (now Leningrad State University), was hanged for conspiring with a revolutionary terrorist group that plotted to assassinate Emperor Alexander III. Suddenly, at age 17, Lenin became the male head of the family, which was now stigmatized as having reared a "state criminal."

Execution of Lenin's brother

Fortunately his mother's pension and inheritance kept the family in comfortable circumstances, despite the frequent imprisonment or exile of her children. Moreover, Lenin's high school principal (the father of Aleksandr Kerensky, who was later to lead the Provisional government deposed by Lenin in November [October, O.S.] 1917) did not turn his back on the "criminal's" family. He courageously wrote a character reference that smoothed Lenin's admission to a university.

In autumn 1887 Lenin enrolled in the faculty of law of the imperial Kazan University (now Kazan [V.I. Lenin] State University), but within three months he was expelled, accused of participating in an illegal student assembly. He was arrested and banished from Kazan to his grandfather's estate in the village of Kokushkino, where his older sister Anna had already been ordered by the police to reside. In the autumn of 1888, the authorities permitted him to return to Kazan but denied him readmission to the university. During his enforced idleness, he met exiled revolutionaries of the older generation and avidly read revolutionary political literature, especially Marx's *Das Kapital.* He became a Marxist in January 1889.

Formation of a revolutionary party. In May 1889 the Ulyanov family moved to Samara (now Kuybyshev). After much petitioning, Lenin was granted permission to take his law examinations. In November 1891 he passed his examinations, taking a first in all subjects, and was graduated with a first-class degree. After the police finally waived their political objections, Lenin was admitted to the bar and practiced law in Samara in 1892–93, his clients being mainly poor peasants and artisans. In his experience at law he acquired an intense loathing for the class bias of the legal system and a lifelong revulsion for lawyers, even those who claimed to be Social-Democrats.

But law was an extremely useful cover for a revolutionary activist. He moved to St. Petersburg (now Leningrad) in August 1893 and, while working as a public defender, made contact with revolutionary Marxists. In 1895 his comrades sent him abroad to make contact with Russian exiles in western Europe, especially with Russia's most commanding Marxist thinker, Georgy Plekhanov. Upon his return in 1895, Lenin and other Marxists, including L. Martov, the future leader of the Mensheviks, succeeded in unifying the Marxist groups of the capital in an organization known as the Union for the Struggle for the Liberation of the Working Class. The Union issued leaflets and proclamations on the workers' behalf, supported workers' strikes, and infiltrated workers' education classes to impart the rudiments of Marxism. In December 1895, the leaders of the Union were arrested. Lenin was jailed for 15 months, then exiled to Shushenskoye, in Siberia, for a term of three years. He was joined there in exile by his fiancée, Nadezhda Krupskaya, a Union member, whom he had met in the capital. They were married in Siberia, and she became Lenin's indispensable secretary and comrade. In exile they conducted clandestine party correspondence and collaborated (legally) on a Russian translation of Sidney and Beatrice Webb's *Industrial Democracy.*

Exile in Siberia

Upon completing his term of exile in January 1900, Lenin went abroad and was joined later by Krupskaya in Munich. His first major task abroad was to join Plekhanov, Martov, and three other editors in bringing out the newspaper *Iskra* ("The Spark"), which they hoped would unify the Russian Marxist groups scattered throughout Russia and western Europe into a cohesive Social-Democratic party.

Up to the point that Lenin joined the *Iskra,* his writings had focussed on three problems: first, he had written a number of leaflets that aimed to shake the workers' traditional veneration of the tsar by showing them that their harsh life was caused, in part, by the support tsarism rendered the capitalists; second, he attacked those self-styled Marxists who urged Social-Democrats and workers to concentrate on wage and hour issues, leaving the political struggle for the present to the bourgeoisie; third, and ultimately most important, he addressed himself to the peasant question.

The principal obstacle to the acceptance of Marxism by many of the intelligentsia was their adherence to the widespread belief of the Populists (Russian pre-Marxist radicals) that Marxism was inapplicable to peasant Russia, in which a proletariat was almost nonexistent. Russia, they held, was immune to capitalism, owing to joint ownership of peasant land by the village commune. This view had been first attacked by Plekhanov in the 1880s. Plekhanov had argued that Russia had already entered the capitalist stage, as evidenced by the rapid growth of industry. Despite the denials of the Populists, the man of the future in Russia was the proletarian, not the peasant. Applying the Marxist scheme of social development to Russia, Plekhanov had come to the view that the revolution in Russia must pass through two discrete stages: first, a bourgeois revolution that would establish a democratic republic and full-blown capitalism; and second, a proletarian revolution after mature capitalism had generated a numerous proletariat that had attained a high level of political organization, socialist consciousness, and culture, enabling them to usher in full Socialism.

It was this view that Lenin adhered to after he read Plekhanov's work in the late 1880s. But, almost immediately, Lenin went a step beyond his former mentor, especially in the peasant question. In an attack on the Populists published in 1894, Lenin charged that, even if they realized their fondest dream and divided all the land among the peasant communes, the result would not be Socialism but capitalism spawned by a free market in agricultural produce. The "Socialism" of the Populists would favour the growth of small-scale capitalism; hence the Populists were not Socialists but "petty bourgeois democrats." Outside of Marxism, which aimed ultimately to abolish the market system as well as the private ownership of the means of production, Lenin concluded, there could be no Socialism.

The peasant question

Even while in exile in Siberia, Lenin had begun research on his investigation of the peasant question, which culminated in his magisterial *Development of Capitalism in Russia* (published legally in 1899). In this work he argued that capitalism was rapidly destroying the peasant commune. The peasantry, for the Populists a homogeneous social class, was in actuality rapidly stratifying into a well-off rural bourgeoisie, a middling peasantry, and an impoverished rural "proletariat and semi-proletariat." In this last group, which comprised half the peasant population, Lenin found an ally for the extremely small industrial proletariat in Russia.

Iskra's success in recruiting Russian intellectuals to Marxism led Lenin and his comrades to believe that the time was ripe to found a revolutionary Marxist party that would weld together all the disparate Marxist groups at home and abroad. An abortive First Congress, held in 1898 in Minsk, had failed to achieve this objective, for most of the delegates were arrested shortly after the congress. The organizing committee of the Second Congress decided to convene the congress in Brussels in 1903, but police pressure forced it to transfer to London.

The congress sessions wore on for nearly three weeks, for no point appeared too trivial to debate. But the main issues quickly became plain: eligibility for membership and the character of party discipline; but, above all, the key issue focussed on the relation between the party and the proletariat, for whom the party claimed to speak.

In his *What Is To Be Done?* (1902), Lenin totally rejected the view that the proletariat was being driven spontaneously to revolutionary Socialism by capitalism and that the party should merely coordinate the struggle of the proletariat's diverse sections on a national and international scale. Capitalism, he contended, predisposed the workers to the acceptance of Socialism but did not spontaneously make them conscious Socialists. The proletariat by its own efforts in the everyday struggle against the capitalist could achieve "trade-union consciousness." But the proletariat could not by its own efforts grasp that it could win complete emancipation only by overthrowing capitalism and building Socialism, unless the party from without infused it with Socialist consciousness.

In his *What Is To Be Done?* and in his other works on party organization, Lenin created one of his most momentous political innovations, his theory of the party as the "vanguard of the proletariat." He saw the vanguard as a highly disciplined, centralized party that worked unremittingly to suffuse the proletariat with Socialist consciousness and served as mentor, leader, and guide, constantly showing the proletariat where its true class interests lie.

The "vanguard of the proletariat"

At the Second Congress the *Iskra* group split, and Lenin found himself in a minority on this very issue. Nevertheless, he continued to develop his view of "the party of a new type," which must be guided by "democratic centralism," or absolute party discipline. The party must be a highly centralized body organized around a small, ideologically homogeneous, hardened core of experienced professional revolutionaries, who were elected to the central committee by the party congress and who led a ramified hierarchy of lower party organizations that enjoyed the support and sympathy of the proletariat and all groups opposed to tsarism. "Give us an organization of revolutionaries," Lenin exclaimed, "and we will overturn Russia!"

Lenin spared no effort to build just this kind of party over the next 20 years, despite fierce attacks on his conception by some of his closest comrades of the *Iskra* days, Plekhanov, Martov, and Leon Trotsky. They charged that his scheme of party organization and discipline tended toward "Jacobinism," suppression of free intraparty discussion, a dictatorship *over* the proletariat, not *of* the proletariat, and, finally, establishment of a one-man dictatorship.

Lenin found himself in the minority in the early sessions of the Second Congress of what was now proclaimed to be the Russian Social-Democratic Workers' Party (RSDWP). But a walkout by a disgruntled group of Jewish Social-Democrats, the Bund, left Lenin with a slight majority. Consequently, the members of Lenin's adventitious majority were called Bolsheviks (majoritarians), and Martov's group were dubbed Mensheviks (minoritarians). The two groups fought each other ceaselessly within the same RSDWP and professed the same program until 1912, when Lenin made the split final at the Prague Conference of the Bolshevik Party.

The Bolsheviks and the Mensheviks

CHALLENGES OF THE REVOLUTION OF 1905 AND WORLD WAR I

The differences between Lenin and the Mensheviks became sharper in the Revolution of 1905 and its aftermath, when Lenin moved to a distinctly original view on two issues: class alignments in the revolution and the character of the post-revolutionary regime.

The outbreak of the revolution, in January 1905, found Lenin abroad in Switzerland, and he did not return to Russia until November. Immediately Lenin set down a novel strategy. Both wings of the RSDWP, Bolshevik and Menshevik, adhered to Plekhanov's view of the revolution in two stages: first, a bourgeois revolution; second, a proletarian revolution (see above). But the Mensheviks argued that the bourgeois revolution must be led by the bourgeoisie, with whom the proletariat must ally itself in order to make the democratic revolution. This would bring the liberal bourgeoisie to full power, whereupon the RSDWP would act as the party of opposition. Lenin defiantly rejected this kind of alliance and post-revolutionary regime. Hitherto he had spoken of the need for the proletariat to win "hegemony" in the democratic revolution. Now he flatly declared that the proletariat was the driving force of the revolution and that its only reliable ally was the peasantry. The bourgeoisie he branded as hopelessly counterrevolutionary and too cowardly to make its own revolution. Thus, unlike the Mensheviks, Lenin henceforth banked on an alliance that would establish a "revolutionary democratic dictatorship of the proletariat and the peasantry."

Nor would the revolution necessarily stop at the first stage, the bourgeois revolution. If the Russian revolution should inspire the western European proletariat to make the Socialist revolution, for which industrial Europe was ripe, the Russian revolution might well pass over directly to the second stage, the Socialist revolution. Then, the Russian proletariat, supported by the rural proletariat and semi-proletariat at home and assisted by the triumphant industrial proletariat of the West, which had established its "dictatorship of the proletariat," could cut short the life-span of Russian capitalism.

Years of discouragement

After the defeat of the Revolution of 1905, the issue between Lenin and the Mensheviks was more clearly drawn than ever, despite efforts at reunion. But, forced again into exile from 1907 to 1917, Lenin found serious challenges to his policies not only from the Mensheviks but within his own faction as well. The combination of repression and modest reform effected by the tsarist regime led to a decline of party membership. Disillusionment and despair in the chances of successful revolution swept the dwindled party ranks, rent by controversies over tactics and philosophy. Attempts to unite the Bolshevik and Menshevik factions came to naught, all breaking on Lenin's intransigent insistence that his conditions for reunification be adopted. As one Menshevik opponent described Lenin: "There is no other man who is absorbed by the revolution twenty-four hours a day, who has no other thoughts but the thought of revolution, and who even when he sleeps, dreams of nothing but revolution." Placing revolution above party unity, Lenin would accept no unity compromise if he thought it might delay, not accelerate, revolution.

Desperately fighting to maintain the cohesion of the Bolsheviks against internal differences and the Mensheviks' growing strength at home, Lenin convened the Bolshevik Party Conference at Prague, in 1912, which split the RSDWP forever. Lenin proclaimed that the Bolsheviks were the RSDWP and that the Mensheviks were schismatics. Thereafter, each faction maintained its separate central committee, party apparatus, and press.

When war broke out, in August 1914, Socialist parties throughout Europe rallied behind their governments despite the resolutions of prewar congresses of the Second International obliging them to resist or even overthrow their respective governments if they plunged their countries into an imperialist war.

Denunciation of pro-war Socialists

After Lenin recovered from his initial disbelief in this "betrayal" of the International, he proclaimed a policy whose audacity stunned his own Bolshevik comrades. He denounced the pro-war Socialists as "social-chauvinists" who had betrayed the international working-class cause by support of a war that was imperialist on both sides. He pronounced the Second International as dead and appealed for the creation of a new, Third International composed of genuinely revolutionary Socialist parties. More immediately, revolutionary Socialists must work to "transform the imperialist war into civil war." The real enemy of the worker was not the worker in the opposite trench but the capitalist at home. Workers and soldiers should therefore turn their guns on their rulers and destroy the system that had plunged them into imperialist carnage.

Lenin's policy found few advocates in Russia or elsewhere in the first months of the war. Indeed, in the first flush of patriotic fervour, not a few Bolsheviks supported the war effort. Lenin and his closest comrades were left an isolated band swimming against the current.

Lenin succeeded in reaching neutral Switzerland in September 1914, there joining a small group of anti-war Bolshevik and Menshevik émigrés. The war virtually cut them off from all contact with Russia and with like-minded Socialists in other countries. Nevertheless, in 1915 and 1916, anti-war Socialists in various countries managed to hold two anti-war conferences in Zimmerwald and Kienthal, Switzerland. Lenin failed at both meetings to persuade his comrades to adopt his slogan: "transform the imperialist war into civil war!" They adopted instead the more moderate formula: "An immediate peace without annexations or indemnities and the right of the peoples to self-determination." Lenin consequently found his party a minority within the group of anti-war Socialists, who, in turn, constituted a small minority of the international Socialist movement compared with the pro-war Socialists.

Lenin's theory of imperialism

Undaunted, Lenin continued to hammer home his views on the war, confident that eventually he would win decisive support. In his *Imperialism, the Highest Stage of Capitalism* (1917), he set out to explain, first, the real causes of the war; second, why Socialists had abandoned internationalism for patriotism and supported the war; and third, why revolution alone could bring about a just, democratic peace.

War erupted, he wrote, because of the insatiable, expansionist character of imperialism, itself a product of monopoly finance capitalism. At the end of the 19th century, a handful of banks had come to dominate the advanced countries, which, by 1914, had in their respective empires brought the rest of the world under their direct or indirect controls. Amassing vast quantities of "surplus" capital, the giant banks found they could garner superprofits on investments in colonies and semi-colonies, and this intensified the race for empire among the great powers. By 1914, dissatisfied with the way the world had been shared out, rival coalitions of imperialists launched the war to bring about a redivision of the world at the expense of the other coalition. The war was therefore imperialist in its origins and aims and deserved the condemnation of genuine Socialists.

Socialist Party and trade-union leaders had rallied to support their respective imperialist governments because they represented the "labour aristocracy," the better paid workers who received a small share of the colonial "superprofits" the imperialists proffered them. "Bribed" by the imperialists, the "labour aristocracy" took the side of their paymasters in the imperialist war and betrayed the most exploited workers at home and the super-exploited in the colonies. The imperialists, Lenin contended, driven by an annexationist dynamic, could not conclude a just, lasting peace. Future wars were inevitable so long as imperialism existed; imperialism was inevitable so long as capitalism existed; only the overthrow of capitalism everywhere could end the imperialist war and prevent such wars in the future. First published in Russia in 1917, *Imperialism* to this day provides the instrument that Communists everywhere employ to evaluate major trends in the non-Communist world.

LEADERSHIP IN THE RUSSIAN REVOLUTION

By 1917 it seemed to Lenin that the war would never end and that the prospect of revolution was rapidly receding. But in the week of March 8–15, the starving, freezing, war-weary workers and soldiers of Petrograd (until 1914, St. Petersburg) succeeded in deposing the Tsar. Lenin and his closest lieutenants hastened home after the German authorities agreed to permit their passage through Germany to neutral Sweden. Berlin hoped that the return of anti-war Socialists to Russia would undermine the Russian war effort.

First return to Petrograd. Lenin arrived in Petrograd

The Provisional Government

on April 16, 1917, one month after the Tsar had been forced to abdicate. Out of the revolution was born the Provisional Government, formed by a group of leaders of the bourgeois liberal parties. This government's accession to power was made possible only by the assent of the Petrograd Soviet, a council of workers' deputies elected in the factories of the capital. Similar soviets of workers' deputies sprang up in all the major cities and towns throughout the country, as did soviets of soldiers' deputies and of peasants' deputies. Although the Petrograd Soviet had been the sole political power recognized by the revolutionary workers and soldiers in March 1917, its leaders had hastily turned full power over to the Provisional Government. The Petrograd Soviet was headed by a majority composed of Menshevik and Socialist Revolutionary (SR), or peasant party, leaders who regarded the March (February, O.S.) Revolution as bourgeois; hence, they believed that the new regime should be headed by leaders of the bourgeois parties.

On his return to Russia, Lenin electrified his own comrades, most of whom accepted the authority of the Provisional Government. Lenin called this government, despite its democratic pretensions, thoroughly imperialist and undeserving of support by Socialists. It was incapable of satisfying the most profound desires of the workers, soldiers, and peasants for immediate peace and division of landed estates among the peasants.

Only a soviet government—that is, direct rule by workers, soldiers, and peasants—could fulfill these demands. Therefore, he raised the battle cry, "All power to the Soviets!"—although the Bolsheviks still constituted a minority within the soviets and despite the manifest unwillingness of the Menshevik–SR majority to exercise such power. This introduced what Lenin called the period of "dual power." Under the leadership of "opportunist" Socialists, the soviets, the real power, had relinquished power to the Provisional Government, the nominal power in the land. The Bolsheviks, Lenin exhorted, must persuade the workers, peasants, and soldiers, temporarily deceived by the "opportunists," to retrieve state power for the soviets from the Provisional Government. This would constitute a second revolution. But, so long as the government did not suppress the revolutionary parties, this revolution could be achieved peacefully, since the Provisional Government existed only by the sufferance of the soviets.

Initially, Lenin's fellow Bolsheviks thought that he was temporarily disoriented by the complexity of the situation; moderate Socialists thought him mad. It required several weeks of sedulous persuasion by Lenin before he won the Bolshevik Party Central Committee to his view. The April Party Conference endorsed his program: the party must withhold support from the Provisional Government and win a majority in the soviets in favour of soviet power. A soviet government, once established, should begin immediate negotiations for a general peace on all fronts. The soviets should forthwith confiscate landlords' estates without compensation, nationalize all land, and divide it among the peasants. And the government should establish tight controls over privately owned industry to the benefit of labour.

Rise of the Bolsheviks

From March to September 1917, the Bolsheviks remained a minority in the soviets. By autumn, however, the Provisional Government (since July headed by the moderate Socialist Aleksandr Kerensky, who was supported by the moderate Socialist leadership of the soviets) had lost popular support. Increasing war-weariness and the breakdown of the economy overtaxed the patience of the workers, peasants, and soldiers, who demanded immediate and fundamental change. Lenin capitalized on the growing disillusionment of the people with Kerensky's ability and willingness to complete the revolution. Kerensky, in turn, claimed that only a freely elected constituent assembly would have the power to decide Russia's political future—but that must await the return of order. Meanwhile, Lenin and the party demanded peace, land, and bread—immediately, without further delay. The Bolshevik line won increasing support among the workers, soldiers, and peasants. By September they voted in a Bolshevik majority in the Petrograd Soviet and in the soviets of the major cities and towns throughout the country.

Decision to seize power. Lenin, who had gone underground in July after he had been accused as a "German agent" by Kerensky's government, now decided that the time was ripe to seize power. The party must immediately begin preparations for an armed uprising to depose the Provisional Government and transfer state power to the soviets, now headed by a Bolshevik majority.

Lenin's decision to establish soviet power derived from his belief that the proletarian revolution must smash the existing state machinery and introduce a "dictatorship of the proletariat"; that is, direct rule by the armed workers and peasants which would eventually "wither away" into a non-coercive, classless, stateless, Communist society. He expounded this view most trenchantly in his brochure *The State and Revolution,* written while he was still in hiding. The brochure, though never completed and often dismissed as Lenin's most "Utopian" work, nevertheless served as Lenin's doctrinal springboard to power.

Until 1917 all revolutionary Socialists rightly believed, Lenin wrote, that a parliamentary republic could serve a Socialist system as well as a capitalist. But the Russian Revolution had brought forth something new, the soviets. Created by workers, soldiers, and peasants and excluding the propertied classes, the soviets infinitely surpassed the most democratic of parliaments in democracy, because parliaments everywhere virtually excluded workers and peasants. The choice before Russia in early September 1917, as Lenin saw it, was either a soviet republic—a dictatorship of the propertyless majority—or a parliamentary republic—as he saw it, a dictatorship of the propertied minority.

Lenin therefore raised the slogan, "All power to the Soviets!", even though he had willingly conceded in the spring of 1917 that revolutionary Russia was the "freest of all the belligerent countries." To Lenin, however, the Provisional Government was merely a "dictatorship of the bourgeoisie" that kept Russia in the imperialist war. What is more, it had turned openly counterrevolutionary in the month of July when it accused the Bolshevik leaders of treason.

From late September, Lenin, a fugitive in Finland, sent a stream of articles and letters to Petrograd feverishly exhorting the Party Central Committee to organize an armed uprising without delay. The opportune moment might be lost. But for nearly a month Lenin's forceful urgings from afar were unsuccessful. As in April, Lenin again found himself in the party minority. He resorted to a desperate stratagem.

Around October 20, Lenin, in disguise and at considerable personal risk, slipped into Petrograd and attended a secret meeting of the Bolshevik Central Committee held on the evening of October 23. Not until after a heated 10-hour debate did he finally win a majority in favour of preparing an armed takeover. Now steps to enlist the support of soldiers and sailors and to train the Red Guards, the Bolshevik-led workers' militia, for an armed takeover proceeded openly under the guise of self-defense of the Petrograd Soviet. But preparations moved haltingly, because serious opposition to the fateful decision persisted in the Central Committee. Enthusiastically in accord with Lenin on the timeliness of an armed uprising, Trotsky led its preparation from his strategic position as newly elected chairman of the Petrograd Soviet. Lenin, now hiding in Petrograd and fearful of further procrastination, desperately pressed the Central Committee to fix an early date for the uprising. On the evening of November 6, he wrote a letter to the members of the Central Committee exhorting them to proceed that very evening to arrest the members of the Provisional Government. To delay would be "fatal." The Second All-Russian Congress of Soviets, scheduled to convene the next evening, should be placed before a *fait accompli.*

Overthrow of the Provisional Government

On November 7 and 8, the Bolshevik-led Red Guards and revolutionary soldiers and sailors, meeting only slight resistance, deposed the Provisional Government and proclaimed that state power had passed into the hands of the Soviets. By this time the Bolsheviks, with their allies among the Left SR's (dissidents who broke with the pro-Kerensky SR leaders), constituted an absolute majority of the Second All-Russian Congress of Soviets. The delegates therefore voted overwhelmingly to accept full power and elected Lenin as chairman of the Council of People's Commissars, the new Soviet Government, and approved his Peace Decree and Land Decree. Overnight, Lenin had vaulted from his hideout as a fugitive to head the Revolutionary government of the largest country in the world. Since his youth he had spent his life building a party that would win such a victory, and now at the age of 47 he and his party had triumphed. "It makes one's head spin," he confessed. But power neither intoxicated nor frightened Lenin; it cleared his head. Soberly, he steered the Soviet government toward the consolidation of its power and negotiations for peace.

Saving the Revolution. In both spheres, Lenin was plagued by breaks within the ranks of Bolshevik leaders. He reluctantly agreed with the right-wingers that it would be desirable to include the Menshevik and Right SR parties in a coalition government—but on Lenin's terms. They must above all accept the soviet form of government, not a parliamentary one; they refused. Only the Left SR's agreed, and several were included in the Soviet government. Likewise, when the freely elected Constituent Assembly met in January 1918, the Mensheviks and Right SR majority flatly rejected sovietism. Lenin without hesitation ordered the dispersal of the Constituent Assembly.

The Allies refused to recognize the Soviet government; consequently it entered alone into peace negotiations with the Central Powers (Germany and her allies Austro-Hungary and Turkey) at the town of Brest-Litovsk. They imposed ruinous conditions that would strip away from Soviet Russia the western tier of non-Russian nations of the old Russian Empire. Left Communists fanatically opposed acceptance and preached a revolutionary war, even if it imperilled the Soviet government. Lenin insisted that the terms, however ruinous and humiliating, must be accepted or he would resign from the government. He sensed that peace was the deepest yearning of the people; in any case, the shattered army could not raise effective resistance to the invader. Finally, in March 1918, after a still larger part had been carved out of old Russia by the enemy, Lenin succeeded in winning the Central Committee's acceptance of the Treaty of Brest-Litovsk. At last Russia was at peace.

Treaty of Brest-Litovsk

But Brest-Litovsk only intensified the determination of counterrevolutionary forces and the Allies who supported them to bring about the overthrow of the Soviet government. That determination hardened when, in 1918, Lenin's government repudiated repayments of all foreign loans obtained by the tsarist and Provisional governments and nationalized foreign properties in Russia without compensation. From 1918 to 1920 Russia was torn by a Civil War, which cost millions of lives and untold destruction. One of the earliest victims was Lenin himself. In August 1918 an assassin fired two bullets into Lenin as he left a factory in which he had just delivered a speech. Because of his robust constitution, he recovered rapidly.

The Soviet government faced tremendous odds. The anti-Soviet forces, or Whites, headed mainly by former tsarist generals and admirals, fought desperately to overthrow the Red regime. Moreover, the Whites were lavishly supplied by the Allies with materiel, money, and support troops that secured White bases. Yet, the Whites failed.

It was largely because of Lenin's inspired leadership that the Soviet government managed to survive against such military odds. He caused the formation and guided the strategy of the Workers' and Peasants' Red Army, commanded by Trotsky. Although the economy had collapsed, he managed to mobilize sufficient resources to sustain the Red Army and the industrial workers. But above all it was his political leadership that saved the day for the Soviets. By proclaiming the right of the peoples to self-determination, including the right to secession, he won the active sympathy, or at least the benevolent neutrality, of the non-Russian nationalities within Russia, because the Whites did not recognize that right. Indeed, his perceptive, skillful policy on the national question enabled Soviet Russia to avoid total disintegration and to remain a huge multinational state. By making the industrial workers the new privileged class, favoured in the distribution of rations, housing, and political power, he retained the loyalty of the proletariat. His championing of the peasants' demand that they take all the land from the gentry, church, and crown without compensation won over the peasants, without whose support the government could not survive.

Because of the breakdown of the economy, however, Lenin adopted a policy toward the peasant that threatened to destroy the Soviet government. Lacking funds or goods to exchange against grain needed to feed the Red Army and the towns, Lenin instituted a system of requisitioning grain surpluses without compensation. Many peasants resisted—at least until they experienced White "liberation." On the territories that the Whites won, they restored landed property to the previous owners and savagely punished the peasants who had dared seize the land. Despite the peasants' detestation of the Soviet's grain requisitioning, the peasants, when forced to choose between Reds and Whites, chose the Reds.

The New Economic Policy

After the defeat of the Whites, the peasants no longer had to make that choice. They now totally refused to surrender their grain to the government. Threatened by mass peasant rebellion, Lenin called a retreat. In March 1921 the government introduced the New Economic Policy, which ended the system of grain requisitioning and permitted the peasant to sell his harvest on an open market. This constituted a partial retreat to capitalism.

From the moment Lenin came to power, his abiding aims in international relations were twofold: to prevent the formation of an imperialist united front against Soviet Russia; but, even more important, to stimulate proletarian revolutions abroad.

In his first aim he largely succeeded. In 1924, shortly after his death, Soviet Russia had won de jure recognition of all the major world powers except the United States. But his greater hope of the formation of a world republic of soviets failed to materialize, and Soviet Russia was left isolated in hostile capitalist encirclement.

Formation of the Third International. To break this encirclement, he had called on revolutionaries to form Communist parties that would emulate the example of the Bolshevik Revolution in all countries. Dramatizing his break with the reformist Second International, in 1918 he had changed the name of the RSDWP to the Russian Communist Party (Bolsheviks), and in March 1919 he founded the Communist, or Third, International. This International accepted the affiliation only of parties that accepted its decisions as binding, imposed iron discipline, and made a clean break with the Second International. In sum, Lenin now held up the Russian Communist Party, the only party that had made a successful revolution, as the model for Communist parties in all countries. One result of this policy was to engender a split in the world labour movement between the adherents of the two internationals.

The Communist International scored its greatest success in the colonial world. By championing the rights of the peoples in the colonies and semi-colonies to self-determination and independence, the International won considerable sympathy for Communism. Lenin's policy in this question still reverberates through the world today. And it offers another example of Lenin's unique ability to find allies where revolutionaries had not found them before. By taking the side of the national liberation movements, Lenin could claim that the overwhelming majority of the world's population, then living under imperialist rule, as well as the European proletariat, were the natural allies of the Bolshevik Revolution.

Thus Lenin's revolutionary genius was not confined to his ability to divide his enemies; more important was his skill in finding allies and friends for the exiguous proletariat of Russia. First, he won the Russian peasants to the side of the proletariat. Second, while he did not win the workers to make successful Communist revolutions in the West, they did compel their governments to curtail armed intervention against the Bolshevik Revolution. Third, while the Asian revolutions barely stirred in his lifetime, they did strengthen the Soviet Communists in the belief that they were not alone in a hostile world.

By 1921 Lenin's government had crushed all opposition parties on the grounds that they had opposed or failed to support sufficiently the Soviet cause in the Civil War. Now that peace had come, Lenin believed that their opposition was more dangerous than ever, since the peasantry and even a large section of the working class had become disaffected with the Soviet regime. To repress opponents of Bolshevism, Lenin demanded the harshest measures, including "show" trials and frequent resort to the death penalty. Moreover, he insisted on even tighter control over dissent within the party. Lenin's insistence on merciless destruction of the opposition to the Bolshevik dictatorship subsequently led many observers to conclude that Lenin, though personally opposed to one-man rule, nevertheless unwittingly cleared the way for the rise of Joseph Stalin's dictatorship.

Suppression of dissent

By 1922 Lenin had become keenly aware that degeneration of the Soviet system and party was the greatest danger to the cause of Socialism in Russia. He found the party and Soviet state apparatus hopelessly entangled in red tape and incompetence. Even the agency headed by Stalin that was responsible for streamlining administration was, in fact, less efficient than the rest of the government. The Soviets of Workers' and Peasants' Deputies had been drained of all power, which had flowed to the centre. Most disturbing was the Great Russian chauvinism that leading Bolsheviks manifested toward the non-Russian nationalities in the reorganization of the state in which Stalin was playing a key role. Moreover, in April 1922 Stalin won appointment as general secretary of the party, in which post he was rapidly concentrating immense power in his hands. Soviet Russia in Lenin's last years could not have been more remote from the picture of Socialism he had portrayed in *State and Revolution.* Lenin strained every nerve to reverse these trends, which he regarded as antithetical to Socialism, and to replace Stalin.

Illness and death. In the spring of 1922, however, Lenin fell seriously ill. In April his doctors extracted from his neck one of the bullets he had received from the assassin's gun in August 1918. He recovered rapidly from the operation, but a month later he fell ill, partially paralyzed and unable to speak. In June he made a partial recovery and threw himself into the formation of the Union of Soviet Socialist Republics, the federal system of reorganization he favoured against Stalin's unitary scheme. But in December he was again incapacitated by semi-paralysis. Although no longer the active leader of the state and party, he did muster the strength to dictate several prescient articles and what is called his political "Testament," dictated to his secretary between December 23, 1922, and January 4, 1923, in which he expressed a great fear for the stability of the party under the leadership of disparate, forceful personalities such as Stalin and Trotsky. On March 10, 1923, another stroke deprived him of speech. His political activity came to an end. He suffered yet another stroke on the morning of January 21, 1924, and died that evening at Gorky, near Moscow.

The last year of Lenin's political life, when he fought to eradicate abuses of his Socialist ideals and the corruption of power, may well have been his greatest. Whether the history of the Soviet Union would have been fundamentally different had he survived beyond his 54th birthday, no one can say with certainty. (Al.Re.)

MAJOR WORKS

BOOKS: *Chto Takoye "Druzya Naroda," kak oni voyuyut protiv Sotsial-Demokratov?* (hectographic editions from 1894, printed from 1920; *What the "Friends of the People" Are, and How They Fight the Social-Democrats,* 1946); *Razvitiye kapitalizma v Rossi* (1899; augmented ed., 1908; *The Development of Capitalism in Russia,* 1956); *Chto delat?* (1902; *What Is To Be Done?,* 1929, 1933, 1950, 1963); *Shag vperyod, dva shaga nazad* (1904; *One Step Forward, Two Steps Back,* 1941); *Dve taktiki Sotsial-Demokraty v demokraticheskoy revolyutsi* (1905; *Two Tactics of Social-Democracy in the Democratic Revolution,* 1935, 1947); *Agrarny vopros i "Kritiki Marksa"* (1908, first chapters already published in periodicals from 1901); *Agrarnaya programma Sotsial-Demokratyi v pervoy russkoy revolyutsi 1905–1907 godov* (1908, 1917; *The Agrarian Programme of Social-Democracy in the First Russian Revolution,* 1954); *Imperializm, kak noveyshy etap kapitalizma* (1917, later retitled *Imperializm, kak vysshaya stadiya kapitalizma; Imperialism: The Latest Stage in the Development of Capitalism,* 1924; and *Imperialism, the Highest Stage of Capitalism,* 1933, 1939, 1947); *Gosudarstvo i revolyutsiya* (1917; *The State and Revolution,* 1919); *Proletarskaya revolyutsiya i renegat Kautsky* (1918; *The Proletarian Revolution and Kautsky the Renegade,* 1920); *Detskaya bolezn "levizny" v kommunizme* (1920; *"Left Wing" Communism: An Infantile Disorder,* 1920, 1934, 1961).

MARXIST PERIODICALS EDITED OR CONTROLLED BY LENIN: *Iskra,* 51 numbers (1900–03); *Zarya,* 4 numbers (1901–02); *Vperyod,* weekly, 18 numbers (1905); *Novaya Zhizn,* daily, 28 numbers (1905); *Proletary,* weekly, 26 numbers (1905); *Volna,* daily, 25 numbers (1906); *Vperyod,* daily, 17 numbers (1906); *Ekho,* daily, 14 numbers (1906); *Proletary,* 50 numbers (1906–09); *Sotsial-Demokrat,* 58 numbers (1908–17; wholly Leninist from 1911); *Zvezda,* at first weekly, then more frequent, 69 numbers (1910–12); *Rabochaya Gazeta,* 9 numbers (1910–12); *Prosveshcheniye,* monthly, 31 numbers (1911–14; and 1 final number, 1917); *Pravda,* daily (1912–14 and from 1917 with numerous changes of name 1913–14 and July-October 1917); *Kommunist,* 1 double number (1915); *Sbornik Sotsial-Demokrata,* 2 numbers (1916).

JOURNALISM AND PARTY THESES: "O zadachakh proletariata v dannoy revolyutsi" (*The April Theses,* 1951), "O dvoevlasti," and "Uroki revolyutsi" (*Lessons of the Revolution,* 1918), all in *Pravda,* 1917; "Ocherednye zadachi Sovietskoy vlasti," *Pravda* (1918; *The Soviets at Work,* 1918; and *The Immediate Tasks of the Soviet Government,* 1951).

BIBLIOGRAPHY

Editions: The fifth and most complete collection of Lenin's works was published in Moscow in 55 vol. (1958–65; with index, 1966; reference volume in two parts, 1967–70). The *Collected Works,* 45 vol. (1960–70, reprinted 1980), is a Soviet English translation of the 4th Russian edition of Lenin's works, enriched by editorial notes from the 5th edition. *Selected Works,* 3 vol. (1970–71), includes most of the works mentioned in this article and many more. Other selections published in the Soviet Union include all the major works, as well as shorter writings, correspondence, and speeches, for the most part taken from the English translation of the 4th edition of the *Collected Works.* C. LEITEIZEN and J.S. ALLEN (eds.), *Lenin on the United States: Selected Writings* (1967, reprinted 1975), includes material from the period 1905–22. Western publications of Lenin's works in English are scarce. They include *The Essentials of Lenin,* 2 vol. (1947, reprinted 1973), which follows the Soviet edition; and ROBERT TUCKER (ed.), *The Lenin Anthology* (1975), with interpretive comments.

Biographical and critical studies: LOUIS FISCHER, *The Life of Lenin* (1964), a major, detailed work, enlivened by the author's own reminiscences of Moscow in Lenin's time; MOSHE LEWIN, *Lenin's Last Struggle* (1968, reprinted 1978; originally published in French, 1967), a brief but incisive account of Lenin's effort to mobilize party forces to curb Stalin and remove him from power; ALFRED MEYER, *Leninism* (1957, reissued 1972), analysis of Lenin's political philosophy; DAVID SHUB, *Lenin,* rev. ed. (1966, reprinted 1977), readable and informative; ADAM B. ULAM, *The Bolsheviks: The Intellectual and Political History of the Triumph of Communism in Russia* (1965, reprinted 1973), a learned political biography of Lenin; BERTRAM D. WOLFE, *Three Who Made a Revolution: A Biographical History,* 4th rev. ed. (1964, reprinted 1978), a pioneering work (biographies of Lenin, Trotsky, and Stalin) that demolishes many myths, and *The Bridge and the Abyss: The Troubled Friendship of Maxim Gorky and V.I. Lenin* (1967, reprinted 1983); NADEZHDA KRUPSKAYA, *Reminiscences of Lenin* (1970; originally published in Russian, 1924, 2nd ed., 1968), reticent, impersonal recollections by Lenin's widow; LEON TROTSKY, *Lenin: Notes for a Biographer* (1971; originally published in Russian, 1924), an appreciation of Lenin of the *Iskra* period and 1917–18, the periods of Trotsky's closest collaboration with Lenin, and *The Young Lenin* (1972, trans. from Russian); NIKOLAI VALENTINOV, *Encounters with Lenin* (1968; originally published in Russian, 1953), and *The Early Years of Lenin* (1969), revealing observations on Lenin's personality by a former associate; DIETRICH GEYER, *Lenin in der Russischen Sozialdemokratie* (1962), a scholarly study of Lenin and the origins of the Bolshevik–Menshevik split. Later studies include LEONARD SCHAPIRO and PETER REDDAWAY (eds.), *Lenin: The Man, the Theorist, the Leader: A Reappraisal* (1967), a collection of essays; E. VICTOR WOLFENSTEIN, *The Revolutionary Personality: Lenin, Trotsky, Gandhi* (1967, reprinted 1971); MICHAEL C. MORGAN, *Lenin* (1971, reprinted 1973), a sympathetic biography; TONY CLIFF, *Lenin,* 4 vol. (1975–79), a political biography, largely hagiographic; and GERDA WEBER and HERMANN WEBER, *Lenin: Life and Works* (1980; originally published in German, 1974).

Lenin is the subject of many historical studies, including HAROLD SHUKMAN, *Lenin and the Russian Revolution* (1967, reprinted 1977), and, with GEORGE KATKOV, *Lenin's Path to*

Power: Bolshevism and the Destiny of Russia (1971); HELMUT GRUBER, *International Communism in the Era of Lenin: A Documentary History* (1967, reissued 1972); BRANKO LAZITCH and MILORAD M. DRACHKOVITCH, *Lenin and the Comintern* (1972); ALFRED ROSMER, *Moscow Under Lenin* (1972; originally published in French, 1953), an insider's account of the period 1920–24, exploring the role of Lenin in the international movement; ALEXANDER RABINOWITCH, *The Bolsheviks Come to Power: The Revolution of 1917 in Petrograd* (1976); THOMAS H. RIGBY, *Lenin's Government: Sovnarkom; 1917–1922* (1979); and HÉLÈNE CARRÈRE D'ENCAUSSE, *Lenin: Revolution and Power* (1982), a study of economic, social, political, and ideological issues.

Interpretive studies include PAUL M. SWEEZY and HARRY MAGDOFF (eds.), *Lenin Today: Eight Essays on the Hundredth Anniversary of Lenin's Birth* (1970), an example of non-Soviet Marxist–Leninist thinking; GEORG LUKÁCS, *Lenin: A Study on the Unity of His Thought* (1971; originally published in German, 1924), an evaluation by a Hungarian Marxist philosopher; DAVID LANE, *Leninism: A Sociological Interpretation* (1981); ALAIN BESANÇON, *The Rise of the Gulag: Intellectual Origins of Leninism* (1981; originally published in French, 1977), a highly critical work aiming to prove the continuity between Leninism and Stalinism; and NEIL HARDING, *Lenin's Political Thought,* 2 vol. (1977–81, reprinted 1983). Official Soviet interpretation of Lenin's role is provided by BORIS N. PONOMAREV, *Lenin and the Revolutionary Process* (1980, trans. from Russian). NINA TUMARKIN, *Lenin Lives!: The Lenin Cult in Soviet Russia* (1983), explores the veneration of Lenin. For bibliography, see DAVID R. EGAN, MELINDA A. EGAN, and JULIE A. GENTHNER, *Lenin: An Annotated Bibliography of English-Language Sources to 1980* (1982).

Leningrad

The second largest city of the Soviet Union and one of the world's major cities, Leningrad has played a vital role in Russian history—for two centuries as capital of the Russian Empire—and it maintains today outstanding importance as an industrial and cultural centre and as a seaport. Founded by Peter I the Great in 1703 as St. Petersburg (Russian Sankt Peterburg), it was renamed Petrograd in 1914 and, finally, Leningrad in 1924. The city is particularly renowned as the scene of the February and October revolutions in 1917, as the besieged and fiercely defended city of World War II, and, architecturally, as one of the most splendid and harmonious cities of Europe. Leningrad lies in the far northwestern corner of the Soviet Union, about 398 miles (640 kilometres) northwest of Moscow and is situated only about 7° south of the Arctic Circle.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

Floods

The city site. Leningrad is located on the delta of the Neva River, at the head of the Gulf of Finland. The city spreads across nearly 42 islands of the delta and across adjacent parts of the mainland floodplain. The very low and originally marshy site has made the city subject to recurrent flooding, especially in the fall, when strong cyclonic winds drive gulf waters upstream, and also at the time of the spring thaw. Exceptionally severe inundations occurred in 1777, 1824, and 1924, the last two being the highest on record and flooding most of the city. To control the destructive floodwaters, the city began construction in the 1980s of an 18-mile-long dike across the Gulf of Finland. A number of canals have also been cut to assist drainage. These, together with the many natural channels, make Leningrad a city of waterways and bridges and have earned it the nickname "Venice of the North."

Greater Leningrad, the city itself with its satellites, forms a horseshoe shape around the head of the Gulf of Finland and includes the island of Kotlin in the gulf. On the north it stretches westward along the shores for nearly 50 miles to include Zelenogorsk. This northern extension is an area of dormitory towns, resorts, sanatoriums, and children's camps set among extensive coniferous forests and fringed by fine beaches and sand dunes. Some influential citizens also have summer cottages, or dachas, in this area. On the southern side of the gulf the metropolitan limits extend westward to include Petrodvorets and Lomonosov. Eastward, Greater Leningrad stretches up the Neva River to Ivanovskoye.

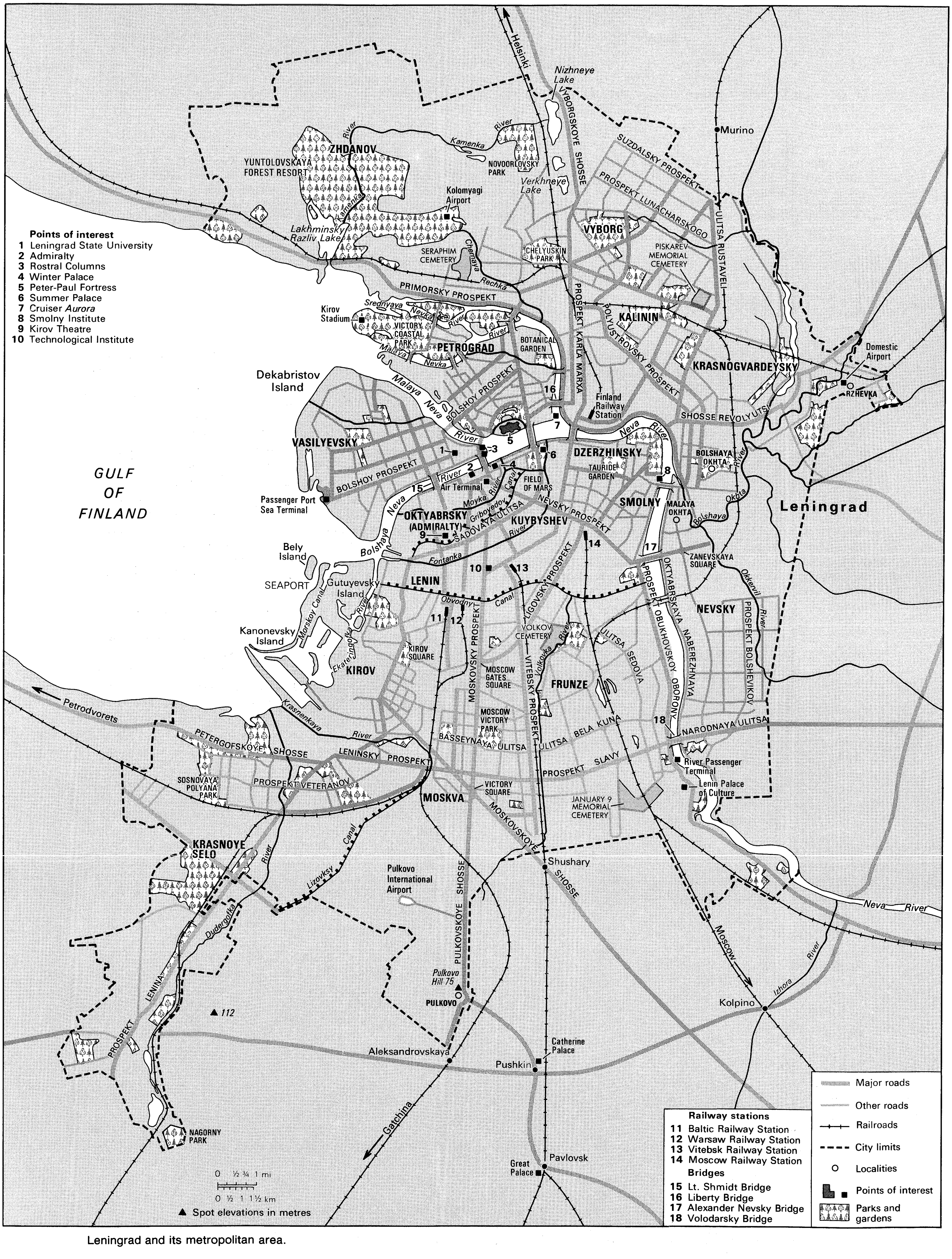
Climate. The mitigating effect of the Atlantic Ocean provides Leningrad with a milder climate than might be expected for its far northern site. Nevertheless, winters are rather cold, with a mean January temperature of about 18° F (−8° C), which is a few degrees warmer than Moscow. Winter temperature, however, can drop below −40° F (−40° C). Snow cover lasts on the average about 132 days. The Neva begins to freeze normally about mid-November, and the ice is solid by the start of December; breakup begins in mid-April and is usually completed by the end of the month. Icebreakers prolong the navigation season. Summers are moderately warm, with an average temperature in July of 64° F (18° C). The mean annual precipitation amounts to about 23 inches (584 millimetres), the maximum coming in summer. Because of Leningrad's northerly location it has long winter nights; in early summer, on the other hand, the city enjoys the half-light "white nights," one of its most renowned features.

The city layout. Central Leningrad is divided by distributaries of the Neva River into four sections. The Admiralty Side lies along the left (south) bank of the Neva itself. Between the two major arms of the Neva, the Bolshaya (Great) Neva and the Malaya (Little) Neva, is Vasilyevsky Island. The Malaya Neva and the distributary known as the Bolshaya Nevka enclose a group of islands known as the Petrograd Side, while east of the Bolshaya Nevka and north of the Neva lies the Vyborg Side.

The Admiralty

The Admiralty Side. Much of Leningrad's historical and cultural heritage is concentrated on the Admiralty Side. The district centres on the Admiralty. This, the nucleus of Peter's original city, was reconstructed in 1806–23 by Andreyan D. Zakharov, as a development of the earlier building of Ivan K. Korobov, remodeled in 1727–38, but retaining the layout of the original. Its elegant spire, topped by a weather vane in the form of a ship, is one of the principal landmarks of the city. The building today houses a naval college.

Just to the east lies the great Palace Square, the city's oldest. The 600-ton granite monolith of the Alexander



Leningrad and its metropolitan area.

Column (1830–34), the tallest of its kind in the world and so finely set that its base is not fastened, thrusts up for 165 feet (50 metres) near the centre of the square.

Between the square and the river rises the huge and massive rectangle of the Winter Palace, the former principal residence of the tsars. The present structure, the fifth to be built, was the Baroque masterpiece of Bartolomeo F. Rastrelli and was built in 1754–62. Both the exterior and the interior of the palace were designed in dazzlingly luxurious style. In 1837 the building was destroyed by fire, and only the adjoining Hermitage survived; the Winter Palace was recreated in 1839 almost exactly according to Rastrelli's plans. The striking appearance of the palace is highlighted by white columns against a green background, with golden stucco moldings; 176 sculptured figures line the roof. The whole complex, now called the Hermitage, or State Hermitage Museum, is a treasure-house of fine art of worldwide significance that originated in 1764 as the private collection of the tsarina Catherine II.

Opposite the Winter Palace, the great crescent of Carlo Rossi's General Staff building (1811–29) dominates the square. The two wings of the building are joined by a huge triumphal arch, topped by heroic figures and crowned by a chariot carrying a figure representing Glory, expressing the Russian victory in the campaign of 1812.

Decembrists' Square

On the western (downstream) side of the Admiralty stretches the expanse that was called Senate Square when the Senate moved there in 1763; it is now called Decembrists' (or Dekabrists') Square in commemoration of the revolt in 1825. The buildings of the former Senate and Synod (now housing archives) dominate the western side of the square, their decorated facades dating from the 1830s and representing the last great work of Rossi. They are separated by an arch looking across to the centre of the square where stands the equestrian statue of Peter the Great, known as the "Bronze Horseman" and created in 1782 by Étienne Falconet. Near the Senate and Synod buildings to the south rises the classical front of the Horse Guards Riding School, or Manezh (1804–07); beyond, dominating the south side of St. Isaac's Square, is the cathedral of the same name. An outstanding monument of late classical Russian architecture built by Auguste Montferrand (1818–58), St. Isaac's is one of the largest domed buildings in the world; its golden cupola, gilded with about 220 pounds (100 kilograms) of pure gold, soars to 331 feet in height and is visible all over Leningrad. It is now a museum.

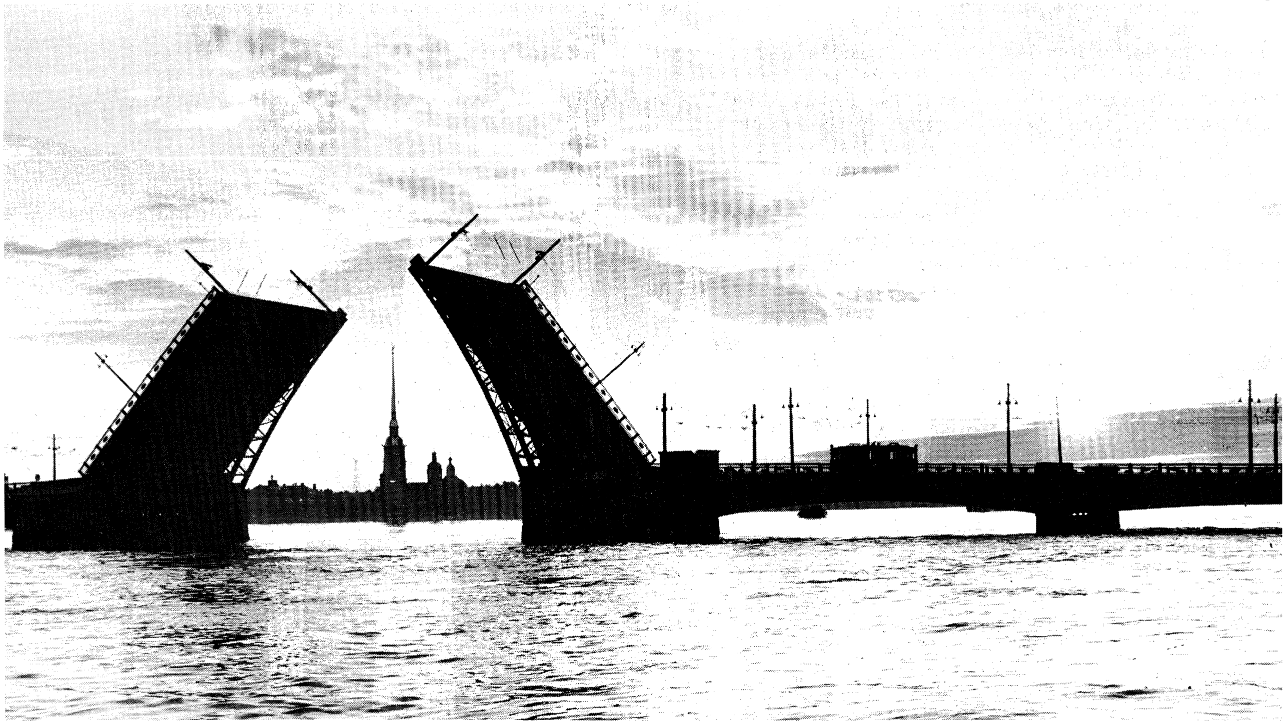
Nevsky Prospekt

From the Admiralty and its surrounding squares radiate three great avenues, of which the most important and best known is the Nevsky Prospekt. One of the world's great thoroughfares, the Nevsky Prospekt cuts southeastward across the peninsula formed by the northward loop of the Neva to the vicinity of the Alexander Nevsky Abbey, crossing the smaller Moyka and Fontanka rivers. The Anichkov Bridge across the latter is graced by four sculptured horses. The street has a special beauty; the architecture is majestic, the buildings are graceful and finely proportioned, the construction is complex. On the Nevsky Prospekt stand the Stroganov, Shuvalov, and Anichkov palaces (former private residences of the nobility) and several churches, of which the most prominent are St. Peter's Lutheran Church (1833–38), St. Catherine's Roman Catholic Church (1763–83), and the Kazan Cathedral (1801–11). The latter edifice, undoubtedly the street's finest feature, was designed by Andrey Voronikhin in Russian classical style and has an interior rich in sculptures and paintings behind a magnificent semicircular frontal colonnade. Another interesting building is the department store Gostiny Dvor (1761–85), originally designed by Jean-Baptiste M. Vallin de la Mothe. This building forms an irregular square and opens onto four streets; formerly it was a mercantile centre. Other department stores line the Nevsky Prospekt, as do many theatres—most notably the Pushkin Academic Drama Theatre—restaurants, and cafés.

At the eastern end of the Nevsky Prospekt, Alexander Nevsky Square fronts the main entrance to the abbey of the same name and its surrounding gardens. Beyond the square's main entrance lie, on the left and right, respectively, monuments and sculptures of the 18th-century Lazarus Cemetery (where Mikhail V. Lomonosov and many of the city's architects are buried) and the 19th-century Tikhvin Cemetery (containing the graves of such writers and composers as Dostoyevsky, Mussorgsky, and Tchaikovsky). Behind rise the spires and cupolas of the Church of the Annunciation (1720, designed by Domenico Trezzini), which is now a museum, and Holy Trinity Cathedral (1778–90, designed by Ivan Starov).

Through the Admiralty Side and intersecting the radial avenues cut the natural channels and canals that so characterize the city. The most important, in outward order

Novosti Press Agency—E. Ettinger



Drawbridge rising over the Neva River, Leningrad. In the background on the right embankment is the Cathedral of St. Peter and St. Paul.

from the Admiralty, are the Moyka and Fontanka rivers and the Griboyedov and Obvodny (Bypass) canals. Downstream from the northern entrance of the Fontanka into the Neva lies the Field of Mars, one of the city's beautiful open spaces. Begun under Peter the Great (when it was known as the Field of Amusement), it was intended for popular festivities and fireworks. It was a favourite haunt of the 18th-century nobility, but its present name derives from a monument erected in 1801 and portraying the great Russian military leader Aleksandr V. Suvorov (buried in the Church of the Annunciation) as the god of war. In the 19th century the space was used for military parades and exercises. The fallen of the 1917 February Revolution and defenders of the city during the civil war and foreign military intervention (1918–20) were buried there. They are commemorated by an eternal flame.

Summer Palace

Just to the east lies the Summer Garden. Founded on an island in 1704, it has parks and gardens that, by the end of the 18th century, contained more than 250 statues and busts, mostly the work of Venetian masters. The Summer Palace, Peter's first in the city, erected 1710–14 in early Russian Baroque style and designed by Trezzini, stands in the northeastern portion. The Neva embankment is fronted by a fence (1784), the iron grille of which is reputed to be among the world's finest examples of wrought iron work. So light and delicate is its design that the grillwork seems almost to be suspended in air.

At the extreme eastern side of the central city, within the sharp bend of the Neva itself, lies the Smolny complex of buildings; these include the former convent, with the five-domed cathedral, designed by Rastrelli and begun in 1748, and the classical building of the Smolny Institute, constructed by Giacomo Quarenghi in 1806–08. The institute was used as Lenin's headquarters in 1917.

The Strelka

Vasilyevsky Island. One of the first areas of Leningrad to be developed because of its defendable position, Vasilyevsky Island forms the northwestern corner of the central city. Opposite the Admiralty and Winter Palace, at the island's eastern tip, is the remarkable architectural complex known as the Strelka (literally, "Pointer"), facing the bifurcation of the Neva. Behind the two great Rostral Columns, decorated by carved ships' prows, and across Pushkin Square, the point rises majestically to the former Exchange building (Thomas de Thomon, 1805–10), the city's finest example of early 19th-century style and reminiscent of a classical Greek temple in appearance; it now houses the Central Naval Museum.

Farther back, the Twelve Colleges building (Trezzini, 1722–42), originally intended to house the supreme governmental bodies of Peter the Great, is now the home of Leningrad A.A. Zhdanov State University. The building is divided into 12 identical but independent sections and runs at right angles to the Neva embankment, which is fronted at that point by the facades of the main building of the Academy of Sciences, the Menshikov Palace, and the Academy of Arts. On the far, or northern, side of the Exchange is the Customs House (now the literary museum and the Institute of Russian Literature known as Pushkin House), designed by Giovanni Luchini (1829–32).

Cathedral of St. Peter and St. Paul

The Petrograd Side. Upstream of the bifurcation of the Neva is the Petrograd Side, where the great Peter–Paul Fortress faces the Strelka across the Malaya Neva. Founded in 1703, this fortification, the city's first structure, initially had earthen walls, but these were soon replaced by stone walls 40 feet high and 12 feet thick, with 300 cannon mounted on the bastions. Above the squat horizontal lines of the fortress's massive walls soars the slender, arrowlike spire of the Cathedral of St. Peter and St. Paul, a golden landmark for the city. The cathedral was built in 1712–33 by Trezzini, and the tsars and tsarinas of Russia from the time of Peter the Great (except for Peter II and Nicholas II) are buried in it. Trezzini also designed St. Peter's (Petrovsky) Gate (1718) as the eastern entrance to the fortress. The Neva Gate, designed by Nikolay A. Lvov, dates from 1787. From the early 19th century the fortress was used as a prison, chiefly for political prisoners. Today it is a museum. At noon each day a cannon is fired from its battlements.

Just to the east of the Peter–Paul Fortress, at the beginning of the distributary known as the Bolshaya Nevka River, the cruiser *Aurora* is permanently moored as a museum and training vessel for the Naval College. It was the *Aurora* that, in 1917, fired a blank shot as a signal to storm the Winter Palace.

Finland Railway Station

The Vyborg Side. The northeastern part of the central city had by the late 19th century developed into an industrial suburb, which in 1917 became a centre of Bolshevik support. One of its most famous features is the Finland Railway Station, which faces the Admiralty Side across the Neva. Lenin returned to Russia in April 1917 via this station, and there he made his initial pronouncement of a new course that would bring the Bolsheviks to power. Honouring that moment is the Statue of Lenin, the first major statue of the Soviet leader, which stands near the station. A major street of the Vyborg Side is the Prospekt Karla Marksa, along which stand such buildings as the modern Leningrad Hotel and the Cathedral of St. Sampson.

The outer region. Leningrad now extends well to the north and south of the original delta site, with arms of growth extending westward along the banks of the Gulf of Finland. The newer outer suburbs include extensive open areas, which help to reduce the overall population density, and parts of the periphery are designated as greenbelt. However, the multiplicity of large housing blocks containing numerous two- or three-room apartments means that population densities in the built-up areas remain extremely high. As in virtually all modern cities, commuting over long distances is the price paid for more living space and the cleaner air of the suburbs. Among the suburbs noteworthy for their historic and cultural value are Petrodvorets, Pushkin, Pavlovsk, and Gatchina.

The Great Palace

Petrodvorets. The most famous of the communities around Leningrad is Petrodvorets (Peterhof before 1944), whose unique garden-park setting, stretching in terraces rising above the Gulf of Finland, contains representative works from two centuries of Russian architectural and park styles. The Great Palace, the former residence of Peter the Great, stands at the edge of the second terrace, its bright yellow walls contrasting with white stucco decorations and the gilt domes of its lateral wings. Built in the Baroque style (1714–28), it was reconstructed and expanded by Rastrelli from the mid-1740s to the mid-1750s. On the north, the building commands a view of the Grand Cascade, a grandiose structure including a grotto, 64 fountains, and two cascading staircases, which lead to an enormous semicircular basin that contains a giant statue of Samson wrestling with a lion. This statue, symbolizing the military glory of Russia, is a copy of the original statue by Mikhail I. Kozlovsky, which was carried off by the Nazis in World War II. In fact, much of the town's treasure was plundered and this magnificent vista becomes all the more remarkable when it is remembered that much of it is a post-World War II restoration.

The Catherine Palace

Pushkin. The town of Pushkin (before 1917 Tsarskoye Selo; called Detskoye Selo in 1918–37) arose in the beginning of the 18th century as one of the tsarist residences. The Catherine Palace (1717–23; enlarged by Aleksey V. Kvasov and Savva I. Chevakinsky, 1743–48; rebuilt by Rastrelli, 1752–57) is notable for its dimensions, the beauty and majesty of its form, and the wealth of its sculptural decoration. The golden suite of splendid halls (including the Amber Room) exemplifies Russian Baroque at its peak. The community also is the site of the Chinese Village (1782–96) in Alexander Park and the gallery (1780–90) named after its architect, Charles Cameron, the terraces of which contain more than 50 busts of figures from ancient Greek and Roman history. The Lycée, a school for the offspring of the nobility, had the great Aleksandr Pushkin as a student, and a famous statue of the poet stands near the town's Egyptian Gates. The town suffered severe destruction during the German onslaught but has been restored to its former glory.

Pavlovsk. Pavlovsk, a southern suburb, is the site of a late 18th- and early 19th-century palace and park in the classical style that was created as a country residence for Tsar Paul I. The central Great Palace (1782–86; Cameron) is crowned by a dome supported by 64 columns. It was severely damaged by the Nazis and has been restored.

Gatchina. Another southern suburb, Gatchina, is noted for the palace built in 1766–81 by Antonio Rinaldi for Count Grigory Orlov, a favourite of Catherine II. Gatchina Park was created at the same time. Its monuments, sculptures, and gardens, like those of all Leningrad, are under state supervision that provides for protection and restoration.

THE PEOPLE

The population of Leningrad is overwhelmingly Russian. Before the Revolution the city had sizable Polish, Baltic, and German and smaller Tatar, Jewish, and Chinese communities. In the interwar period it continued to act as a magnet for Russian peasant labour, and, even in the more homogeneous postwar city, newcomers have tended to outnumber native Leningraders. Nevertheless, observers have commonly ascribed certain traits to the people—politeness, a sophistication, a slight reserve—that have seemingly passed from generation to generation. The old intelligentsia is no more, but many Leningraders, living in a city designed as a cultural centre, consider themselves to be the most cultivated of Russians.

THE ECONOMY

Industry. Leningrad is one of the major industrial centres in the Soviet Union, with more than half of its working population employed in factories and the building trades; its products are distributed throughout the Soviet Union. The city also makes important contributions in scientific and technical research.

Engineering industries

Engineering accounts for more than half of the city's industrial output. Special emphasis is placed on those branches of engineering that require a skilled labour force and relatively small quantities of metal. Even so, Leningrad uses about 10 percent of all steel made in the Soviet Union. Leningrad's original industry, shipbuilding, is still important and remains one of the largest of its kind in the Soviet Union; it produces icebreakers (some atomic-powered), tankers, timber carriers, and fishing vessels. Other sectors of heavy engineering make armaments and rolling stock. Of national importance are a plant producing nuclear reactors and others that manufacture electrical and power machinery, such as steam, hydraulic, and gas turbines, armatures, and generators. Precision engineering includes tools and instruments, refrigerators, radios, televisions, and other electrical and electronic goods. Machinery is produced for automated factories and for the knitwear and footwear industries.

Chemical-based industries rank second to engineering in importance. These make a wide range of items, such as superphosphate fertilizers (using apatite from the Kola Peninsula), tires and other synthetic rubber goods, plastics and plastic goods, artificial and synthetic fibres, paints, and pharmaceutical preparations. There are industries producing a wide range of consumer goods, with the city itself as their principal market. These include the production of cotton and woolen textiles, clothing, footwear, tobacco products, beer, and foodstuffs. Leningrad has a long-established and sizable printing industry.

Power sources

Electrical power for these industries comes from hydroelectric plants on the Volkhov, Svir, and Vuoksa rivers, and, more recently, from nuclear power stations. Natural gas is piped to Leningrad from the southern regions of European Russia and Soviet Central Asia.

Waterways

Transportation. Leningrad is one of the Soviet Union's most important hubs of transportation. Its port, the nation's largest, is of international significance. The main harbour is protected by breakwaters and is reached by a dredged channel from Kronshtadt. Imports include metal pipes, factory equipment, chemicals, sugar, cotton, and fruit, while machinery, timber, coal, potassium salts, and pyrites form the bulk of exports. Passenger liners maintain regular summer services to Stockholm and to Tilbury in Britain. Smaller seagoing ships have access by way of the Neva to Lake Ladoga and thence throughout the inland waterway system of European Russia. From Lake Ladoga the Svir River, Lake Onega, and the White Sea Canal (Belomor Kanal) lead to the White Sea and the Soviet Arctic coast. From Lake Onega the Lenin Volga–Baltic Waterway system leads to the Volga Basin and Caspian Sea and, via the Volga–Don Canal, to the Black and Azov seas. The main airfield, Pulkovo International Airport, is located 11 miles south of the city.

Local transport

The city is a focus of rail routes, with trunk lines radiating to Helsinki and Warsaw as well as to Moscow and other major cities of the Soviet Union. There are five principal passenger rail terminals—the Moscow, Vitebsk, Warsaw, Baltic, and Finland stations. A network of suburban electric services connects the outer parts of Leningrad and its satellite towns. Internal city traffic is carried by a subway system (opened in 1955) and a well-developed network of bus, streetcar, and trolleybus lines.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Since 1931 Leningrad has been designated a "city of republican subordination"—that is, its City Council (Soviet) is directly subordinate to the government of the Russian Soviet Federated Socialist Republic. The council members are elected for two-year terms. The city itself is divided into 16 administrative wards. In addition to these, there are five towns in the outer part of Greater Leningrad administered by the local government, which have status equivalent to that of the administrative wards; these include Kolpino, Petrodvorets, Pushkin, Sestroretsk, and Kronshtadt, on the island of Kotlin. Two other towns, Zelenogorsk and Pavlovsk, and 17 urban districts are subordinate to the administrative ward councils. Leningrad is also the administrative centre of Leningrad *oblast* (region). As a residue of the city's former status as capital, certain organizations still maintain their national headquarters in Leningrad, among them the all-union geographical, chemical, and medical societies.

Services and health. Much of the housing in Leningrad was destroyed during World War II and was replaced by massive postwar building programs. As a result a large proportion of the people live in relatively new and modern apartments, some in high-rise buildings. Virtually all housing has central heat and is tied to the city's sanitation and power services.

Leningrad is fully equipped with the health services of a modern city. There are several hundred clinics providing medical and dental care and maternity and nursing services for the residential regions and suburbs. Medical care also is provided by more than 150 general and specialized hospitals. As in the rest of the Soviet Union, health care in Leningrad is free.

Education. Leningrad is one of the most important centres for education and scientific research in the Soviet Union; a sizable proportion of the employed population is engaged in education, the arts, and the sciences. Heading the educational establishments is the Leningrad A.A. Zhdanov State University, founded in 1819. No less renowned and even older are the Academy of Arts, founded in 1757, the Institute of Mines (1773), and the Military Medical Academy (1798). The Leningrad M.I. Kalinin Polytechnic Institute (1899) is also noteworthy. The city has numerous other higher educational establishments and also a large number of general schools, as well as specialist and technical secondary schools.

Library of the Academy of Sciences

A focus for research is the library of the Academy of Sciences of the U.S.S.R., which remained in Leningrad when the academy's headquarters moved back to Moscow after the Revolution. The research establishments of the Academy of Sciences in the city include the Botanical, Geological, Forestry, and Zoological institutes and the Pulkovo Observatory. The city is the principal Soviet centre for arctic research, notably at the Arctic and Antarctic Research Institute and at the Institute for the Study of Permafrost.

CULTURAL LIFE

Leningrad evolved as a city of culture, and the number and quality of its cultural institutions remains one of its enduring attractions. It has many large and grand theatres and auditoriums. The Kirov State Academic Theatre of Opera and Ballet, formerly the Mariinsky Theatre, has long enjoyed an international reputation, and its resident company is frequently on tour abroad. Other important

Museums

theatres are the Maly (Little), Gorky, Pushkin, and Musical Comedy theatres. The largest of several concert halls is the October Great Concert Hall, which seats 4,000 people. Notable museums include the Hermitage and the State Russian Museum, which specializes in Russian painting. Leningrad is a significant centre of the Soviet motion-picture industry. There are a large number of libraries in the city, headed by the Saltykov–Shchedrin Public Library on the Nevsky Prospekt, established in 1795; among libraries in the Soviet Union, it is second only to Moscow's Lenin Library. Another important specialized collection is in the Pushkin House literary museum on Vasilyevsky Island.

Leningraders have abundant recreational facilities at their disposal. Among the stadiums in the area are the Kirov Stadium, the largest, seating about 100,000, and a 25,000-seat stadium in the Lenin sports complex. Other opportunities for outdoor recreation are provided by the Kirov Park of Culture and Rest, the zoo, the botanical gardens, and numerous other smaller parks and gardens.

History

THE EARLY PERIOD

Foundation and early growth. Settlement of the region around the head of the Gulf of Finland by Russians began in the 8th or 9th century AD. Known then as Izhorskaya Zemlya, or more commonly as Ingermanland, or Ingria, the region came under the control of Novgorod, but for long it remained thinly populated. In the 15th century the area passed, with Novgorod, into the possession of the grand princes of Moscow. Sweden annexed Ingria in 1617 and established fortresses along the Neva. During the Second Northern War of 1700–21, Peter the Great, seeking a sea outlet to the West, constructed a fleet on the Svir River (which connects Lakes Onega and Ladoga) and, sailing across Lake Ladoga, launched an attack on the fortress of Noteburg (now Petrokrepost), where the Neva flows out of Ladoga. In 1703 Noteburg fell to Peter; afterward he captured the Swedish fortress of Nienshants on the lower Neva, thus gaining control of the delta.

Peter–Paul Fortress

On May 16 (May 27, new style), 1703, shortly after the fall of Nienshants, Peter himself laid the foundation stones for the Peter–Paul Fortress on Zayachy Island. This date is taken as the founding date of St. Petersburg. In the spring of the following year Peter established the fortress of Kronshlot, later Kronshtadt, on Kotlin Island in the Gulf of Finland, to protect the approaches to the delta. At the same time he founded the Admiralty shipyard on the riverbank opposite the Peter–Paul Fortress; in 1706 its first warship was launched. Around the fortress and shipyard Peter began the building of a new city to serve as his "window on Europe." Just upstream of the Peter–Paul Fortress, the first small house, built for Peter himself in the early days of the city's construction, is preserved as a museum.

Capital of Russia

Although the first dwellings were single-storied and made of wood, it was not long before stone buildings were erected. The first stone palace, still preserved, was completed in 1714 for Prince Aleksandr Danilovich Menshikov, first governor of the city. From the start the city was planned as an imposing capital, on a regular street pattern, with spacious squares and broad avenues radiating out from the Admiralty. Architects, craftsmen, and artisans were brought from all over Russia and from many foreign countries to construct and embellish the new town. In 1712 the capital of Russia was transferred there from Moscow, although it was not until 1721 that Sweden, in the Treaty of Nystad, formally ceded sovereignty of the area to Russia. Members of the nobility and merchant class were compelled by Peter to move to the new capital and to build houses for themselves. Government buildings and private palaces and houses arose swiftly; among the earliest buildings were the Merchants' Exchange (now the Naval Museum), Customs House (now the Museum of Literature), and marine hospital, together with the Summer Palace. Canals for drainage were cut through the marshy left bank of the Admiralty Side. The first floating bridge over the Neva was constructed in 1727, and soon more than 370 bridges had been built across the many canals and river channels. Marshy, flood-prone land and an inhospitable climate made construction expensive in terms of human life; St. Petersburg, it was later suggested, rested on a swamp of human bones.

A harbour was constructed, and Peter took measures to curtail traffic through Arkhangelsk on the White Sea, previously Russia's major port. In consequence, as early as 1726 St. Petersburg was handling 90 percent of Russia's foreign trade. In 1703 work began on the Vyshnevolotsky Canal in the Valdai Hills, the first link in a chain that by 1709 gave the capital a direct water route to central Russia and all of the Volga Basin. Industry soon began to develop. The original and flourishing Admiralty shipyard was joined by enterprises to supply its needs and those of the growing fleet—a foundry to produce cannon, a gunpowder factory, and a tar works. Merchantmen as well as warships were built, and before the end of the 18th century papermaking, printing, and food, clothing, and footwear industries had been established; as early as the 1740s a factory was set up to make china. By 1765 the population numbered 150,300 and by the end of the century it had reached 220,200, of whom more than a third were in the armed forces or the administration.

Russian Baroque architecture

The rise to splendour. The growing city displayed a remarkable richness of architecture and harmony of style. Initially the style was one of simple but elegant restraint, represented in the cathedral of the Peter–Paul Fortress and in the Summer Palace. In the mid-18th century an indelible stamp was put on the city's appearance by the architects Bartolomeo F. Rastrelli, Savva I. Chevakinsky, and Vasily P. Stasov, working in the Russian Baroque style, which combined clear-cut, even austere lines with richness of decoration and use of colour. To this period belong the Winter Palace, the Smolny Convent, and the Vorontsov and Stroganov palaces, among others; outside the city were built the summer palaces of Peterhof and of Tsarskoye Selo. After a transitional period, dominated by the architecture of Jean-Baptiste M. Vallin de la Mothe and Aleksandr Kokorinov, toward the end of the 18th century a pure classical style emerged under the architects Giacomo Quarenghi, Carlo Rossi, Andrey Voronikhin, and others. The Kazan and St. Isaac's cathedrals, the Smolny Institute, the new Admiralty, the Senate, and the Mikhaylovsky Palace (now the State Russian Museum) are representative of the splendid buildings of this period.

Within this grand architectural setting cultural life developed and flourished. In 1773 the Institute of Mines was established. The University of St. Petersburg (now Leningrad State University) was founded in 1819. Many of the most celebrated names in Russia in the spheres of learning, science, and the arts are associated with the city: Mikhail V. Lomonosov, Dmitry I. Mendeleyev, Ivan Pavlov, Aleksandr Pushkin, Leo Tolstoy, and Fyodor Dostoyevsky, among others. Dostoyevsky's *Crime and Punishment* was set in the city, and the buildings described in the novel are a focus of tourism. As early as 1738 the first ballet school in Russia was opened in St. Petersburg; in the 19th century, under Marius Petipa, the Russian ballet rose to worldwide renown and produced such dancers as Vaslav Nijinsky, Tamara Karsavina, and Anna Pavlova. In 1862 the first conservatory of music in Russia opened its doors, and there the premieres of works by Tchaikovsky, Rimsky-Korsakov, Rachmaninoff, and other composers were performed. Over all, as focus and patron of the city's cultural life, stood the imperial court; its ostentatious splendour and wealth were legendary throughout Europe.

EVOLUTION OF THE MODERN CITY

New lines of communication

The road to revolution. The imperial magnificence, centred on the tsarist autocracy, was in sharp contrast to the other side of St. Petersburg's development, the growth of its industrial proletariat. During the 19th century there was much industrial growth in the city, accelerated by improvements in communications and extension of trade. Navigation was opened on the Mariinsky (1810) and Tikhvin (1811) canal systems, which replaced an old and inadequate system. In 1813 the first Russian steamship was built in St. Petersburg, and in 1837 the first Russian railway was constructed from the city to the Summer Palace in Tsarskoye Selo (now Pushkin). Five years later work started on the railway to Moscow, opened to traffic in 1851. A line to Warsaw was built in 1861–62, followed by still others. In particular, the cotton textile and metalworking industries flourished, the former using imported raw materials. By the 1840s more than 60 percent of Russian imports were entering by way of St. Petersburg. In 1885, a channel was dredged to give larger ships access to the port. City growth and industrialization were stimulated by the emancipation of the Russian serfs in 1861, which allowed far greater mobility of labour. From 539,400 inhabitants in 1864 the population rose to about 1,500,000 in 1900, largely by migration from rural areas (as late as 1910 only one-third of the population had been born in the city). By 1917 the total had risen to about 2,500,000.

Politicized work force

The factory environment in Leningrad became a breeding ground for revolution. With the development of metalworking and engineering as the primary industries, there arose a skilled labour force, increasingly alert politically. Moreover, the factory workers, who numbered nearly 250,000 in 1914, tended to be concentrated in plants of far larger size than was usual in Russia; the Putilov (now Kirov) armaments works alone employed about 13,000. It was thus easier for revolutionaries to spread their ideas and for workers' groups to organize than it was elsewhere in Russia. At the same time the growth of the city was characterized by a belated and slow development of public transport, making it necessary for workers to live close to their place of work, in conditions of appalling overcrowding (more than 180,000 per square mile in the centre), squalor, and lack of sanitation. Throughout the period before 1917 the city administration was lacking in efficiency and often in funds, and the provision of all public services—even a water supply—was inadequate. Outbreaks of serious epidemics were commonplace, and St. Petersburg became notorious for its unhealthy conditions.

Decembrist insurrection

The first serious revolutionary outbreak in St. Petersburg came on Dec. 14 (Dec. 26, N.S.), 1825, the Decembrist insurrection, organized largely by liberal aristocrats and army officers seeking a liberal constitution and an end to serfdom. It was ruthlessly suppressed. During the rest of the 19th century, workers' revolutionary activity and unrest steadily increased, with ever more frequent strikes and outbreaks of violence. These culminated in the general strike of January 1905, when some 150,000 workers took part. On what became known as Bloody Sunday, January 9 (22, N.S.), a mass march to the Winter Palace, bearing a petition to the Tsar, was met by troops, who opened fire; more than 100 people were killed and hundreds more wounded. The situation developed into revolution, spreading throughout Russia. Although it was again crushed, underground revolutionary activity continued.

Return of Lenin

The outbreak of World War I in 1914 brought an upsurge of patriotic fervour centred on the Tsar. The Germanic form of the city's name was changed to its Russian version, Petrograd. The military disasters of the war and the worsening economic situation, however, revived and intensified discontent. Transport inefficiencies led to severe shortages of food and other supplies. On Feb. 26 (March 11, N.S.), 1917, with a general strike in effect, disorders broke out. The authorities were slow to act and lost all control. The Petrograd Soviet of Workers' and Soldiers' Deputies was formed on February 27 (March 12, N.S.). On March 2 (15, N.S.) the Tsar abdicated. A provisional government was set up, eventually under the premiership of Aleksandr Kerensky. On April 3 (16, N.S.) Lenin returned to Petrograd from Switzerland and set about organizing the overthrow of the provisional government. Demonstrations in July were suppressed, but on October 25 (November 7, N.S.) Bolshevik-led workers and sailors stormed the Winter Palace, deposing the provisional government and establishing the Bolshevik Party in power.

The Russian Revolution of 1917, which changed the course of history, was spearheaded by the Petrograd proletariat and the sailors from Kronshtadt. In January 1918 a Constituent Assembly met in Petrograd, but the Bolsheviks, who had won only a minority of seats, dispersed it and consolidated their authority.

The Soviet period. Civil war reigned in Russia from 1918 to 1920, during which the Bolsheviks successfully defended their government against various Russian and foreign elements. In March 1918 the capital of the young Soviet state had been moved back to Moscow. The years of the civil war after the Revolution had a disastrous effect on the city's economy. Industry came very nearly to a standstill. The population fell sharply to 722,000 in 1920, a mere third of the pre-Revolutionary size. Many died of starvation. Recovery began when the war ended. In 1924, following Lenin's death, the city was renamed Leningrad. When in 1928 the era of five-year plans began, much of the initial burden of developing the national economy fell on Leningrad and its established industrial plant and work force, especially in the provision of power equipment and machinery. This stimulated further growth, and by 1939 the city was responsible for 11 percent of all Soviet industrial output and its population had passed 3,000,000.

The 900-day siege

Then once again Leningrad was struck by a period of loss and destruction. It was one of the initial targets of the German invasion in 1941; by September of that year German troops were on the outskirts of the city and had cut off communication with the rest of the Soviet Union, while Finnish troops advanced from the north. Many of the inhabitants and nearly three-quarters of the industrial plant were evacuated eastward ahead of the German advance. The remainder of the population and the garrison then began to endure what has become known as the 900-day siege; the German blockade in fact lasted 872 days, from Sept. 8, 1941, to Jan. 27, 1944. Leningrad put up a desperate and courageous resistance in the face of many assaults, constant artillery and air bombardment, and appalling suffering from shortages of supplies. An estimated 660,000 people died, a very high proportion from scurvy and starvation. In particular the exceptionally bitter winter of 1941–42, when temperatures fell to −40° F (−40° C), was one of extreme hardship and loss of life. The only route for supplies was the "road of life" across the ice of Lake Ladoga; later an oil pipeline and electric cables were laid on the lake bed. The blockade proper was broken in January 1943, but it was another year before the Germans had been driven back from the outskirts. Enormous damage had been caused by the bombardment, and before retreating the Germans destroyed the palaces at Petrodvorets and Pushkin. For its role in the war Leningrad was honoured with the title of Hero City, a special defense medal, and the Order of Lenin. Not until the 1960s did the city regain its prewar size of 3,000,000; by the 1980s the population had passed 4,000,000.

In Leningrad, the first postwar Five-Year Plan was devoted to reconstruction of the city's industry and restoration of its architectural heritage. In the late 1950s a program of housing construction in the new suburbs got under way; renovation of highly sought after inner-city apartments began in the 1970s. In the face of continuing construction and expansion, maintenance of the old city and modernization of the infrastructure have become major problems. In response the city planners have pioneered new forms of industrial administration, drawing on the city's strength as a scientific and technical centre and emphasizing the need to preserve its unique cultural heritage.

BIBLIOGRAPHY. General descriptive works in English are NIGEL GOSLING, *Leningrad: History, Art, Architecture* (1965); and K. NEUBERT and J. NEUBERT, *Portrait of Leningrad* (1966). Two helpful books, both translated from Russian, are PAVEL KANN, *The Environs of Leningrad* (1981); and A. BEREZINA, *Leningrad: A Short Guide* (1980). Practical information can be found in EVAN MAWDSLEY and MARGARET MAWDSLEY (eds.), *Moscow and Leningrad* (1980). On the historical background, see JAMES H. BATER, *St. Petersburg: Industrialization and Change* (1967); and E.M. ALMEDINGEN, *Tomorrow Will Come* (1941, reprinted 1983), a memoir spanning the revolutionary period. On the siege of Leningrad, see HARRISON E. SALISBURY, *The 900 Days* (1969, reprinted 1985). An account of the city government is provided by DAVID T. CATTELL, *Leningrad: A Case Study of Soviet Urban Government* (1968). For politics and city planning in the 1970s, see DENIS J.B. SHAW, "Planning Leningrad," *Geographical Review,* 68(2):183–200 (April 1978); and BLAIR A. RUBLE, "Romanov's Leningrad," *Problems of Communism,* 32(6):36–48 (Nov.–Dec. 1983). The cultural

history is discussed in JOHN GREGORY and ALEXANDER UKLADNIKOV, *Leningrad's Ballet: Maryinsky to Kirov* (1980). LOGAN ROBINSON, *An American in Leningrad* (1982), is an account by a Harvard law student.

Works in Russian include Л.С. ШАУМЯН (ed.), *Ленинград: энциклопедический справочник* (1957), a survey of Leningrad's history, economics, public education, cultural and educational institutions, architecture and construction, public health service, science, literature, and art; С.М. СЕРПОКРЫЛ (comp.), *Ленинград: Достопримечательности,* 3rd enlarged ed. (1974), an account of the basic architectural aggregations, squares, embankments, avenues, and suburban parks; and М.П. ВЯТКИН (ed.), *Очерки истории Ленинграда*, 6 vol. (1955–70), an analysis of the development of the city and its construction, with an account of the stages of its revolutionary development. See also *Ленинград: историко-географический атлас* (1981), a historic-geographical atlas.

(Y.M.D./R.A.F./M.McA.)

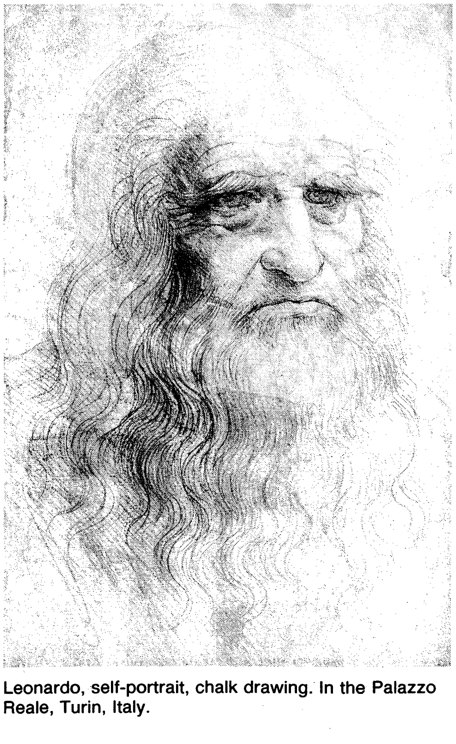
Leonardo da Vinci

The unique fame that the Florentine artist and scientist Leonardo da Vinci enjoyed in his lifetime and that, filtered and purified by historical criticism, has remained undimmed to the present day is based on the equally unique universality of his spirit. Leonardo's universality is more than many-sidedness. True, at the time of the Renaissance and the period of Humanism, many-sidedness was a highly esteemed quality; but it was by no means rare. Many other good artists possessed it. Leonardo's universality, on the other hand, was a spiritual force, peculiarly his own, that generated in him an unlimited desire for knowledge and guided his thinking and behaviour. An artist by disposition and endowment, he found that his eyes were his main avenue to knowledge; to Leonardo, sight was man's highest sense organ because sight alone conveyed the facts of experience immediately, correctly, and with certainty. Hence, every phenomenon perceived became an object of knowledge. *Saper vedere* ("knowing how to see") became the great theme of his studies of man's works and nature's creations. His creativity reached out into every realm in which graphic representation is used: he was painter, sculptor, architect, and engineer. But he went even beyond that. His superb intellect, his unusual powers of observation, and his mastery of the art of drawing led him to the study of nature itself, which he pursued with method and penetrating logic—and in which his art and his science were equally revealed.

LIFE AND WORKS

Early period: Florence. The illegitimate son of Ser Piero, a Florentine notary and landlord, Leonardo was born in 1452 on his father's family estate in Vinci, near Empoli. His mother, Caterina, was a young peasant woman who shortly thereafter married an artisan from that region. Not until his third and fourth marriages did Ser Piero's wives have children, the first one in 1476, when Leonardo was already an adult. Thus, Leonardo grew up in his father's house, where he was treated as a legitimate son and received the usual elementary education of that day: reading, writing, and arithmetic. As for Latin, the key language of traditional learning, Leonardo did not seriously study it until much later, when he acquired a working knowledge of it on his own. Not until he was 30 years old did he apply himself to higher mathematics—advanced geometry and arithmetic—which he studied with diligent tenacity; but here, too, he did not get much beyond the beginning stages.

Alinari—Art Resource/EB Inc.



Leonardo, self-portrait, chalk drawing. In the Palazzo Reale, Turin, Italy.

Leonardo's artistic inclinations must have appeared early. When he was about 15, his father, who enjoyed a high reputation in the Florence community, apprenticed him to Andrea del Verrocchio. In Verrocchio's renowned workshop Leonardo received a many-sided training that included not only painting and sculpture but the technical-mechanical arts as well. He also worked in the next-door workshop of Antonio Pollaiuolo, where he was probably first drawn to the study of anatomy. In 1472 Leonardo was accepted in the painters' guild of Florence but remained five years more in his teacher's workshop. Then he worked independently in Florence until 1481. In the few extant works of this early period one may clearly trace the development of the artist's remarkable talent. Keenness of observation and creative imagination stand out. His early mastery is revealed in an angel and a segment of landscape executed by him in Verrocchio's painting the "Baptism of Christ" (Uffizi, Florence) and in two Annunciations (Uffizi, as well as the Louvre, Paris), both of them done in Verrocchio's workshop, as were the "Madonna with the Carnation," the "Madonna Benois," and the "Portrait of Ginevra de' Benci." This mastery reached its peak in two paintings that remained unfinished: "St. Jerome" and a large panel painting of "The Adoration of the Magi." In addition to these few paintings there are a great many superb pen and pencil drawings, in which Leonardo's mastery blazed new trails for this graphic art. Among the drawings are many technical sketches—for example, pumps, military weapons, mechanical apparatus—evidence of Leonardo's interest in and knowledge of technical matters at the outset of his career.

Evidence of early mastery

Unfolding of Leonardo's genius: first Milanese period (1482–99). In 1482 Leonardo entered the service of the Duke of Milan—a surprising step when one realizes that the 30-year-old artist had just received his first substantial commissions from his native city of Florence: the above-mentioned unfinished panel painting of "The Adoration of the Magi" for the monastery of S. Donato a Scopeto (1481) and an altar painting for the St. Bernard Chapel in the Palazzo della Signoria, which was never fulfilled. That he gave up both projects despite the commitments he had undertaken—not even starting on the second named—seems to indicate deeper reasons for his leaving Florence. It may have been that the rather sophisticated spirit of Neoplatonism prevailing in the Florence of the Medici went against the grain of his experience-oriented mind and that the more realistic academic atmosphere of Milan attracted him. Moreover, there was the fascination of Ludovico Sforza's brilliant court and the meaningful projects awaiting him there.

Leonardo spent 17 years in Milan, until Ludovico's fall from power in 1499. He was listed in the register of the royal household as *pictor et ingeniarius ducalis* ("painter and engineer of the duke"). Highly esteemed, Leonardo was constantly kept busy as a painter and sculptor and as a designer of court festivals. He was also frequently consulted as a technical adviser in the fields of architecture, fortifications, and military matters, and he served as a hydraulic and mechanical engineer.

In this phase of his life Leonardo's genius unfolded to the full, in all its versatility and creatively powerful artistic and scientific thought, achieving that quality of uniqueness that called forth the awe and astonished admiration of his contemporaries. At the same time, in the boundlessness of the goals he set himself, Leonardo's genius bore the mark of the unattainable so that, if one traces the outlines of his lifework as a whole, one is tempted to call it a grandiose "unfinished symphony."

The six paintings completed in the Milanese years

Painting and sculpture. As a painter Leonardo completed only six works in the 17 years in Milan: portraits of Cecilia Gallerani ("Lady with an Ermine") and a musician, an altar painting of "The Virgin of the Rocks" (two versions), a monumental wall painting of the "Last Supper" in the refectory of the monastery of Sta. Maria delle Grazie (1495–97), and the decorative ceiling painting of the Sala delle Asse in the Milan Castello Sforzesco (1498). Three other pictures that, according to old sources, Leonardo was commissioned to do have disappeared or were never done: a "Nativity" said to have belonged to Emperor Maximilian; a "Madonna" that Ludovico Sforza announced as a gift to the Hungarian king Matthias Corvinus; and the portrait of one of Ludovico's mistresses, Lucrezia Crivelli.

Also unfinished was a grandiose sculptural project that seems to have been the real reason Leonardo was invited to Milan: a monumental equestrian statue in bronze to be erected in honour of Francesco Sforza, the founder of the Sforza dynasty. Leonardo devoted 12 years—with interruptions—to this task. Many sketches of it exist, the most impressive ones discovered only in the mid-20th century, when two of Leonardo's notebooks came to light again in Madrid. They reveal the sublimity but also the almost unreal boldness of his conception. In 1493 the clay model of the horse was put on open display on the occasion of the marriage of Emperor Maximilian with Bianca Maria Sforza, and preparations were made to cast the colossal figure, which was to be 16 feet (five metres) high—double the size of Verrocchio's equestrian statue of Bartolomeo Colleoni! But, because of the imminent danger of war, the metal, ready to be poured, was used for cannon instead, and so the project came to a halt. Ludovico's fall in 1499 sealed the fate of this abortive undertaking, which was perhaps the grandest concept of a monument in the 15th century. The ravages of war left the clay model a heap of ruins.

As a master artist Leonardo maintained an extensive workshop in Milan, employing apprentices and students. The role of most of these associates is unclear. Their activity involves the question of Leonardo's so-called apocryphal works, in which the master collaborated with his assistants. Scholars have been unable to agree in their attributions of these works, which include such paintings as "La Belle Ferronnière" in the Louvre, the so-called "Lucrezia Crivelli" in the Pinacoteca Ambrosiana, Milan, and the "Madonna Litta" in the Hermitage, Leningrad. Among Leonardo's pupils at this time were Giovanni Antonio Boltraffio, Ambrogio de Predis, Bernardino de' Conti, Francesco Napoletano, Andrea Solari, Marco d'Oggiono, and Salai.

Art and science: the notebooks. The Milan years also saw Leonardo's decided turn toward scientific studies. He began to pursue these systematically and with such intensity that they demanded more and more of his time and energy and developed into an independent realm of creative productivity. Within him there arose now a growing need to note and write down in literary form every one of his perceptions and experiences. It is a unique phenomenon in the history of art. Undoubtedly, the several treatises on art that appeared or were made available during those decades provided an external stimulus. Leon Battista Alberti's *De re aedificatoria* (*Ten Books on Architecture*) was first printed in 1485; Francesco di Giorgio's treatise on architecture was available in its first manuscript versions, and Leonardo had received a copy from the author as a gift. Moreover, Piero della Francesca in his *De prospectiva pingendi* ("On Perspective in Painting") had provided for his contemporaries a model text on the theory of perspective. Finally, there was the mathematician Lucas Pacioli, who had become an acquaintance of Leonardo's. In 1494 Pacioli published his *Summa de arithmetica geometria proportioni et proportional ità,* followed by his *Divina proportione* ("On Divine Proportion"), for which Leonardo drew figures of symmetrical bodies.

Stimulus for Leonardo's scientific studies

In this ambience Leonardo began to nourish the desire to write a theory of art of his own, and there arose in him the far-reaching concept of a "science of painting." Alberti and Piero della Francesca had already offered proof of the mathematical basis of painting in their analysis of the laws of perspective and proportion and thereby buttressed painting's claim to being a science. But Leonardo's claims went much further. Proceeding from the basic conviction that sight is the human being's most unerring sense organ, yielding immediate, accurate, and reliable data of experience, Leonardo—equating "seeing" with "perceiving"—arrived at a bold conclusion: the painter, doubly endowed with subtle powers of perception and the complete ability to pictorialize them, was the prime person qualified to achieve knowledge by observing and to reproduce that knowledge authentically in a pictorial manner. Hence, Leonardo conceived the staggering plan of observing all objects in the visible world, recognizing their form and structure, and pictorially describing them exactly as they are. Thus, drawing became the chief instrument of his didactic method.

In the years between 1490 and 1495 the great program of Leonardo the writer (author of treatises) began. In it, four main themes, which were to occupy him for the rest of his life, could be discerned and gradually took shape: a treatise on painting, a treatise on architecture, a book on the elements of mechanics, and a broadly outlined work on human anatomy. His geophysical, botanical, hydrological, and aerological researches also belong to this period and constitute parts of the "visible cosmology" that loomed before Leonardo as a distant goal. Against speculative book knowledge, which he scorned, he set irrefutable facts gained from experience—from *saper vedere.*

The notebooks

All these studies and sketches were written down in Leonardo's notebooks and on individual sheets of paper. Altogether they add up to thousands of closely written pages abundantly illustrated with sketches—the most voluminous literary legacy any painter has ever left behind. Of more than 40 codices mentioned in the older sources—often, of course, rather inaccurately—21 have survived; these in turn sometimes contain notebooks originally separate and now bound together so that 31 in all have been preserved. To these should be added several large bundles of documents: an omnibus volume in the Biblioteca Ambrosiana, Milan, called Codex Atlanticus because of its size, was collected by the sculptor Pompeo Leoni at the end of the 16th century; its sister volume, after a roundabout journey, fell into the possession of the English crown and was placed in the Royal Library, Windsor Castle. Finally there is the Arundel Manuscript (British Museum, MS. 263), which contains a number of Leonardo's fascicles on various themes.

It was during his years in Milan that Leonardo began the earliest of these notebooks. He would first make quick sketches of his observations on loose sheets or on tiny paper pads he kept in his belt; then he would arrange them according to theme and enter them in order in the notebook. Surviving are a first collection of material for the painting treatise (MSS. A and B in the Institut de France, Paris), a model book of sketches for sacred and profane architecture (MS. B, Institut de France, Paris), the treatise on elementary theory of mechanics (MS. 8937, Biblioteca Nacional, Madrid), and the first sections of a treatise on the human body (Anatomical MS. B; Windsor Castle, Royal Library).

Two special features make Leonardo's notes and sketches unusual: his use of mirror writing and the relationship between word and picture.

Leonardo was left-handed; so mirror writing came easily and naturally to him. It should not be looked upon as a secret handwriting. Though somewhat unusual, his script can be read clearly and without difficulty with the help of a mirror—as his contemporaries testified. But the fact that Leonardo used mirror writing throughout, even in his fair copies drawn up with painstaking calligraphy, forces one to conclude that, although he constantly addressed an imaginary reader in his writings, he never felt the need to achieve easy communication by using conventional handwriting. Yet occasional examples of normal handwriting (drafts of letters, notes, and comments to be submitted to third parties) show that Leonardo was completely at home in it. In the overwhelming majority of his notes in mirror writing, therefore, one gets the strong impression of "monologues in writing." Finally, then, his writings must be interpreted as preliminary stages of works destined for eventual publication, which Leonardo never got around to completing. In a sentence in the margin of one of his late anatomy sketches, he implores his followers to see that his works are printed.

Function of Leonardo's illustration vis-à-vis text

The second unusual feature in Leonardo's writings is the new function given to illustration vis-à-vis the text. Leonardo strove passionately for a language that was clear yet expressive. The vividness and wealth of his vocabulary were the result of intense self-study and represented a significant contribution to the evolution of scientific prose in the Italian vernacular. On the other hand, in his teaching method Leonardo gave absolute precedence to the illustration over the written word; hence, the drawing does not illustrate the text; rather, the text serves to explain the picture. In formulating his own principle of graphic representation—which he himself called *dimostrazione* ("demonstrations")—Leonardo was a precursor of modern scientific illustration.

Thus, during Leonardo's years in Milan the two "action fields"—the artistic and the scientific—developed and shaped his future creativity. It was a kind of "creative dualism," with mutual encouragement but also mutual pressure from each field.

Second Florentine period (1500–06). In December 1499 or at the latest January 1500—three months after the victorious entry of the French into Milan—Leonardo left that city in the company of Lucas Pacioli. He stopped first at Mantua, where, in February 1500, he drew a portrait of his hostess, Marchioness Isabella d'Este, and then proceeded to Venice (in March), where the Signoria (governing council) sought his advice on how to ward off a threatened Turkish incursion in Friuli. Leonardo recommended that they prepare to flood the menaced region. From Venice he returned to Florence, where, after a long absence, he was received with acclaim and honoured as a renowned native son. In that same year he was appointed an architectural expert to a committee investigating damages to the foundation and structure of the church of S. Francesco al Monte. A guest of the Servite order in the cloister of SS. Annunziata, Leonardo began there a cartoon for a painting of the "Virgin and Child with St. Anne," the composition of which won admiration from artists and art lovers of the city. He also painted (1501) a "Madonna with the Yarn-Winder," which has survived only in copies and which he probably never finished. Mathematical studies seem to have kept him away from his painting activity much of the time, or so Isabella d'Este, who sought in vain to obtain a painting done by him, was informed by Fra Pietro Nuvolaria, her representative in Florence.

Service with Cesare Borgia

Only his omnivorous "appetite for life" can explain Leonardo's decision, in the summer of the following year (1502), to leave Florence and enter the service of Cesare Borgia as "senior military architect and general engineer." Borgia, the notorious son of Pope Alexander VI, had, as commander in chief of the papal army, sought with unexampled ruthlessness to gain control of the Papal States of Romagna and the Marches. Now he was at the peak of his power and, at 27, was undoubtedly the most compelling and at the same time most feared person of his time. Leonardo, twice his age, must have been fascinated by his personality. For 10 months he travelled across the condottiere's territories and surveyed them. In the course of his activity Leonardo sketched some of the city plans and topographical maps that laid the groundwork for modern cartography. At the court of Cesare Borgia, Leonardo also met Niccolò Machiavelli, temporarily stationed there as a political observer for the city of Florence.

In the spring of 1503 Leonardo returned to Florence to make an expert survey of a project for diverting the Arno River behind Pisa so that the city, then under siege by the Florentines, would be deprived of access to the sea. The plan proved unworkable, but Leonardo's activity led him to a much more significant theme, one that served peace rather than war; the project, first advanced in the 13th century and now again under consideration, was to build a large canal that would bypass the unnavigable stretch of the Arno and connect Florence by water with the sea. Leonardo developed his ideas in a series of studies; with panoramic views of the river bank, which are also landscape sketches of great artistic charm, and with exact measurements of the terrain, he produced a map in which the route of the canal (with its transit through the mountain pass of Serravalle) was shown. The project, considered time and again in subsequent centuries, was never carried out, but centuries later the express highway from Florence to the sea was built over the exact route Leonardo chose for his canal.

That same year (1503), however, Leonardo also received a prized commission: to paint a mural for the Hall of the Five Hundred in Florence's Palazzo Vecchio; a historical scene of monumental proportions (at 23 × 56 feet [7 × 17 metres], it would have been twice as large as the "Last Supper"). For three years he worked on this "Battle of Anghiari"; like its intended complementary painting, Michelangelo's "Battle of Cascina," it remained unfinished. But the cartoon and the copies showing the main scene of the battle, the fight for the standard, were for a long time, to quote the sculptor Benvenuto Cellini, "the school of the world." These same years saw the portrait of "Mona Lisa" and a painting of a standing "Leda," which was not completed and has survived only in copies.

Scientific study during the Florentine period

The Florentine period was also, however, a time of intensive scientific study; Leonardo did dissections in the hospital of Sta. Maria Nuova and broadened his anatomical work into a comprehensive study of the structure and function of the human organism. He made systematic observations of the flight of birds, concerning which he planned a treatise. Even his hydrological studies, "on the nature and movement of water," broadened into research on the physical properties of water, especially the laws of currents, which he compared with those pertaining to air. These were also set down in his own collection of data, contained in the so-called Leicester Codex in Holkham Hall, Norfolk, England.

Second Milanese period (1506–13). Thus, during these years in Florence, Leonardo's productivity was also marked by his "creative dualism." Only sporadically did he work at his paintings. When, in May 1506, Charles d'Amboise, governor of the King of France in Milan, asked and was granted permission by the Signoria in Florence for Leonardo to go for a time to Milan, the artist had no hesitation about accepting the invitation. But what was originally a limited period of time became a permanent move under the stress of political circumstances. Florence let Leonardo go, and the monumental "Battle of Anghiari" remained unfinished. Unsuccessful technical experiments with paints seem to have impelled Leonardo to stop working on the mural. One cannot otherwise explain his abandonment of this great work—great both in conception and in realization.

Leonardo spent six years in Milan, interrupted only by a six-month stay in Florence in the winter of 1507–08, where he helped the sculptor Giovanni Francesco Rustici execute his bronze statues for the Florence Baptistery but did not resume work on the "Battle of Anghiari." Honoured and admired by his patrons Charles d'Amboise and King Louis XII, who gave him a yearly stipend of 400 ducats, Leonardo never found his duties onerous. They were limited to advice in architectural matters, tangible evidence of which are plans for a palace–villa for Charles d'Amboise and perhaps also sketches for an oratory for the church of Sta. Maria alla Fontana, which Charles funded. Leonardo also looked into an old project revived by the French governor: the Adda canal that would link Milan with Lake Como by water.

In Milan he did very little as a painter: two Madonnas, which he promised the King of France, were never painted. He continued to work on the paintings of the "Virgin and Child with St. Anne" and "Leda," which he had brought with him from Florence, as copies from the Lombard school of that period attest. Again Leonardo gathered pupils around him. With Ambrogio de Predis he completed a second version of "The Virgin of the Rocks" (1508), in the course of which protracted litigation between the purchasers and the artists had a happy ending. Of his older disciples, Bernardino de' Conti and Salai were again in his studio; new pupils came, among them Cesare da Sesto, Giampetrino, Bernardino Luini, and the young nobleman Francesco Melzi, Leonardo's most faithful friend and companion until his death.

The tomb sculpture for Trivulzio

An important commission in sculpture came his way. Gian Giacomo Trivulzio had returned victoriously to Milan as marshal of the French army and a bitter foe of Ludovico Sforza. He commissioned Leonardo to sculpt his tomb, which was to take the form of an equestrian statue and be placed in the mortuary chapel donated by Trivulzio to the church of S. Nazaro Maggiore. But after years of preparatory work on the monument, for which a number of significant sketches have survived, the Marshal himself gave up the plan in favour of a more modest one; so this undertaking, too, remained unfinished. Leonardo must have felt keenly this second disappointment in his work as a sculptor.

Compared with his almost cursory work in art, Leonardo's scientific activity flourished. His studies in anatomy achieved a new dimension in his collaboration with a famous anatomist from Pavia, Marcantonio della Torre. He outlined a plan for an overall work that would include not only exact, detailed reproductions of the human body and its organs but would also include comparative anatomy and the whole field of physiology. He even thought he would finish his anatomical manuscript in the winter of 1510–11. Beyond that, his manuscripts are replete with mathematical, optical, mechanical, geological, and botanical studies that must be understood as data for his "perceptual cosmology." This became increasingly actuated by a central idea: the conviction that force and motion as basic mechanical functions produce all outward forms in organic and inorganic nature and give them their shape and, furthermore, the recognition that these functioning forces operate in accordance with orderly, harmonious laws.

Last years (1513–19). In 1513 political events—the temporary ouster of the French from Milan—caused the now 60-year-old Leonardo to move again. At the end of the year he went to Rome, accompanied by his pupils Melzi and Salai as well as by two studio assistants, hoping to find employment there through his patron, Giuliano de' Medici, brother of the new pope Leo X. Giuliano gave him a suite of rooms in his residence, the Belvedere, in the Vatican. He also gave him a considerable monthly stipend, but no large commissions came to him. For three years Leonardo remained in the Eternal City, off to one side, while Donato Bramante was building St. Peter's, Raphael was painting the last rooms of the Pope's new apartments, Michelangelo was struggling to complete the tomb of Pope Julius, and many younger artists such as Peruzzi, Timoteo Viti, and Sodoma were active there. Drafts of embittered letters betray the disappointment of the aging master who worked in his studio on mathematical studies and technical experiments or, strolling through the city, surveyed ancient monuments. A magnificently executed map of the Pontine Marshes (Royal Library, Windsor Castle; 12684) suggests that Leonardo was at least a consultant for a reclamation project that Giuliano de' Medici ordered in 1514. On the other hand, there were sketches for a spacious residence for the Medici in Florence, who had returned to power there in 1512. But this did not go beyond the stage of preliminary sketches and never came to pass. Leonardo seems to have resumed his friendship with Bramante, but the latter died in 1514. And there is no record of Leonardo's relations with any other artists in Rome.

Service with King Francis I of France

In a life of such loneliness, it is easy to understand why Leonardo, despite his 65 years, decided to accept the invitation of the young king Francis I to enter his service in France. At the end of 1516 he left Italy forever, together with his most devoted pupil, Francesco Melzi. Leonardo spent the last three years of his life in the small residence of Cloux (later called Clos-Lucé), near the King's summer palace at Amboise on the Loire. *Premier peintre, architecte et méchanicien du Roi* ("first painter, architect, and mechanic of the King") was the proud title he bore; yet the admiring King left him complete freedom of action. He did no more painting or at most completed the painting of the enigmatic, mystical "St. John the Baptist," which the Cardinal of Aragon, when he visited Amboise, saw in Leonardo's studio along with the "Mona Lisa" and the "Virgin and Child with St. Anne."

For the King he drew up plans for the palace and garden of Romorantin, destined to be the widow's residence of the Queen Mother. But the carefully worked-out project, combining the best features of Italian-French traditions in palace and landscape architecture, had to be halted because the region was threatened with malaria.

Leonardo still made sketches for court festivals, but the King treated him in every respect as an honoured guest. Decades later, Francis I talked with the sculptor Benvenuto Cellini about Leonardo in terms of the utmost admiration and esteem. Leonardo spent most of his time arranging and editing his scientific studies. The final drafts for his treatise on painting and a few pages of the anatomy appeared. Consummate drawings such as the "Floating Figure" (Royal Library, Windsor Castle; 12581) are the final testimonials to his undiminished genius. In the so-called "Visions of the End of the World," or "Deluge" (Royal Library, Windsor Castle), he depicts with overpowering pictorial imagination the primal forces that rule nature.

On May 2, 1519, Leonardo died at Cloux. He was laid to rest in the palace church of Saint-Florentin. But the church was devastated during the French Revolution and completely torn down at the beginning of the 19th century. Hence, his grave can no longer be located. Francesco Melzi fell heir to his artistic and scientific estate.

ANALYSIS AND EVALUATION OF LEONARDO'S ACHIEVEMENT

Painting. Leonardo's total output in painting is really not large; only 17 of the paintings that have survived can be definitely attributed to him, and several of them are unfinished. Two of his most important works—the "Battle of Anghiari" and the "Leda," neither of them completed—have only survived in copies. Yet these few creations have established the unique fame of a man whom Vasari, in his *Lives,* dividing art history into three ages, placed in the last "golden age of the arts." His works, unaffected by all the vicissitudes of aesthetic doctrines in subsequent centuries, have stood out in all periods and all countries as consummate masterpieces of painting.

Leonardo's art of expression

The many testimonials to Leonardo, ranging from Vasari to Peter Paul Rubens, Johann Wolfgang von Goethe, and Eugène Delacroix, make it unmistakably clear that it has been, above all, Leonardo's art of expression that has called forth the utmost admiration. It is, in fact, the core of his formation as a painter—from his earliest beginnings to his last work. This expression was nurtured by his power of invention but also by every technical means: drawing, colour, use of light and shadow. To Leonardo, expression became a key concept of art; it also included the basic demands of truth, beauty, and accuracy in everything depicted.

What Leonardo was striving for was already revealed in his angel in Verrocchio's "Baptism of Christ" (*c.* 1474–75): in the natural structuring of the angel's body based on movement in several directions, in the relaxation of his attitude, and in his glance, which takes in what is occurring but at the same time is directed inward. In his landscape segment in the same picture, Leonardo also found a new expression for "nature experienced," in reproducing the forms he perceived as if through a veil of mist. The landscape study (Uffizi, Florence) dated 1473, a pen drawing, foreshadows in its treatment of transparent atmosphere by a 21-year-old his telling ability to transform perceived phenomena into convincing graphic forms.

In the "Madonna Benois" (1478) Leonardo succeeded in giving an old traditional type of picture a new, unusually charming, and expressive mood by showing the child Jesus reaching for the flower in Mary's hand in a sweet and tender manner.

His "Portrait of Ginevra de' Benci" (*c.* 1475–78) opened new paths for portrait painting with his singular linking of nearness and distance.

The emaciated body of his "St. Jerome" (*c.* 1480) is presented with realistic truth based on his sober and objective studies in anatomy; gesture and look give Jerome an unrivalled expression of transfigured sorrow.

The interplay of mimicry and gesture—"physical and spiritual motion," in Leonardo's words—is also the chief concern of his first large creation containing many figures, "The Adoration of the Magi" (1481). Never finished, the painting nevertheless affords rich insight into the master's subtle methods of work. The various aspects of the scene are "built up" from the base with very delicate, paper-thin layers of paint in chiaroscuro (the balance of light and shadow) relief. The main treatment of the Virgin and Child group and the secondary treatment of the surrounding groups are clearly set apart with a masterful sense of composition; yet thematically they are closely interconnected: the bearing and expression of the figures—most striking in the group of praying shepherds—depict all degrees and levels of profound amazement.

The two versions of "The Virgin of the Rocks"

"The Virgin of the Rocks" in its first version in the Louvre is the work that reveals Leonardo's painting art at its purest. The painting, according to Leonardo's first contract with the Confraternity of the Immaculate Conception, was to be the central panel of a large work for their chapel in the church of S. Francesco and was done in the years *c.* 1483–85. It never arrived, however, at the place it was originally destined for. It seems to have been prematurely taken from the Confraternity, perhaps by some highly placed interested party who removed it from Leonardo's workshop. Instead of this first painting, Leonardo and Ambrogio de Predis painted a second, slightly revised version, probably begun around 1494. This one gave rise to a 10-year litigation between the artist and the Confraternity regarding the price, a dispute that was not settled until 1506 in favour of Leonardo; whereupon, two years later, the painting was delivered as per contract. This second version remained in the chapel of S. Francesco until the Confraternity was dissolved (1781), and then, after changing owners frequently, it came finally in 1880 to the National Gallery in London.

"The Virgin of the Rocks" depicts the apocryphal legend of the meeting in the wilderness between the boy John and the equally young Jesus returning home from Egypt. Leonardo's artistry makes of this theme a vision that the true believer experiences when he contemplates the devotional picture. In the visionary character of the picture lies the secret of its effect: it presents not a "reality" but a "manifestation." Leonardo uses every artistic means at his disposal to emphasize the visionary nature of the scene. The soft colour tones (his famous sfumato), the dim light of the cave from which the figures emerge bathed in light, their quiet attitude, the meaningful gesture with which the angel (the only one facing the viewer) points to John as the intercessor between the Son of God and humanity—all this combines, in a patterned and formal way, to achieve an effect of the highest expressiveness.

The "Last Supper"

Leonardo's "Last Supper" is among the most famous paintings in the world. In its monumental simplicity, the composition of the scene is masterful; the power of its effect comes from the striking contrast in the attitudes of the 12 disciples as counterposed to Christ. Leonardo did not choose the portrayal of the traitor Judas customary in the iconographic tradition; he portrayed, rather, that moment of highest tension as related in the New Testament, "One of you which eateth with me will betray me." All of the Apostles—as human beings who do not understand what is about to occur—are agitated, whereas Christ alone, conscious of his divine mission, sits in lonely, transfigured serenity. Only one other being shares the secret knowledge: Judas, who is both part of and yet excluded from the movement of his companions; in this isolation he becomes the second lonely figure—the guilty one—of the company.

In the profound conception of his theme, in the perfect yet seemingly simple arrangement of the individuals, in the temperaments of the Apostles highlighted by gesture and mimicry, in the drama and at the same time the sublimity of the treatment, Leonardo attained a height of expression that has remained a model of its kind. Untold painters in succeeding generations, among them great masters such as Rubens and Rembrandt, marvelled at Leonardo's composition and were influenced by it. The painting also inspired some of Goethe's finest pages of descriptive prose. It has become widely known through countless reproductions and prints, the most important being those produced by Raffaello Morghen in 1800. Thus, the "Last Supper" has become part of humanity's common heritage and remains today one of the world's outstanding paintings.

Technical deficiencies in the execution of the work have not lessened its fame. Leonardo was uncertain about the technique he should use. He bypassed fresco painting, which, because it is executed on fresh plaster, demands quick and uninterrupted painting, in favour of another technique he had developed: tempera on a base mixed by himself on the stone wall. This procedure proved unsuccessful, inasmuch as the base soon began to be loosened from the wall. Damage appeared by the beginning of the 16th century, and deterioration soon set in. By the middle of the century the work was called a ruin. Later, inadequate attempts at restoration only aggravated the situation, and not until the most modern restoration techniques were applied after World War II was the process of decay halted.

In the Florence years between 1500 and 1506, four great creations appeared that confirmed and heightened Leonardo's fame: the "Virgin and Child with St. Anne" (Louvre); "Mona Lisa," "Battle of Anghiari," and "Leda." Even before it was completed, the "Virgin and Child with St. Anne" won the critical acclaim of the Florentines; the monumental plasticity of the group and the calculated effects of dynamism and tension in the composition made it a model that inspired Classicists and Mannerists in equal measure.

The "Mona Lisa"

The "Mona Lisa" became the ideal type of portrait, in which the features and symbolic overtones of the person painted achieved a complete synthesis. The young Raphael sketched the work in progress, and it served as a model for his "Portrait of Maddalena Doni." Similarly, the "Leda" became a model of the *figura serpentinata* ("sinuous figure")—that is, a figure built up from several intertwining views. It influenced such classical artists as Raphael, who drew it, but it had an equally strong effect on Mannerists such as Jacopo Pontormo.

In the "Battle of Anghiari" (1503–06) Leonardo's art of expression reached its high point. The preliminary drawings—many of which have been preserved—reveal Leonardo's lofty conception of the "science of painting"; the laws of equilibrium that he had probed in his studies in mechanics were put to artistic use in this painting. The "centre of gravity" lies in the group of flags fought for by all the horsemen. For a moment the intense and expanding movement of the swirl of riders seems frozen; this passing moment, the transition from one active movement to the next, is uniquely interpreted.

On the other hand, Leonardo's studies in anatomy and physiology influenced his representation of human and animal bodies, particularly when they were in a state of excitement. He studied and described extensively the baring of teeth and puffing of lips as signs of animal and human anger. On the painted canvas, rider and horse, their features distorted, are remarkably similar in expression.

The highly imaginative trappings take the event out of the sphere of the historical into a timeless realm. Thus, the "Battle of Anghiari" became the standard model for a cavalry battle. Its composition has influenced many painters: from Rubens in the 17th century, who made the most impressive copy of the scene from Leonardo's now-lost cartoon, to Delacroix in the 19th century.

After 1507—in Milan, Rome, and France—Leonardo did very little painting. He did resume work on the Leda theme during his years in Milan and sketched a variation, the "Kneeling Leda." The drawings he prepared—revealing examples of his late style—have a curious, enigmatic sensuality. Perhaps in Rome he began the "St. John the Baptist," which he completed in France. Bursting all the boundaries of usual painting tradition, he presented Christ's forerunner as the herald of a mystic oracle; his was an "art of expression" that seemed to strive consciously to bring out the hidden ambiguity of the theme.

Last manifestation of Leonardo's art of expression

The last manifestation of Leonardo's art of expression was in his "Visions of the End of the World," a series of pictorial sketches that took the end of the world as its theme. Here Leonardo's power of imagination—born of reason and fantasy—attained its highest level. The immaterial forces in the cosmos, invisible in themselves, appear in the material things they set in motion. What Leonardo had observed in the swirling of water and eddying of air, in the shape of a mountain boulder and in the growth of plants now assumed gigantic shape in cloud formations and rainstorms. The framework of the world splits asunder, but even its destruction occurs—as the monstrously "beautiful" forms of the unleashed elements show—in accordance with the self-same laws of order, harmony, and proportion that presided at its creation and that govern the life and death of every created thing in nature. Without any model, these "visions" are the last and most original expressions of Leonardo's art—an art in which his perception based on *saper vedere* seems to have come to fruition.

Sculpture. That Leonardo worked as a sculptor from his youth on is borne out by his own statements and those of other sources. In the introduction to his *Treatise on Painting* he gives painting precedence over sculpture in the hierarchy of the arts; yet he emphasizes that he practices both arts equally. A small group of generals' heads in marble and plaster, works of Verrocchio's followers, are sometimes linked with Leonardo because a lovely drawing on the same theme from his hand suggests such a connection. But the inferior quality of this group rules out an attribution to the master. Not a trace has remained of the heads of women and children that, according to Vasari, Leonardo modelled in clay in his youth.

The two great sculptural projects to which Leonardo devoted himself wholeheartedly stood under an unlucky star; neither the huge, bronze equestrian statue for Francesco Sforza, on which he worked until 1494, nor the monument for Marshal Trivulzio, on which he was busy in the years 1506–11, were brought to completion. Leonardo kept a detailed diary about his work on the Sforza horse; it came to light with the rediscovery of the Madrid MS. 8936. Text and drawings both show Leonardo's wide experience in the technique of bronze casting but at the same time reveal the almost utopian nature of the project. He wanted to cast the horse in a single piece, but the gigantic dimensions of the steed presented insurmountable technical problems. Indeed, Leonardo remained uncertain of the problem's solution to the very end.

The greatness of Leonardo's concept of sculpture

The drawings for these two monuments reveal the greatness of Leonardo's concept of sculpture. Exact studies of the anatomy, movement, and proportions of a live horse—Leonardo even seems to have thought of writing a treatise on the horse—preceded the sketches for the monuments. Leonardo pondered the merits of two types, the galloping or trotting horse, and in both cases decided in favour of the latter. These sketches, superior in the suppressed tension of horse and rider to the achievements of Donatello's Gattamelata and Verrocchio's Colleoni sculptures, are among the most beautiful and significant examples of Leonardo's art. Unquestionably—as ideas—they exerted a very strong influence on the development of equestrian statues in the 16th century.

A small bronze of a galloping horseman in Budapest is so close to Leonardo's style that, if not from his own hand, it must have been done under his immediate influence (perhaps by Giovanni Francesco Rustici). Rustici, according to Vasari, was Leonardo's zealous student and enjoyed his master's help in sculpting his large group in bronze of "St. John the Baptist Teaching" over the north door of the Baptistery in Florence. There are, indeed, discernible traces of Leonardo's influence in John's stance, with the unusual gesture of his upward pointing hand, and in the figure of the bald-headed Levite. Moreover, an echo of Leonardo's inspiration is unmistakable in the much-discussed and much-reviled wax bust of "Flora" in Berlin. It may have been made in France, perhaps in the circle of Rustici, who entered Francis I's service in 1528.

Architecture. Leonardo, who in a letter to Ludovico Sforza applying for service described himself as an experienced architect, military engineer, and hydraulic engineer, was concerned with architectural matters all his life. But his effectiveness was essentially limited to the role of an adviser. Only once—in the competition for the cupola of the Milan cathedral (1487–90)—did he actually consider personal participation; but he gave up this idea when the model he had submitted was returned to him. In other instances, his claim to being a practicing architect involved sketches for representative secular buildings: for the palace of a Milanese nobleman (around 1490), for the villa of the French governor in Milan (1507–08), and for the Medici residence in Florence (1515). Finally, there was his big project for the palace and garden of Romorantin in France (1517–19). Especially in this last named, Leonardo's pencil sketches clearly reveal his mastery of technical as well as artistic architectural problems; the view in perspective (at Windsor Castle) gives an idea of the magnificence of the site.

Activity as a military engineer

Leonardo was also quite active as a military engineer, beginning with the years of his stay in Milan. But no definite examples of his work can be adduced. Not until the discovery of the Madrid notebooks was it known that in 1504, sent probably by the Florence governing council, he stood at the side of the Lord of Piombino when the city's fortifications system was repaired and that Leonardo suggested a detailed plan for overhauling it. Finally, his studies for large-scale canal projects in the Arno region and in Lombardy show that he was also an expert in hydraulic engineering.

But what really characterizes Leonardo's architectural studies and makes them stand out is their comprehensiveness; they range far afield and embrace every type of building problem of his time. Furthermore, there frequently appears evidence of Leonardo's impulse to teach: he wanted to collect his writings on this theme in a theory of architecture. This treatise on architecture—the initial lines of which are in MS. B (Institut de France, Paris), a model book of the types of sacred and profane buildings—was to deal with the entire field of architecture as well as with the theory of forms and construction and was to include such items as urbanism, sacred and profane building, and a compendium of the important individual elements (for example, domes, steps, portals, and windows).

In the fullness and richness of their ideas, Leonardo's architectural studies offer an unusually wide-ranging insight into the architectural achievements of his epoch. Like a seismograph, his observations sensitively register all themes and problems. For almost 20 years he was associated with Bramante at the court of Milan and again met him in Rome in 1513–14; he was closely associated with such other distinguished architects as Francesco di Giorgio, Giuliano da Sangallo, Giovanni Antonio Amadeo, and Luca Fancelli. Thus, he was brought in closest touch with all of the most significant building undertakings of the time. Since Leonardo's architectural drawings extend over his whole life, they span precisely that developmentally crucial period—from the 1480s to the second decade of the 16th century—in which the principles of the classical style were formulated and came to maturity. That this genetic process can be followed in the ideas of one of the greatest men of the period lends Leonardo's studies their distinctive artistic value and their outstanding historical significance.

Science. *Science of painting.* Notwithstanding Leonardo's abundant scientific activity, one must never lose sight of the fact that it was the intellectual output of a man who proudly and consciously felt himself an artist throughout his life. And he described himself as such. He first came in contact with science as an artist, in the task he set himself of writing a treatise on painting.

Treatise on Painting

Leonardo's famous book on painting, in the form known and read today, is not an original work by the master but a compilation of texts from various manuscripts by Leonardo, collected and arranged with loving care by his disciple and heir, Francesco Melzi. It is the Codex Urbinas Latinus 1270, now in the Vatican Library. It was prepared around 1540–50, but from its form one can see that it was still an unfinished rather than a completed manuscript. Many original texts known to exist are missing; whole sections of Leonardo's overall plan are not included.

The first printed edition of the treatise in Melzi's version, omitting the long introductory chapter concerning the "pecking order" among the arts, appeared in a luxurious binding in 1651 in Paris, published by Raffaelo du Fresne with illustrations after drawings by Nicolas Poussin. The first complete edition of Melzi's text did not appear until 1817, published by Guglielmo Manzi in Rome. The two standard modern editions are that of Emil Ludwig, three volumes, Vienna, 1882 (with German translation); and that of A. Philip McMahon, Princeton, 1956, two volumes (facsimile of the Codex Urbinas and English translation).

Leonardo's plan envisaged a much broader treatment of the theme, as his own allusions to it indicate. For, in addition to detailed practical instructions for painting and drawing, the treatise was to deal with every area involving the artist's perception and experience, which he could then convey as acquired criteria. Three main problems form the keynote of the work: the definition of painting as a science, which is briefly outlined above; the theory of the mathematical basis of painting—that is, geometry, perspective, and optics—with the systematic study of light and shadow, colour, and aerial perspective; and the theory of forms and functions in organic and inorganic nature, as they are explained and made comprehensible to the painter trained in *saper vedere.* This theory of the forms and functions of the visible world sought first of all to describe the animal world, including man; next it sought to include the plant world; finally it endeavoured to explain how such phenomena of inorganic nature as water and earth, air and fire came into being.

Drawings for the *Treatise*

In the drawings for the *Treatise on Painting,* extending from the earliest Milan period to the final years of Leonardo's life in France, the progressive broadening and deepening of the theme can be followed. Many drawings were placed by the side of the text, and some of them were coloured; many studies of nature that are admired as art works, such as the famous rain landscape (Windsor Castle; 12409) or the "Foliage" (Royal Library, Windsor Castle; 12431), can be identified as illustrations for the treatise. Manuscript C in the Institut de France, Paris, with its diagrams of the blending of lights and shadows, likewise represents a segment of this textbook. Leonardo's so-called grotesque heads are also closely linked with the treatise. They have often been erroneously described as caricatures; but actually, for the most part, they represent types and only occasionally individuals. They are variations of the human face in its gradations between the poles of the beautiful and ugly, the normal and abnormal, the dignified and vulgar. They are also related to anatomical-physiological studies, in which old age—with wrinkled skin and bulging tendons—is contrasted with youth. Representation of the human being was to be treated at length: his body, his proportions, his organs and their functions but also his attitudes in physical and spiritual movement. Here Leonardo's artistic and scientific aims intertwine.

Anatomical studies and drawing. Leonardo's anatomical studies are perhaps the best way of revealing the process by which, in Leonardo's mind, an increasing differentiation set in among his diverse spheres of interest; but it was a differentiation in which the seemingly divergent areas of study—likewise on a higher level—always remained interrelated. Thus, Leonardo's study of anatomy, originally pursued for his training as an artist, quickly grew into an independent area of research. As his sharp eye uncovered the structure of the human body, Leonardo became fascinated by the *figura istrumentale dell' omo* ("man's instrumental figure"), and he sought to probe it and present it as a creation of nature. The early studies dealt chiefly with the skeleton and muscles; yet even at the outset Leonardo combined anatomical with physiological researches. From observing the static structure, Leonardo proceeded to study the functions exercised by the individual parts of the body as they bring into play the organism's mechanical activity. This led him finally to the study of the internal organs; among them he probed most deeply into the brain, heart, and lungs as the "motors" of the senses and of life. He did practical work in anatomy on the dissection table in Milan, then in the hospital of Sta. Maria Nuova in Florence, and again in Milan and Pavia, where he received counsel and inspiration from the physician-anatomist Marcantonio della Torre. By his own admission he dissected 30 corpses in his lifetime, thus acquiring an astonishing range of experience on his own. This experience was distilled in the famous anatomical drawings, which are among the most significant achievements of Renaissance science. These drawings, among his *dimostrazione,* are based on a curious connection between natural and abstract representation; sections in perspective, reproduction of muscles as "strings" or the indication of hidden parts by dotted lines, and finally a specifically devised hatching system enable him to represent any part of the body in transparent layers that afford an "insight" into the organ. Here Leonardo's mastery of drawing proved most useful. The genuine value of these *dimostrazione* and their superiority to descriptive words—as Leonardo proudly emphasized—lay in the fact that they were able to synthesize a multiplicity of individual experiences at the dissecting table and make the data immediately and accurately visible. The effect is unlike that of all dead anatomical preparations; in this way the "live quality" of the organism is retained.

Value of the *dimostrazione*

This great picture chart of the human body was what Leonardo envisaged as a *cosmografia del minor mondo* ("cosmography of the microcosm"). From the advanced portions that have survived, it is apparent how much and how long it occupied his mind. And it provided the basic principles for modern scientific illustration. Leonardo has not sufficiently received his due in this domain. Thanks to a method of seeing that was peculiarly his own, he elevated the art of drawing into a means of scientific investigation and teaching of the highest quality.

Mechanics and cosmology. With Leonardo, mechanics also proceeds from artistic practice, with which he became quite familiar as an architect and engineer. Throughout his life Leonardo was an inventive builder; he was thoroughly at home in the principles of mechanics of his epoch and contributed in many ways to advancing them.

His model book on the elementary theory of mechanics, which appeared in Milan at the end of the 1490s, was discovered in the Madrid Codex 8937. Its importance lay less in its description of specific machines or work tools than in its use of demonstration models to explain the basic mechanical principles and functions employed in building machinery. Leonardo was especially concerned with problems of friction and resistance. These elements—screw threads, gears, hydraulic jacks, swivelling devices, transmission gears, and the like—are described individually or in various combinations; and here, too, drawing takes precedence over the written word. As in his anatomical drawings, Leonardo develops definite principles of graphic representation—stylization, patterns, and diagrams—that guarantee a precise demonstration of the object in question.

Importance of primal mechanical forces to his thought

In the course of years his interest in pure mechanics merged increasingly with an interest in applied mechanics. Leonardo realized that the mechanical forces at work in the basic laws of mechanics operate everywhere in the organic and inorganic world. They determine animate and inanimate nature alike as well as man. Leonardo wrote on a page of his treatise on anatomy:

> See to it that the book of the principles of mechanics precedes the book of force and movement of man and the other living creatures, for only in that way will you be able to prove your statements.

So, finally, "force" became the key concept for Leonardo; as *virtù spirituale* ("spiritual property"), it shaped and ruled the cosmos.

Wherever Leonardo probed the phenomena of nature, he recognized the existence of primal mechanical forces that govern the shape and function of the universe: in his studies on the flight of birds, in which his youthful idea of the feasibility of a flying apparatus took shape and led to exhaustive research into the element of air; in his studies of water, the *vetturale della natura* ("conveyor of nature"), in which he was as much concerned with the physical properties of water as with its laws of motion and currents; in his researches on the laws of growth of plants and trees as well as the geological structure of earth and hill formations; and finally in his observation of air currents, which evoked the image of the flame of a candle or the picture of a wisp of cloud and smoke. In his drawings, especially in his studies of whirlpools, based on numerous experiments he undertook, Leonardo again found a stylized form of representation that was uniquely his own: this involved breaking down a phenomenon into its component parts—the traces of water or eddies of the whirlpool—yet at the same time preserving the total picture, analytic and synthetic vision.

Thus, for all the separate individual realms of his knowledge, Leonardo's science offered a unified picture of the world: a cosmogony based on *saper vedere.* Its final wisdom is that all the workings of nature are subject to a law of necessity and a law of order that the *Primo Motore,* the divine "Prime Mover," created. "Marvelous is Thy justice, O Prime Mover! Thou hast seen to it that no power lacks the order and value of your necessary governance."

Leonardo as artist-scientist. As the 15th century expired, Scholastic doctrines were in decline, and Humanistic scholarship was on the rise. Leonardo, however, was part of an intellectual circle that developed a third, specifically modern form of cognition. In his view the artist—as transmitter of the true and accurate data of experience acquired by visual observation—played a significant part. With this sense of the artist's high calling, Leonardo approached the vast realm of nature to probe its secrets. His utopian idea of transmitting in encyclopaedic form the knowledge thus won was still bound up with medieval Scholastic conceptions, but the results of his research were among the first great achievements of the thinking of the new age because they were based on the principle of experience in an absolutely new way and to an unprecedented degree.

Finally, Leonardo, although he made strenuous efforts to teach himself and become erudite in languages, natural science, mathematics, philosophy, and history, as a mere listing of the wide-ranging contents of his library demonstrates, remained an empiricist of visual observation. But precisely here—thanks to his genius—he developed his own "theory of knowledge," unique in its kind, in which art and science form a synthesis. In the face of the overall achievements of Leonardo's creative genius, the question of how much he finished or did not finish becomes pointless. The crux of the matter is his intellectual force—self-contained and inherent in every one of his creations. This force has remained constantly operative to the present day.

(L.H.H.)

MAJOR WORKS

PAINTINGS: "The Annunciation" (*c.* 1472–77; Uffizi, Florence); "The Annunciation" (*c.* 1472–77; Louvre, Paris); "Madonna with the Carnation" (*c.* 1474; Alte Pinakothek, Munich); "Portrait of Ginevra de' Benci" (*c.* 1475–78; National Gallery of Art, Washington, D.C.); "Madonna Benois" (1478–after 1500; Hermitage, Leningrad); "St. Jerome" (*c.* 1480; Vatican Museums, Rome); "The Adoration of the Magi" (1481; Uffizi); "The Virgin of the Rocks" (*c.* 1483–85; Louvre); "The Musician" (*c.* 1490; Pinacoteca Ambrosiana, Milan); "Lady with an Ermine" ("Cecilia Gallerani"; *c.* 1490; Muzeum Narodowe, Kraków, Poland); "The Virgin of the Rocks" (1494–1508; National Gallery, London); "Last Supper" (1495–97; Sta. Maria delle Grazie, Milan); decoration of the Sala delle Asse (1498; Castello Sforzesco, Milan); "The Virgin and Child with St. Anne" (cartoon, *c.* 1499; National Gallery, London); "Virgin and Child with St. Anne" (*c.* 1501–12; Louvre); "Mona Lisa" ("La Gioconda"; 1503–06; Louvre); "St. John the Baptist" (before 1517; Louvre). Lost: "Madonna with the Yarn-Winder" (1501; best copy in the Duke of Buccleuch Collection, Boughton, Kettering); "Leda" (1503–06; best copy at Galleria Borghese, Rome); "Battle of Anghiari" (1503–06; copy at Palazzo Vecchio, Florence).

DRAWINGS AND NOTEBOOKS: Main collections: Institut de France, Paris; British Museum; Uffizi, Florence; Biblioteca Ambrosiana, Milan; Accademia, Venice; Royal Library, Windsor Castle; Biblioteca Reale, Turin; Biblioteca Nacional, Madrid; Victoria and Albert Museum, London.

BIBLIOGRAPHY. ANGELA OTTINO DELLA CHIESA (ed.), *The Complete Paintings of Leonardo da Vinci* (1969; originally published in Italian, 1967), catalogs the paintings. The most informative account of Leonardo's workshop and pupils is W. SUIDA, *Leonardo und sein Kreis* (1929). ARTHUR E. POPHAM (ed.), *The Drawings of Leonardo da Vinci* (1945, reissued 1973), is important for the study of Leonardo as a draftsman; the standard publication on the drawings is KENNETH CLARK and CARLO PEDRETTI, *The Drawings of Leonardo da Vinci in the Collection of Her Majesty the Queen at Windsor Castle,* 2nd ed., 3 vol. (1968). A. MARIONI, "I manoscritti di Leonardo da Vinci," in *Leonardo: Saggi e richerche* (1954), is a concise summary of all manuscripts, their facsimile editions, and chronology. This collection contains other excellent essays by various authors on Leonardo as artist and scientist. The Madrid Codices are presented in LEONARDO DA VINCI, *The Madrid Codices,* 5 vol. (1974), of which vol. 1 and 2 contain the facsimiles, vol. 3 has commentary by LADISLAO RETI, and vol. 4 and 5 contain Reti's transcription and translation into English. See also two articles on the rediscovered codices by Reti in *Burlington Magazine,* vol. 110 (1968). Aspects of Leonardo's personality and creativity made evident in the Madrid Codices are discussed in LADISLAO RETI (ed.), *The Unknown Leonardo* (1974), 10 essays; RICHARD MCLANATHAN, *Images of the Universe: Leonardo da Vinci: The Artist as Scientist* (1966); and EMANUEL WINTERNITZ, *Leonardo da Vinci as a Musician* (1982). AMOS P. MCMAHON (ed.), *Treatise on Painting,* is a facsimile edition of *Codex Urbinas Latinus 1270,* accompanied by an English translation, 2 vol. (1956). The best anthologies of Leonardo's literary heritage are JEAN P. RICHTER, *The Literary Works of Leonardo da Vinci,* 3rd ed., 2 vol. (1970, reissued 1977); and EDWARD MCCURDY (ed.), *The Notebooks of Leonardo da Vinci,* 2 vol. (1955, reissued 1977). For additional sources of information, see ETTORE VERGA, *Bibliografia Vinciana, 1493–1930,* 2 vol. (1931, reprinted 1970); and *Raccolta Vinciana,* fasc. 1–20 (1905–64). Two institutions devoted exclusively to the study of Leonardo are the Raccolta Vinciana, Castello Sforzesco, Milan, and the Elmer Belt Library of Vinciana at the University of California at Los Angeles. *Leonardo da Vinci,* 2 vol. folio, ed. by the ISTITUTO GEOGRAFICO DE AGOSTINI, NOVARA (1964), contains numerous essays and is a richly illustrated compendium of Leonardo's artistic and scientific activity. The two standard publications on Leonardo sources are LUCA BELTRAMI, *Documenti e memorie riguardanti la vita e le opere di Leonardo da Vinci in ordine cronologico* (1919); and GEROLAMO CALVI, *I manoscritti di Leonardo da Vinci, dal punto di vista cronologico, storico e biografico* (1925). KENNETH D. KEELE and CARLO PEDRETTI (eds.), *Leonardo da Vinci: Corpus of the Anatomical Studies in the Collection of Her Majesty the Queen at Windsor Castle,* 3 vol. (1978–80), includes a volume of facsimile plates. A selection, with drawings, is found in EMERY KELEN (ed.), *Fantastic Tales, Strange Animals, Riddles, Jests, and Prophesies of Leonardo da Vinci* (1971). The following six monographs are among the most valuable of the countless studies: GABRIEL SEAILLES, *Léonardo de Vinci: l'artiste et le savant* (1892); WOLDEMAR VON SEIDLITZ, *Leonardo da Vinci,* rev. ed. (1935), accompanied by extensive documentation (in German); KENNETH CLARK, *Leonardo da Vinci: An Account of His Development as an Artist,* 2nd ed. (1952, reprinted 1980); LUDWIG H. HEYDENREICH, *Leonardo da Vinci,* 2 vol. (1954; originally published in German, 1953); VASILI P. ZUBOV, *Leonardo da Vinci* (1968; originally published in Russian, 1961); and PETER R. RITCHIE-CALDER, *Leonardo and the Age of the Eye* (1970). MORRIS PHILIPSON (ed.), *Leonardo da Vinci: Aspects of the Renaissance Genius* (1966), contains valuable contributions to the historical and psychological aspects of Leonardo. Other works include CHARLES D. O'MALLEY (ed.), *Leonardo's Legacy: An International Symposium* (1969), a collection of essays: CARLO PEDRETTI, *Leonardo da Vinci: The Royal Palace at Romorantin* (1972), and *Leonardo: A Study in Chronology and Style* (1973, reprinted 1982); CECIL GOULD, *Leonardo: The Artist and the Nonartist* (1975); ROBERT PAYNE, *Leonardo* (1978), an account of Leonardo's career, with several new interpretations; and MARTIN KEMP, *Leonardo da Vinci: The Marvelous Works of Nature and Man* (1981).

(L.H.H./Ed.)

Libraries

A library (from Latin *liber*, "book") is a collection of written, printed, or other graphic or visual material (including films, photographs, tapes, phonograph records, videodiscs, microforms, and computer programs) organized and maintained for reading, study, and consultation. In this article the general term books is used to denote collectively these various forms of the contents of libraries, except when a particular form, such as manuscript or film, needs to be specified.

The English word "library" has a long history, occurring in a prose translation of the Roman philosopher Boethius' *Consolation of Philosophy* that Geoffrey Chaucer made in about 1374. The word *librairie* in French (and its counterpart in other Romance languages) does not have the same meaning, being used to denote a bookshop or, by extension, a publisher; the word used in many other countries to signify a collection of books, public or private, is derived from a Latinized Greek word, *bibliotheca:* hence *bibliothèque* in French, *biblioteca* in Italian and Spanish, *Bibliothek* in German, *biblioteka* in Russian. In Japanese the word is *toshokan* ("building of books"). The use of the word library to denote a building, room, or set of rooms in which a collection of books is contained goes back to the 15th century.

Libraries may be roughly classified in two ways: by ownership or purpose (*e.g.*, national, county or municipal, university, research, school, industrial, club, private) or by subject content (*e.g.*, general or special, the latter including medical, legal, theological, scientific, engineering, music, etc.). General libraries frequently contain special collections or departments. Organization ranges from the complex system of a great library—with catalogs, indexes, and a large staff—to the simple arrangement or listing that may suffice for the owner of a small private library. Faced with an ever-increasing quantity of information in a variety of technical forms, modern libraries have developed highly sophisticated mechanical and electronic systems for processing and retrieval. These systems have both widened the librarian's horizon and complicated the organizational problems of the library. The result has been the increased emphasis on a variety of active information services among libraries of all types.

The history of these developments is discussed in this article as a part of the general history of libraries; further information on the application of the theory and technology of information science in libraries and related fields is covered in the article INFORMATION SCIENCE. (Ed.)

The article is divided into the following sections:

HISTORICAL PERSPECTIVES

The ancient world. In earliest times there was no distinction between a record room (or archive) and a library, and in this sense libraries can be said to have existed for almost as long as records have been kept. A temple in the Babylonian town of Nippur, dating from the first half of the 3rd millennium BC, was found to have a number of rooms filled with clay tablets, suggesting a well-stocked archive or library. Similar collections of Assyrian clay tablets of the 2nd millennium BC were found at Tell el-Amarna in Egypt. Ashurbanipal (668–627 BC), the last of the great kings of Assyria, maintained an archive of some 25,000 tablets, comprising transcripts and texts systematically collected from temples throughout his kingdom. The majority carry his name, and some state that the King had "edited" them and caused them to be collected in his palace at Nineveh.

Ashurbanipal's library

Many collections of records were destroyed in the course of wars or were purposely purged when rulers were replaced or when governments fell. In ancient China, for example, the emperor Shih huang-ti (Shi Huangdi), a member of the Ch'in (Qin) dynasty and ruler of the first unified Chinese empire, ordered that historical records other than those of the Ch'in be destroyed so that history might be seen to begin with his dynasty. Repression of history was lifted, however, under the Han dynasty, which succeeded the Ch'in in 206 BC; works of antiquity were recovered, and the writing of literature as well as record keeping were encouraged. Private libraries flourished in some later dynasties, and texts held in these libraries were more likely to survive periods of war than were those held in government collections. Eventually the proliferation of collections in government offices, schools, and monasteries helped to assure the survival—and to increase availability—of valuable records and texts.

Greece and Alexandria. In the West the idea of book collecting, and hence of libraries as they are now understood, had its origin in the classical world. Most of the larger Greek temples seem to have possessed libraries, even in quite early times; many certainly had archive repositories. The tragic playwright Euripides was known as a private collector of books, but the first important institutional libraries came in Athens during the 4th century BC with the great schools of philosophy. Their texts were written on perishable material such as papyrus and parchment, and much copying took place. The Stoics, having no property, owned no library; the schools of Plato and of the Epicureans did possess libraries, the influence of which lasted for many centuries. But the most famous collection was that of the Peripatetic school, founded by Aristotle and systematically organized by him with the intention of facilitating scientific research. A full edition of Aristotle's library was prepared from surviving texts by Andronicus of Rhodes and Tyrannion in Rome *c.* 60 BC. The texts had reached Rome as war booty carried off by Sulla when he sacked Athens in 86 BC.

Aristotle's library formed the basis, mainly by means of copies, of the library established at Alexandria, which became the greatest in antiquity. It was planned by Ptolemy I Soter (died 283 BC) and brought into being by his son Ptolemy II Philadelphus (308–246 BC) with the collaboration of Demetrius of Phaleron, their adviser. The founders of this library apparently aimed to collect the whole body of Greek literature in the best available copies, arranged in systematic order so as to form the basis of published commentaries. Its collections of papyrus and vellum scrolls are said to have numbered hundreds of thousands. Situated in a temple of the Muses called the Mouseion, it was staffed by many famous Greek writers and scholars, who included the grammarian and poet Callimachus (died 240 BC), Eratosthenes (died *c.* 194 BC), the philosopher Aristophanes of Byzantium (died *c.* 180 BC), and Aristarchus of Samothrace (died 145 BC), the foremost critical scholar of antiquity.

Pergamum. In Asia Minor, a library rivaling that of Alexandria was set up at Pergamum during the reigns of Attalus I (died 197 BC) and Eumenes II (died 159 BC). Parchment (*charta pergamena*) was said to have been developed there after the copying of books was impeded by Ptolemy Philadelphus' ban on the export of papyrus from Egypt. (Parchment proved to be more durable than papyrus and so marks a significant development in the impact of technical advances on the dissemination of knowledge.) The library was bequeathed with the whole of the kingdom of Pergamum to the Roman people in 133 BC, and Plutarch records an allegation that Antony gave its 200,000 volumes to Cleopatra, to become part of the Alexandrian library.

Rome. There were many private libraries in classical Rome, including that of Cicero. Indeed, it became highly fashionable to own a library, judging from the strictures of the moralizing statesman Seneca and the spiteful jibes by the poet Lucian on the uncultured "book clown." Excavations at both Rome and Herculaneum have revealed what were undoubtedly library rooms in private houses, one at Herculaneum being fitted with bookcases around the walls. A Roman statesman and general, Lucullus, who was reckoned one of the richest men in the Roman world at that time and was famous for his luxurious way of life, acquired as part of his war booty an enormous library, which he generously put at the disposal of those who were interested. His biographer, Plutarch, speaks appreciatively of the quality of his book collection, and Cicero tells of an episode when he visited the library to borrow a book to find his friend, Cato, already ensconced there surrounded by books of the Stoic philosophy.

The library planned by Julius Caesar

Julius Caesar planned the creation of a public library and entrusted the implementation of his plans to an outstanding scholar and writer, Marcus Terentius Varro, also the author of a treatise on libraries, *De bibliothecis* (which has not survived). Caesar died before his plans were carried out, but a public library was built within five years by the literary patron Asinius Pollio. Describing its foundation in his *Natural History,* Pliny coins a striking phrase which has application to libraries generally: *ingenia hominum rem publicam fecit* ("He made men's talents a public possession"). Libraries were also set up by Tiberius, Vespasian, Trajan, and many of the later emperors; the Bibliotheca Ulpia, which was established by Trajan in about AD 100 and continued until the 5th century, was also the Public Record Office of Rome.

Byzantium and Islām. In the East the library tradition was picked up at Constantinople. It was probably at Caesarea that Constantine the Great's order for 50 copies of the Christian Scriptures was carried out. Under Constantine himself, Julian, and Justinian, the imperial, patriarchal, and scholarly libraries at Constantinople amassed large collections; and their real significance is that for a thousand years they preserved, through generations of uncritical teachers, copyists, and editors, the treasures of the schools and libraries of Athens, Alexandria, and Asia Minor. Losses occurred, but these were mostly due to the habit, noticeable especially in the 9th century, of replacing original texts with epitomes, or summaries. By far the greater part of the Greek classics, however, was faithfully preserved and handed on to the schools and universities of western Europe, and for this a debt is owed to the great libraries and the rich private collections of Constantinople.

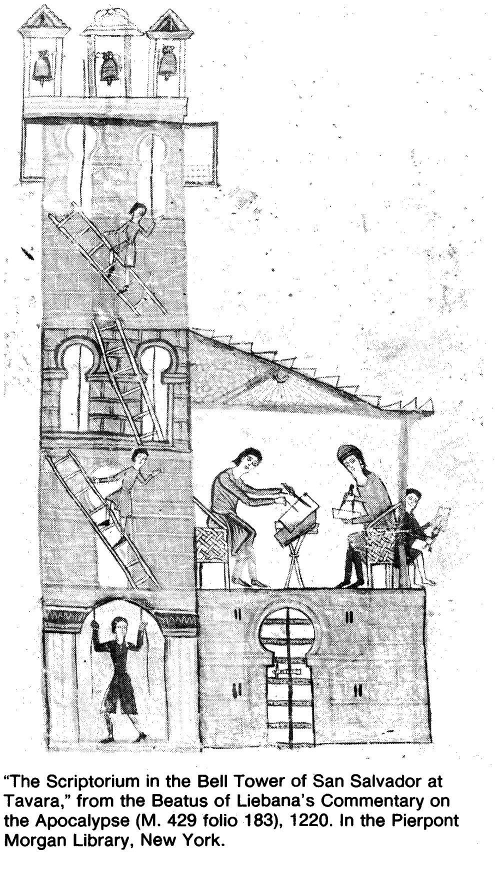
The links with classical Greece provided by Constantinople were supplemented by the libraries of Islām, which incorporated Arabic versions of Greek medical, mathematical, and scientific works, including those of Aristotle. From the 9th century richly equipped libraries existed throughout the Islāmic world, including Baghdad under Hārūn ar-Rashid, Cairo, Alexandria, and also Spain, where there was an elaborate system of public libraries centred on Córdoba, Toledo, and Granada. Arabic works from these libraries began to reach Western scholars in the 12th century, about the time that Greek works from Constantinople were filtering through to the West.

Middle Ages and Renaissance. *Role of the monasteries.* As monastic communities were set up (from as early as the 2nd century AD) books were found to be essential to the spiritual life. The rule laid down for observance by several monastic orders enjoined the use of books: that of the Benedictine Order, especially, recognized the importance of reading and study, making mention of a "library" and its use under the supervision of a precentor, one of whose duties was to issue the books and make an annual check of them. Scriptoria, the places where manuscripts were copied out, were a common feature of the monasteries—again, especially in those of the Benedictine Order, where there was a strict obligation to preserve manuscripts by copying them. Many—Monte Cassino (529) and Bobbio (614) in Italy; Luxeuil (*c.* 550) in France; Reichenau (724), Fulda (744), and Corvey (822) in Germany; Canterbury (597), Wearmouth (674), and Jarrow (681) in England—became famous for the production of copies. Rules were laid down for the use of books, and curses invoked against any person who made off with them. Books were, however, lent to other monasteries and even to the secular public against security. In this sense, the monasteries to some extent performed the function of public libraries.

Contents of monastic libraries

The contents of these monastic libraries consisted chiefly of the Scriptures, the writings of the early Church Fathers and commentaries on them, chronicles, histories such as Bede's *Ecclesiastical History of the English Peoples,* philo-

By courtesy of the Pierpont Morgan Library, New York



"The Scriptorium in the Bell Tower of San Salvador at Tavara," from the Beatus of Liebana's Commentary on the Apocalypse (M. 429 folio 183), 1220. In the Pierpont Morgan Library, New York.

sophical writings such as those of Anselm, Abelard, St. Thomas Aquinas, and Roger Bacon, and possibly some secular literature represented by the Roman poets Virgil and Horace and the orator Cicero. After the universities were founded, beginning in the 11th century, monkish students, on returning to their monasteries, deposited in the libraries there the lecture notes they had made on Aristotle and Plato, on law and medicine, and so forth, and in this way expanded their contents.

The new learning. The libraries of the newly founded universities—along with those of the monasteries—were the main centres for the study of books until the late Middle Ages; books were expensive and beyond the means of all but a few wealthy people. The 13th, 14th, and 15th centuries, however, saw the development of private book collections. Philip the Good, duke of Burgundy, and the French kings Louis IX and Charles V (who may be looked upon as the founder of the Bibliothèque du Roi [King's Library], which later became the Bibliothèque Nationale [National Library] in Paris) were great collectors, as were also such princes of the church as Richard de Bury, bishop of Durham (died 1345), who wrote a famous book in praise of books, *Philobiblon* (first printed in Cologne, 1473). But new cultural factors—including the growth of commerce, the new learning of the Renaissance (that was based on newly discovered classical texts), Gutenberg's invention of movable type for printing, and a substantial expansion of lay literacy—widened the circle of book collectors to include wealthy merchants whose libraries contained books of law and medicine, herbals, books of hours, and other devotional works. Italian humanists, such as Petrarch, searched for and copied manuscripts of classical writings to establish their scholarly libraries. Niccolò Niccoli (librarian to Cosimo de' Medici, the 14th-century ruler of Florence and a considerable patron of the arts) and Gian Francesco Poggio Bracciolini were two scholars who shared Petrarch's enthusiasm and ransacked Europe and the Middle East for manuscripts of the writers of Greece and Rome. Notable collections of books were made outside Italy, too (though Florence remained the centre of the rising book trade): by Diane de Poitiers, mistress of Henry II of France; by Jean Grolier, a high French official and diplomat, who was a great patron of bookbinders; by John Tiptoft, earl of Worcester; by Henry VII and Henry VIII of England and by many others.

Petrarch had wished to bequeath his collection to the municipality of Venice as a public library, but his intention was not realized. Cosimo de' Medici, however, set up, on the basis of Niccolò Niccoli's library, the Biblioteca Marciana in Florence in the convent of San Marco. The rich library of Lorenzo the Magnificent, grandson of Cosimo and an even greater patron of learning and the arts, also became a public library. It was opened in 1571 in a fine building designed by Michelangelo and still exists as the Biblioteca Laurenziana (though in 1808 it was amalgamated with the Marciana to form the Biblioteca Medicea-Laurenziana). Many other princely libraries were formed at this time, including that of Matthias I Corvinus of Hungary. Like the Medici, Corvinus employed agents to purchase manuscripts in the Levant, and he was also a customer of a well-known Florentine dealer in manuscripts, Vespasiano da Bisticci. Corvinus' library was destroyed when the Turks overran his capital of Buda in 1526 (a few books that survived are in the Austrian National Library at Vienna). The library of the Escorial in Madrid, founded in 1557, was based on the collections of Philip II. The Vatican library also dates its foundation from this time, the real founder of the collections as they are today having been Pope Nicholas V (reigned 1447–55), who dispatched agents to Germany, England, and Greece in search of manuscripts; Sixtus IV (reigned 1471–84) reassembled the Vatican collection in its present location.

Renaissance public libraries

Effects of the Reformation and religious wars. In England the end of the monastic libraries came in 1536–40 when the religious houses were suppressed by Henry VIII and their treasures dispersed. No organized steps were taken to preserve their libraries: it seems likely that a few—though only a very small proportion of the whole—were taken for the Royal Library. Even more wholesale destruction came in 1550: Henry VIII and Edward VI aligned with the "new learning" of the humanists; and university, church, and school libraries were purged of books embodying the "old learning" of the Middle Ages. The losses were incalculable. A change of attitude came about with Elizabeth's reign when those in authority realized that the kind of books that had been dispersed would form useful propaganda for the government's policy; the archbishop of Canterbury, Matthew Parker, and the Queen's principal adviser, William Cecil, took the lead in seeking out and acquiring the scattered manuscripts. Many other collectors were also active, including Sir Robert Cotton and Sir Thomas Bodley. As a result, some considerable portion of the libraries that had been scattered at the suppression was, by 1660, reassembled in collections—Parker's eventually went to Corpus Christi College at Cambridge; Cotton's to the British Museum; and Bodley's to form the Bodleian Library at Oxford—where they remain to this day.

Elsewhere in Europe, the period of the Reformation also saw many of the contents of monastic libraries destroyed, especially in Germany and the northern countries. The Reformation leader Martin Luther, however, did himself passionately believe in the value of libraries, and in a letter of 1524 to all German towns he insisted that neither pains nor money should be spared in setting up libraries, or book houses, particularly in the larger towns. As a consequence, many town libraries in Germany, including those at Hamburg (1529) and Augsburg (1537), date from this time. These, and the libraries of the newly created universities (such as Königsberg, Jena, and Marburg), were partly, at any rate, built up on the basis of the old monastic collections. In Denmark, similarly, some books from the churches and monasteries were incorporated with the new university library, though many were destroyed.

Libraries in Germany suffered severely in the Thirty Years' War. The Bibliotheca Palatina at the University of Heidelberg, for example, which had been founded in 1386, was taken as the spoil of war by Maximilian I of Bavaria, who offered it to Pope Gregory XV in 1623; and Gustavus Adolphus sent whole libraries to Sweden, most of them to swell the library of the University of Uppsala, which he had founded in 1620. The collections of the Royal Library in Stockholm were similarly enriched by the war booty that fell to Sweden during the reigns of Queen Christina and Charles X.

In France, Italy, Southern Germany, and Austria, where the Catholic faith remained unshaken, the old libraries remained and were supplemented by new ones set up for educational purposes by the Society of Jesus (the Jesuits).

17th and 18th centuries and the great national libraries. In the 17th and 18th centuries book collecting became more widespread. The motive sometimes was sheer ostentation, but often it was genuine love of scholarship. Several fine private collections were assembled, many of which were eventually to become the core of today's great national and state libraries—for this was also the period that saw new national and university collections springing up all over Europe.

There were, of course, other developments. In England were established a number of parish libraries, attached to churches and chiefly intended for the use of the clergy (one of the earliest, at Grantham in Lincolnshire, was set up as early as 1598, and some of its original chained books are still to be seen there). They were sometimes the result of lay donation: a Manchester merchant, Humphrey Chetham, left money in 1653 for the foundation of parish libraries in Bolton and Manchester and also for the establishment of a town library in Manchester (which still exists, housed in its original bookcases, in its original building). Later, in the 18th century, especially in England (though also elsewhere in Europe) and America, there was a great vogue for the circulating and subscription libraries—societies that provided reference service and lending collections for their members and had much influence on the formation of popular literary taste, especially in fiction.

Parish, town, and subscription libraries

Library planning. The private libraries of powerful and influential collectors such as Cardinal Mazarin in France were so large that a new approach to library organization was needed. The Escorial library in Madrid, erected in

1584, had been the first to do away with book bays on the medieval pattern and to arrange its collection in cases lining the walls. The old practice of chaining books to their cases was gradually abandoned; and the change to the present arrangement of standing them, spines facing outward, began in France, probably with the personal library of the lawyer, councillor of state, historian, and bibliophile Jacques-Auguste de Thou (died 1617). Mazarin's library was in the charge of Gabriel Naudé, who produced the first modern treatise on library economy, *Advis pour dresser une bibliothèque* (1627; *Instructions Concerning Erecting of a Library*). This work marked the transition to the age of modern library practice. One of its first fruits was the library of the diarist Samuel Pepys; in the last 14 years of his life Pepys devoted much time to the organization of his collection, and he left it to Magdalene College, Cambridge.

Naudé's concept of a scholarly library, systematically arranged, displaying the whole of recorded knowledge and open to all scholars, took root. It was above all absorbed by the philosopher Gottfried Wilhelm Leibniz (died 1716), a prominent librarian of his age, who conceived the idea of a national bibliographical organization that would provide the scholar with easy access to all that had been written on his subject. He realized that scientific progress is encouraged by communication among scholars and that an important purpose of great libraries and museums is to provide and keep open the channels of communication.

Emergence of national collections. The scope of scholarship and inquiry expanded rapidly during the 17th and 18th centuries, especially in the field of historical studies and in philosophy, resulting in a vast outpouring of books that went to swell the libraries of private collectors. In France, Jacques-Auguste de Thou, highly qualified as a collector, was made director in 1593 of the Bibliothèque du Roi (founded by Charles V and largely reorganized during the 15th century by Louis XII). Mazarin's library was scattered when he was compelled to leave France during the period of unrest known as the Fronde, but it was reassembled when he returned to power in 1661. Rehoused in a new building, it was opened to the public in 1691 and remained one of France's great libraries until after the French Revolution when it was incorporated with other collections (including the Bibliothèque du Roi) to form the Bibliothèque Nationale, today one of the world's great libraries. August, duke of Brunswick, established a library in 1604 that later became the Herzog August Bibliothek at Wolfenbüttel, one of the finest libraries in Europe (Leibniz was its librarian from 1690 to 1716). A library assembled by the elector Friedrich Wilhelm of Brandenburg was founded in 1659 and later became the Prussian State Library. The collections of the English book collectors Sir Hans Sloane, Sir Robert Cotton, and Edward and Robert Harley, earls of Oxford, formed the basis of the British Museum collection, founded in 1753, which from the start was regarded as the national library; it was further enlarged by the addition in 1757 of the Royal Library, containing books collected by the kings of England from Edward IV to George II.

The great national collections

By courtesy of the Master and Fellows of Magdalene College, Cambridge University

The library of Samuel Pepys at Magdalene College, Cambridge.

The effects of the French Revolution. On the continent of Europe the anticlerical movement that found expression in revolution sealed the fate of many monastic and church libraries: those in France, for example, were expropriated in 1789; in Germany in 1803; in Spain in 1835. The dispersal of books was, however, better organized than that resulting from the dissolution of the English monasteries under Henry VIII. In France books were collected in the main towns of the *départements* in what were called *dépots littéraires.* In 1792 the same fate befell the collections of aristocratic families, and these, too, were added to the *dépots.* The enormous accumulations caused problems, and many books were lost, but the plan of coordinating library resources throughout the country was carried out. The Bibliothèque Nationale received some 300,000 volumes; and new libraries were set up in many important provincial cities. Napoleon's successful campaigns throughout Europe also brought acquisitions, in the form of war booty, to the Bibliothèque Nationale, but most of the books were returned to their original owners after the Congress of Vienna in 1814–15. In Bavaria the state library was greatly enriched by the contents of over 150 confiscated libraries, and many of the provincial libraries were similarly enlarged. In Austria, as a result of confiscations, *Studienbibliotheken* (study libraries) were set up at Linz, Klagenfurt, and Salzburg, the university libraries at Graz and Innsbruck were substantially enlarged, and many valuable acquisitions accrued to the Hofbibliothek in Vienna.

Later developments. The difficulties of library management grew more noticeable in the 19th century. Libraries had grown in size, but their growth had been haphazard; administration had become weak, standards of service almost nonexistent; funds for acquisition tended to be inadequate; the post of librarian was often looked on as a part-time one; and cataloging was frequently in arrears and lacked proper method.

The university library at Göttingen was a notable exception. Johann Gesner, the first librarian, working in close association with the curator of the university, G.A. von Münchhausen, and proceeding on the principles laid down by Leibniz, made strenuous efforts to cover all departments of learning; the library provided good catalogs of carefully selected literature and was available to all as liberally as possible. The next director, C.G. Heyne, enthusiastically followed the same principles, with the result that Göttingen became the best organized library in the world.

By the middle of the century new conceptions of the purpose and scope of learned libraries were everywhere taking shape. Within a few years the picture of library service had been transformed, and a leading figure in this development was Antonio (later Sir Anthony) Panizzi, a political refugee from Italy who began working for the British Museum in 1831 and was its principal librarian from 1856 to 1866. From the start he revolutionized library administration, demonstrating that the books in a library should match its declared objectives and showing what these objectives should be in the case of a great national library. He perceived the importance of a good catalog and to this end elaborated a complete code of rules for catalogers. He also saw the potentiality of libraries in a modern community as instruments of study and research, available to all, and, by his planning of the British Museum reading room and its accompanying bookstacks, showed how this potential might be realized. His ideas have dominated library thought in the field of scholarly—or, as they are now called, research—libraries and have achieved major expression in the Library of Congress in Washington, D.C.

The revolution in library administration

By the middle of the 19th century the idea had been accepted that community libraries might be provided by local authorities at public expense. This proved a significant stage in the development of library provision. Panizzi had stated that he wanted the facilities of a great library to be available to poor students so that they could indulge their "learned curiosity"; in England in 1850 an act of Parliament was passed enabling local councils to levy a rate for the provision of free library facilities. From the first tentative beginnings there has been continuous growth in the providing of reading and other services and in the use that is made of them by the public.

KINDS OF LIBRARIES

Library services available throughout the world vary so much in detail from country to country that it is impossible to present anything but the most general picture of their activities. Nevertheless, they follow a broad but discernible pattern that has evolved over the years.

Aims of a national library

National libraries. In most countries there is a national or state library or group of libraries maintained by national resources, usually bearing responsibility for publishing a national bibliography and for maintaining a national bibliographical information centre. National libraries strive principally to collect and to preserve the nation's literature, though they try to be as international in the range of their collections as possible.

Most national libraries receive, by legal right, one free copy of each book and periodical printed in the country. Certain other libraries throughout the world share this privilege, though many of them receive their legal (or copyright) deposit only by requesting it.

The Bibliothèque Nationale in Paris, the British Library in London, the Library of Congress in Washington, D.C., and the Lenin Library in Moscow are the most famous and possibly the most important national libraries in the Western world. Their importance springs from the quality, size, and range of their collections, which are comprehensive in scope and attempt to maintain their comprehensiveness. This latter they achieve with diminishing success in view of the vastly increased number of publications that daily appear throughout the world, the failure of publishers to provide legal-deposit copies, and the difficulty of ensuring adequate representation of publications issued in the developing countries.

Bibliothèque Nationale. The Bibliothèque Nationale was before the Revolution known as the Bibliothèque du Roi and owes its origin (as is indicated above) to Charles V. It was the recipient during the 15th and 16th centuries of a number of important collections of manuscripts; in 1617, under the librarianship of the great collector de Thou, its right to legal deposit was reaffirmed and continued to be rigidly enforced. In the first quarter of the 18th century four of the library's departments (of prints, coins, printed books, and manuscripts) were created; it was opened to the public in 1735. Enormous additions accrued to the library as a result of the French Revolution and the confiscation of aristocratic and church private collections. The catalog of the library on cards was completed under the librarianship (1874–1905) of Léopold Delisle, and in 1897 he made a start to the task of compiling a printed catalog in volume form.

The present-day Bibliothèque Nationale plays a leading role in the French national library service. It houses the Direction des Bibliothèques, which oversees all public libraries, and participates in the training of library professionals. The library has undertaken the retrospective conversion of its catalog into machine-readable form.

The British Library. For more than two centuries the British Museum combined a great museum of antiquities with a great comprehensive library. The library was founded in 1753 by the acceptance of the bequest of the collections of Sir Hans Sloane, physician to King George II and president of the Royal Society. The library was built up on the basis of two other important collections, that of Sir Robert Cotton and that of Edward and Robert Harley, earls of Oxford; to these were added the Royal Library, given by George II in 1757. With this collection came also the right to legal deposit of one copy of every book published in the British Isles; this right is generally enforced, yet many titles arrive only slowly and some not at all. These four basic collections were notably enlarged during the first century of the library's history by the addition of many private collections, including the libraries of King George III (1823) and of Thomas Grenville (1846). Sir Anthony Panizzi reorganized the library; he was also responsible for its printed catalog, made between 1881 and 1905.

British Library Act of 1972

The British Museum Library was separated from the Museum under the British Library Act of 1972 and by July 1, 1973, was reorganized as the British Library Reference Division. The British Library Lending Division was formed from the amalgamation of two previously existing libraries: the National Central Library, which grew out of the Central Library for Students founded in 1916 and was the centre for interlibrary lending from 1927, and which had a collection of some 400,000 books and periodicals, mainly in the humanities and social sciences; and the National Lending Library for Science and Technology, which had been opened in 1962 by the Department of Scientific and Industrial Research. The Lending Division is located in Yorkshire and operates an extensive lending service through the mail.

The British Library Bibliographic Services Division was formed from the British National Bibliography Ltd., an independent organization set up in 1949 to publish a weekly catalog of books published in the United Kingdom and received at the British Museum by legal deposit. The *British National Bibliography,* as this weekly catalog was called, quickly established itself as a foremost reference work, both for book selection and cataloging and for reference retrieval. Since the reorganization of 1973 the division has continued and expanded the computerizing of current cataloging and the central provision of both printed cards and machine-readable entries. The BLAISE service (British Library Automated Information Service) offers a cataloging facility to any library wishing to participate, and the Bibliographic Services Division and its predecessor, the British National Bibliography, have cooperated closely with the U.S. Library of Congress in the Project for Machine-Readable Cataloging (MARC), which provides on-line access to the catalogs of the current acquisitions of the British Library Reference Division and the Library of Congress.

Library of Congress. The U.S. Library of Congress, located in Washington, D.C., probably is the largest of the national libraries, and its collection of modern books is particularly extensive. It was founded in 1800 but lost many of its books by fire during a bombardment of the Capitol by British troops in 1814. These losses were to some extent made good by the purchase of Thomas Jefferson's library shortly thereafter. The library remained a strictly congressional library for many years, but as the collections were notably enlarged by purchases and by additions under the copyright acts, the library became and remained—in effect, although not in law—the national library of the United States. The public has access to many of the collections.

The Library of Congress makes its catalog available to many thousands of subscribing American libraries and institutions. The service was begun by Herbert Putnam, librarian from 1889 to 1939. At the program's inception, in 1902, printed cards were used, and in the first year there were 212 subscribers.

The library's impact on librarianship has always been of the highest value. Through the Library of Congress Classification, the printed catalog cards, and the Project for Machine-Readable Cataloging (MARC; see below *Technical services*), the library's practices are widely followed. It publishes the *National Union Catalog,* its many editions totaling several hundred volumes and representing the stock of several thousand libraries. The library began producing most of the catalog on microfiche in 1983.

Lenin Library. Of a size and importance comparable to the Library of Congress, the Lenin Library of the U.S.S.R., in Moscow, is the national library of the Soviet Union. It receives several copies of all publications from the constituent republics of the Soviet Union and distributes copies to specialist libraries. It issues printed cards for the *Bibliography of Periodicals, 1917–1947* and for a cooperative catalog called *Books Published in the USSR, 1917–1945,* which lists the holdings of the Lenin Library, the

Saltykov-Shchedrin Public Library in Leningrad, the Library of the U.S.S.R. Academy of Sciences in Leningrad, and the Central Book Office. It has produced the Soviet Library–Bibliographical Classification scheme based on a Marxist–Leninist classification of knowledge. It organizes domestic and international lending and exchanges and offers courses of lectures for professional education and also for readers.

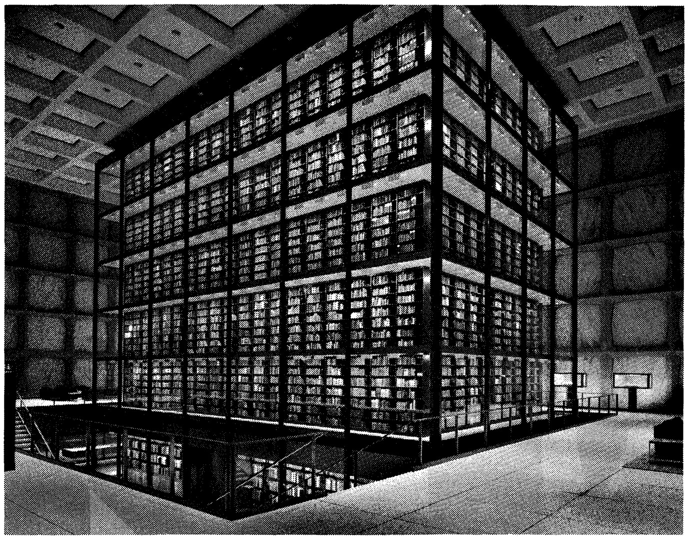
Other national collections. There are many other national libraries with important collections and very long histories. The Bibliothèque Royale Albert I, in Brussels, founded in 1837 and centred on the 15th-century collection of the dukes of Burgundy, is the national library of Belgium and the centre of the country's library network; it maintains a regular lending service with the university libraries and with the large town library of Antwerp. The Dutch Royal Library in The Hague was founded in 1798, and it, too, is the centre of a well-developed interlibrary loan system. Because the unification of Italy in the 19th century brought together many city-states that had major libraries, the country has a number of national libraries, the chief being the Biblioteca Nazionale Centrale Vittorio Emanuele II in Rome, founded in 1875, and the historically richer Biblioteca Nazionale Centrale at Florence, founded in 1747. Other Italian national libraries are at Milan, Naples, Palermo, Turin, and Venice. Germany was equally remarkable before World War II both for the importance of its state or provincial libraries and for the lack of a recognized national library. The former Preussische Staatsbibliothek was given national status in 1919; after World War II, under the name Deutsche Staatsbibliothek, it became the national library of East Germany. West Germany has no national library as such, the functions of such a library being performed by the Deutsche Bibliothek in Frankfurt and the Staatsbibliothek Preussischer Kulturbesitz in West Berlin. The Austrian National Library, founded by the emperor Maximilian I in 1493, has rich collections—notably of manuscripts from the Austrian monasteries and from the library of Matthias I Corvinus, dispersed after the capture of his capital, Buda, by the Turks in 1526. The National Library of Australia in Canberra, formally created by legislation in 1960, grew out of the Commonwealth Parliamentary Library, established in 1901.

National libraries in Asia

Many nations in Asia have national libraries, some of them—including the National Library of Peking and the National Diet Library in Tokyo—with holdings of more than 2,000,000 volumes. The National Library of India (formerly the Imperial Library) in Calcutta was founded in 1903. In some countries, such as Iceland, Norway, and Israel, the national library is combined with a university library. In Nigeria the library of the University of Ibadan served as the national library until the national library in Lagos was founded.

University and research libraries. Before the invention of printing, it was common for students to travel long distances to hear famous teachers. Printing made it possible for copies of a teacher's lectures to be widely disseminated, and from that point universities began to create great libraries. The Bodleian Library at Oxford University and Harvard University Library at Cambridge, Mass., are superior to many national libraries in size and quality. The collections of the libraries of the University of California together nearly equal those of the Library of Congress. In addition to a large central library, often spoken of as the heart of a university, there are often smaller, specialized collections in separate colleges and institutes. The academies of science in the Soviet Union and the countries of eastern Europe consist of groups of these specialized institutes, and while not all act as universities in awarding degrees, their research function has the same importance. Some, as in Hungary and Romania, serve as the national library.

In a university library many users may seek the same books at the same time. The difficulty of providing multiple copies has vexed most university librarians, who must balance slender resources against sometimes vociferous demand. To handle the problem, many libraries have set up a short-loan collection from which books may be borrowed for as little as a few hours. The use of a microcomputer for short-loan records has brought some relief through its great flexibility of operation and capacity for instant recall of information on the whereabouts of a particular work.



Glass-enclosed bookstack of the Beinecke Rare Book and Manuscript Library at Yale University, New Haven, Conn.

Ezra Stoller © ESTO

Cooperation among libraries

The range of research carried out at a traditional university may encompass every aspect of every discipline, and even the largest university libraries have long recognized the need for cooperation with others, first in cataloging and later in acquisitions. Automation has helped, too, by making it possible for readers in one library to consult the catalogs of others by means of computer networks. Previously, much effort was expended in printing volumes of union catalogs, especially for periodicals, and although these were costly to maintain, they proved the value to scholars of sharing information on catalogs and collections. Many universities have published catalogs of their special collections and have arranged for the production of microfilms both of rare individual works and of complete collections. An example is the Goldsmiths'–Kress collection of early works in economics, which combines the holdings of the Goldsmiths' Library at the University of London and the Kress Library at Harvard.

Role of public libraries

Public libraries. Public libraries are now acknowledged to be an indispensable part of community life as promoters of literacy, providers of a wide range of reading for all ages, and centres for community information services. Yet although the practice of opening libraries to the public has been known from ancient times, it was not without considerable opposition that the idea became accepted, in the 19th century, that a library's provision was a legitimate charge on public funds. It required legislation to enable local authorities to devote funds to this cause.

Public libraries now provide well-stocked reference libraries and wide-ranging loan services based on systems of branch libraries. They are further supplemented by traveling libraries, which serve outlying districts. Special facilities may be provided for the old and the disabled, and in many cases library services are organized for local schools, hospitals, and jails. The services provided vary in proportion to the size of the municipality or the area covered. In the case of very large municipalities, library provision may be on a considerable scale, including a reference library, which has many of the features associated with large research libraries. The New York Public Library, for example, has rich collections in many research fields; and Boston Public Library, the first of the great city public libraries in the United States (and the first to be supported by direct public taxation), has had from the first a twofold character as a library for scholarly research as well as for general reading. In the United Kingdom, the first tax-supported public libraries were set up in 1850; they provide a highly significant part of the country's total national library service. The importance of public library activities has been recognized in many countries by legislation designed to ensure that good library services are available to all without charge.

In many cases, public libraries build up collections that relate to local interests, often providing information for local industry and commerce. It is becoming more and more usual for public libraries to provide, for home loan, music scores, phonograph records, and, in some countries—notably in Sweden and the United Kingdom—original works of art for enjoyment, against a deposit, in the home.

Not all countries provide public library services of an equally high standard, but there has been a tendency everywhere to recognize their value and to improve services where they exist or to introduce them where they do not. Public librarians work strenuously, through such organizations as the International Federation of Library Associations, for such developments.

Some public libraries have begun to promote a program called public lending right, under which authors of books lent from public libraries receive a fee for each loan. The programs are popular in small countries with minority languages, where writers have little hope of earning their living from royalties on sales. The principle has been widely accepted, despite fierce opposition, even in countries with well-established book trade and the advantages of an international language.

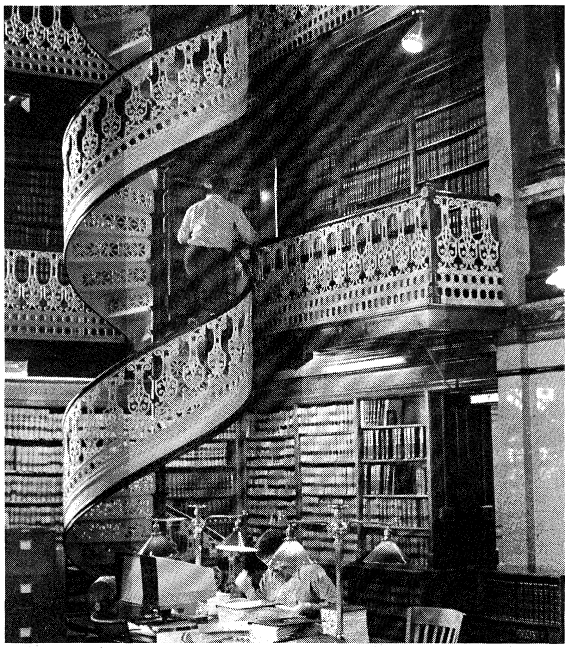
Special libraries. The national, university, and public libraries form the network of general libraries more or less accessible to the general public. The libraries take pride in special collections, which are built around a special subject interest. Outside of the network are a large number of libraries established by special groups of users to meet their own needs. Many of these originated with learned societies and especially with the great scientific and engineering societies founded during the 19th century to provide specialist material for their members. Thus some special libraries were founded independently of public libraries and before major scientific departments were developed in national libraries. For example, the National Reference Library of Science and Invention, now the Science Reference Library and part of the British Library, was originally established at the U.K. Patent Office.

With the coming of the Industrial Revolution arose the need for a working class educated in technology, and industrialists and philanthropists provided facilities and books of elementary technical instruction. In the United Kingdom the Mechanics' Institutes were founded in the rapidly growing industrial towns to provide books and lectures to workers and tradesmen at prices lower than those of the subscription libraries.

Functions and services of specialist libraries

Special libraries are often attached to official institutions such as government departments, hospitals, museums, and the like. For the most part, however, they come into being to meet specific needs in commercial and industrial organizations. They are planned on strictly practical lines, with activities and collections carefully controlled in size and scope, though they may be and often are large and wide-ranging in their activities; they cooperate widely with other libraries. They are largely concerned with communicating information to specialist users, in response to, or preferably in anticipation of, their needs. They have therefore been much concerned with the theoretical investigation of information techniques, including the use of computers for indexing and retrieval. It was in this area that the concept of a science of information flow and transfer emerged as a new field of fundamental theoretical study. The concept underpins the practices not only of special libraries but of all types of library and information services.

School libraries. Where public libraries and schools are provided by the same education authority, the public library service may include a school department, which takes care of all routine procedures, including purchase, processing with labels, and attaching book cards and protective covers; the books are sent to the schools ready for use. This is done in Denmark and in some parts of the United Kingdom. In other countries, the United States for example, processing may be contracted out to a specialist supplier. In the Soviet Union the All-Union Book Chamber prepares bibliographical data for distribution to all school libraries in the country. Everywhere, in fact, school and public libraries cooperate closely.



The law library serving the special needs of state legislators in Des Moines, the capital of Iowa.

Erich Hartmann—Magnum

Teachers who take an interest in the school library make a considerable contribution to its progress, and many have acquired qualifications in librarianship, recognizing that a modern library requires full-time attention and a variety of skills. Like the special librarian, the school librarian must have a close knowledge of and sympathy with the work of the teaching staff; and school libraries have been the scene of significant research and experiment with many different media, so much so that some school libraries have been promoted to become resource centres. Teachers accustomed to using visual aids, often indeed to making their own, have come to expect the library to provide collections of photographs, slides, films and filmstrips, and artifacts for work in subjects such as history and mathematics. Some school librarians use the term "realia" to describe these resources.

Private libraries. The libraries owned by private individuals are as varied in their range of interest as the individuals who collected them, and so they do not lend themselves to generalized treatment. The phrase private library is anyway unfortunate because it gives little idea of the public importance such libraries may have. Private collectors are often able to collect in depth on a subject to a degree usually impossible for a public institution; being known to booksellers and other collectors, they are likely to be given early information about books of interest to them; they can also give close attention to the condition of the books they buy. In these ways they add greatly to the sum of bibliographical knowledge (especially if they make their collections available to scholars).

Henry Clay Folger, for example, collected no fewer than 70 copies of one book—the first collected edition of Shakespeare's plays. (In 1932 he opened the Folger Shakespeare Library in Washington, D.C., which had been built to house his collection.) As a result of his collecting he added greatly to the sum of knowledge about the printing of Shakespeare's plays and about 17th-century printing in general. Collectors of private libraries have sometimes benefited posterity by leaving their collections to public institutions or founding a library. Examples in the United States include Henry E. Huntington, John Carter Brown, William L. Clements, and J.P. Morgan. The tradition has long been established in Europe, where many important libraries have been built up around the nucleus of a private collection.

The main reading room in the Folger Shakespeare Library, Washington, D.C., a private library established by Henry Clay Folger in 1932.

© Peter Aaron/ESTO

Subscription libraries. Part public, part private, these libraries enjoyed much popularity from the late 17th to the 19th century. Many of them were set up by associations of scholarly professional groups for the benefit of academies, colleges, and institutions; but their membership was also open to the general public. Some of them are still in existence: perhaps the most famous are the Library Company of Philadelphia, founded by Benjamin Franklin in 1731; the Boston Athenaeum, founded in 1807; and the London Library, opened largely at the instance of Thomas Carlyle in 1841, which today has a wide-ranging collection for loan to its members in their homes.

Famous subscription libraries

During the 19th century, the great size of many subscription libraries enabled them to wield much influence over publishers and authors: Mudie's Circulating Library, for instance, established in London in 1842, would account for the sale of as much as 75 percent of a popular novel's edition. Nevertheless, these libraries were for the most part unable to survive, and the service they gave is now largely provided by the free public libraries.

Archives. Archives are created in the course of conducting business activities of a public or private body; they are accumulations of documentary material. Originally, such records were not distinguished from library materials and were preserved in the same places as other manuscripts, up until the mid-15th century and the invention of printing. The importance nowadays accorded to public records has been recognized as one outcome of the French Revolution, when for the first time an independent national system of archive administration was set up, for whose preservation and maintenance the state was responsible and to which there was public access.

The science of archive administration embraces the study of records management, records appraisal, accessioning and arrangement, archival buildings and storage facilities, preservation and rehabilitation, and reference services, including exhibition and publication. Academic courses offering instruction in archive work are available in many parts of the world.

While the administration of archives shares with libraries the basic obligation to collect, to preserve, and to make available, it has to employ different principles and management techniques. Libraries might be described as collecting agencies, whereas archival institutions are receiving agencies: they do not select—their function is to preserve documents as organic bodies of documentation. They must respect the integrity of these bodies of documents and maintain as far as possible the order in which they were created. And, of course, the documents need catalogs and finding aids, or guides.

A distinction has to be drawn between public and private archives. Every state, broadly speaking, now recognizes the need to preserve its own official records and is expected to maintain a system of archive administration, which has the function of collecting them, preserving them, and making them publicly available after the appropriate lapse of time. Among the best known are the Archives Nationales in France, the U.S. National Archives, and the British Public Record Office. Nonofficial archives—the records of the day-to-day activities of an institution or a business—are now recognized as having great value for socioeconomic history, and they are frequently sought by libraries for their historical value and preserved in manuscript and similar collections. It is the practice of many institutions such as universities, professional and commercial organizations, and ecclesiastical establishments to set up their own archive departments. (F.C.F./D.J.F.)

Distinction between public and private archives

LIBRARY ARCHITECTURE

Function and design. The basic function of a library building is to house the library's collections, to provide adequate space for staff administration and procedures, and to offer acceptable accommodation in which the collections can be used. The form given to the library is undoubtedly of great importance, but its success will in the long run be measured by its ability to meet the basic functional requirements. These requirements have to be interpreted differently according to the objectives and needs of different types of libraries, whether large research, university, public, or special.

In planning a research or university library, due regard to the following considerations is essential: (1) The space set aside for the collections must not only be adequate for immediate needs but also allow for planned growth over a reasonable period of time. (2) Adequate provision should be made to ensure the preservation and safety of the collections, such as proper temperature and humidity control, air conditioning, and a building plan calculated to provide security against misuse. (3) Provision should be made for the convenience and comfort of the readers and of the library staff, including the avoidance of unnecessary and time-wasting movement within the building.

The same strictures apply, on the whole, to the planning of other kinds of libraries. A good public library would pay more attention to providing an attractive exterior, a convenient layout, and pleasing conditions for users so that they can browse at leisure and discuss their reading and information requirements comfortably with the staff employed for that purpose. The functional emphasis in a special library would be on ease and speed of consultation.

The final form to be aimed at in all these cases is a library building both functionally and architecturally distinguished. Hence, the closest collaboration is essential between librarian and architect, the former supplying as complete a brief as possible (after careful discussion with library colleagues) and the latter endeavouring to understand the librarian's brief and interpret it in building terms. Difficulties and misconceptions are removed by close and constant discussion. Cooperation of this kind has become normal practice only in recent years; indeed, it is only in recent years that the concept of a fully documented brief prepared by the librarian has been accepted or even understood by architects and librarians alike. In the early days, libraries were often associated with temples; and it is still common to look upon a library as a central feature in a municipal or university complex, hence the temptation to erect a "prestigious" building, which may and often does subordinate function to appearance.

Essential collaboration between architect and librarian

Historical developments. The clay-tablet rooms in the temple at Nippur in Babylonia may be looked upon as functionally designed for their purpose, as was an "archive room" excavated at Herculaneum; and, in the cloister of the early monasteries, the *armarium,* or book cupboard, was appropriately placed for its purpose in the well-lighted cloister. Monastery and cathedral libraries of the later medieval period employed the chained library stall system and were designed to meet the readers' need for well-lit workspace. Each consisted of a long room that ran the length of the building, usually on the first upper floor, lighted on both sides with rows of windows. Between the windows and at right angles to the walls were rows of back-to-back presses, or cupboards, with two or three fixed shelves on each side and with desks attached to the presses at the level of the lowest shelf; the books were chained to a long bar running the width of the presses by chains just long enough to enable them to lie open on the desk. When not in use, the books, usually at the time bulky folios and quartos, stood on the shelves, with their front edges facing outward so that the chains, which were normally attached to the front edge of the upper or lower cover, should not damage the binding of the other books on the shelves. A centre gangway ran the length of the room between the presses; and in the alcoves formed by the presses there were fixed benches for readers, who were thus able to get at the books on the desks and have enough light from the windows to read by.

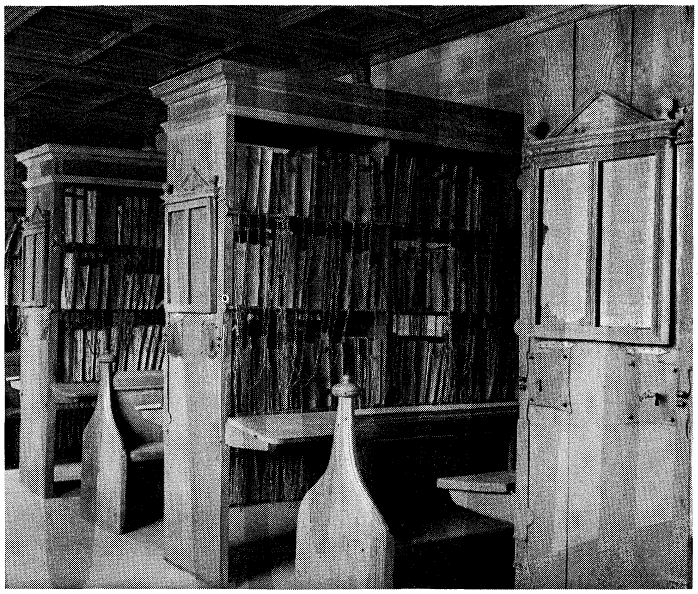
17th and 18th centuries. The "long room" was retained even in 17th- and 18th-century libraries, a reminder of the older system. The abandonment of chaining, however, coupled with a change in the pattern of publication—marked by decreasing cost and greater numbers of publications, particularly in smaller sizes—made possible the adoption of wall shelving. Wall shelving, in turn, made possible the greater use of the height of a building—the higher the walls, the greater the number of books that could be shelved. Besides this, the disappearance of chains made it possible for books to be taken away from the shelves for reading, and seats could be disposed as desired; the walls thus left unencumbered with desks and benches could be experimented with, and the main floor area could be left empty or provide an elegant setting for statuary or exhibits of other kinds, such as coin cabinets. Sometimes spaciousness and display were more highly prized than convenience for study. An example of this architectural emphasis is to be found in the Prunksaal in the Austrian National Library in Vienna, erected in 1723–26 after a plan by Johann Fischer von Erlach, which consists of a magnificent Baroque hall, decorated with marble Corinthian columns; a statue of Charles VI stands in the middle and other statues of princes of the empire are placed along the walls. A number of the libraries in the Oxford and Cambridge colleges—the Wren Library at Trinity College, Cambridge, and that at Christ Church, Oxford—also provide excellent examples. Perhaps the last great flourish of library design of this kind was the King's Library in the British Museum, built between 1823 and 1826 to house the library of King George III, acquired by the museum in 1823. It consists of a room 300 feet (about 100 metres) in length with a gallery; the walls are lined with bookshelves from floor to ceiling.

Wall shelving

British Museum reading room. A complete innovation in functional planning was made with the celebrated reading room in the British Museum. By the middle of the 19th century, readers were too numerous for the space provided for them in the existing reading rooms. In 1854 Sir Anthony Panizzi, by a brilliant application of the "engineering age" to library construction, planned a bookstack in cast iron with exterior enclosing walls and widely spaced brick piers, to hold some 1,500,000 volumes; it was designed to surround and support a huge, circular reading room whose radiating rows of reading desks with places for 450 readers naturally followed the circular plan, facilitating supervision, and whose walls provided space for an open access general reference library of 25,000 volumes. The "iron library," which provided an extensive library in close proximity to the reading area, reduced the risk of fire and demonstrated immense saving of space by eliminating brick supports and substituting metal. Modern stack construction has greatly improved on it, substituting steel for cast iron and generally economizing on space; but the basic plan of this stack was followed in subsequent stacks.

Bookstacks and reading rooms

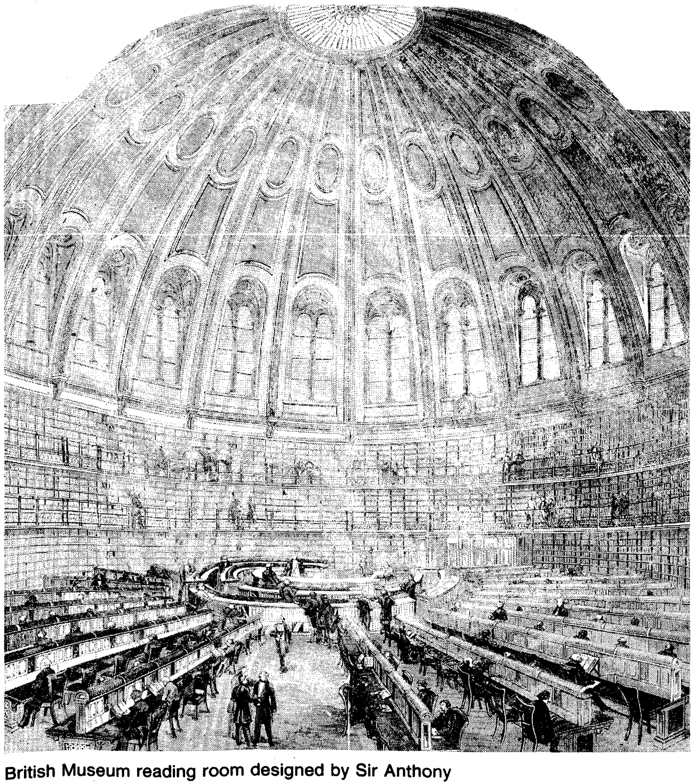
The modern period. Panizzi's creation focused attention once more on functionalism in library planning. From that time up to World War II, the design of libraries was conditioned largely by architects' interpretation of the functions to be performed by the library, and their buildings rarely gave any indication of what was actually performed within them. General style tended toward the monumental, as befitted (so it seemed to them) a building devoted to noble and scholarly pursuits. The librarian had little say—indeed, as a rule probably had little to say—on the design of the building, except to indicate the proportion of space to be allotted to bookstacks, reading rooms, and administrative quarters.


The chained library of the Cathedral Church of the Blessed Virgin Mary and St. Aethelberht, Hereford, England.

By courtesy of the Dean and Chapter of Hereford Cathedral, England; photograph, F.C. Morgan

A specific part of the building was set aside for bookstacks, and, not surprisingly in view of the constant growth in library collections, the stack frequently came to be designed as a tower. Reading rooms were treated as separate architectural units, in many cases, related neither in proportion nor access to the bookstacks. Finally, the

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.


British Museum reading room designed by Sir Anthony Panizzi, 1854. Illustration by Sydney Smirke, from *The Illustrated London News,* 1857.

space provided for administrative purposes and the routine library procedures (such as control and recording of acquisitions, cataloging, binding, and so on) was frequently planned without properly appreciating the importance of a relationship between bookstacks, reading rooms, and public. Fixed function planning produced buildings with little possibility of the flexible use of space and as a rule with only limited, if any, possibility of expansion. Much frustration was caused to the librarian, who had to put up with (and provide) light and heat for unnecessarily high reading rooms, spacious corridors, and, sometimes, monumental halls and staircases. The reader was expected to climb unnecessary stairs and to walk long distances from entrance to reading room and often from the catalog and the reference shelves back to a reading place.

Access to the stacks by readers was rarely provided in the large research libraries (nor is it always desirable that it should be). Public libraries, however, though they too were often housed in monumental buildings, found it possible to introduce open access to their stock of books very early on (in England during the 1890s). Open access made bookstacks with high shelves impracticable for readers, and it encouraged the use of classified arrangement of books so that readers could easily locate the subjects and the volumes of special interest to them.

A remarkable change of attitude toward the design and construction of library buildings took place following World War II. It is of considerable interest for the history of libraries and their development to consider what the reasons for this change may have been. Scientists, technologists, and others had, during the war, needed precise information, speedily supplied; there was undoubtedly much dissatisfaction with the traditional library services and hence a more critical attitude toward libraries and their characteristic features. Afterward, the needs of users and their convenience were given greater consideration and more careful regard. There were also many more people using libraries—more students in a greater number of universities and colleges and more members of the general public looking to libraries for information, instruction, and recreation. Significant improvements in building techniques, moreover, made new building designs and methods possible, whereas very many new libraries were needed, partly because of the interruption of building during the war, partly because many libraries in Europe and other parts of the world had been destroyed or damaged. The old kinds of library, with their monumental and inflexible structure, were no longer acceptable.

The modular system of building

The new look at library design, thus prompted, has been further encouraged by development of the modular system of building. In the modular building, the basic floor area is divided into equal rectangles defined by structural columns at the corners, which are the only weight-bearing structures within the building. Subdivisions of these areas can be created by non-weight-bearing walls, bookstacks, and freestanding furniture; it follows, in theory, that nothing within the building is fixed and immovable except the columns (and of course stairways, elevators, heating facilities, ducts, and plumbing) so that the use made of an area can be extended or modified at will. (F.C.F./Ed.)

LIBRARY MATERIALS

Historically library services have depended on what materials were available to build collections. The evolution of libraries in antiquity involved the search for a material durable enough to survive as a permanent record and relatively easy to use. Clay and stone provided permanence, but inscribing the records required considerable labour. Palm leaves, bamboo strips, and papyrus offered a flat surface that more readily accepted handwriting, and it was said that parchment came into use in Asia Minor after the export of papyrus from Egypt was banned. In about AD 105 the invention of paper was announced by Ts'ai Lun to the Chinese emperor Ho Ti, and the British Museum has a paper fragment dated about 137. The use of paper spread slowly, however, and most of the oldest surviving manuscripts are of other materials, particularly vellum.

Samples of ancient writing are rare and therefore are highly valued, and national and other scholarly libraries collect and preserve them as part of their responsibility to the preservation of history and the advancement of learning. Most universities have collections of rare books. Eton College, for example, has a fine collection of incunabula, some purchased when they were first printed. A Gutenburg Bible is one of its finest examples. Some, such as the Duke Humphrey Library in the Bodleian at Oxford and the Beinecke Library at Yale, contain collections of manuscripts, and wealthy private collectors have established world-famous institutions such as the Henry E. Huntington Library in San Marino, Calif., the Folger Shakespeare Library in Washington, D.C., and the Cotton and Harley collections in the British Library Reference Division.

Uses of roll film and microfiche

The invention of photography in the 19th century made possible a new kind of record, and collections of photographs are popular, particularly in public libraries with an interest in local history. Specialized picture libraries, such as the BBC Hulton Picture Library, are regularly used to provide illustrative material for film and television programs. For the general use of libraries, however, microphotography has played a much more important role. Many leading newpapers and periodicals have reproduced their entire sets of back issues on roll film, which offers a considerable saving of space and makes it feasible for even a small library to house an entire set. The disadvantage of roll film is that the user must start searching the roll film from the beginning of the reel, no matter where the relevant pages may be on the reel. A considerable advance was achieved by the invention of the transparent Microcard, or microfiche. This is a piece of film cut to a specified size and shape usually approximating a library catalog card but available in more than one size (although the most favoured size is five by three inches). The microfiche offers the advantage of random access; that is, instead of starting at the beginning, the user can bring any section of the microfiche directly into view on the screen. Microfiche also are more convenient to store and handle, and they have become very popular for the production of catalogs and bibliographies as well as for reproduction of texts.

There are various forms of audiovisual media. The most common in libraries is the audio recording on disc or tape, and most libraries, especially public and school libraries, have built up extensive collections of nonbook materials, from the recordings of symphony orchestras on long-playing records to tape-recorded oral history interviews. Cooperation among school and public libraries in this field has made a considerable contribution to local history. The importance attached to these media, particularly in schools, is indicated by the use of the name resource centre for what was formerly called the school library. When teachers are eager to use audiovisual materials, they are often also eager to create materials, and this enthusiasm can enable school librarians to build up a strong and productive relationship between the library and the teaching program.

New articles requiring library storage are the machine-readable magnetic tape and disk. These need such specialized treatment, for the safeguarding of their contents from accidental erasure, that most computer centres employ their own specialist librarian. Like roll film, magnetic tapes and disks do not readily yield information about their contents and therefore require particular care in labeling and indexing.

These are some of the physical materials to be found in library stocks. There are other variations in the form of the contents of records, and these too usually require some form of specialist attention. The literature of science and technology, more than other subjects, contains a multitude of other forms. The printed specifications issued by patent offices and standards institutions form a class with problems peculiar to itself, in that they are written in a highly esoteric language that requires expertise to decipher and index. Patents also illustrate one of the most complex forms of subject classification known, but they are almost always sought by their serial number, and this is how they are usually filed and retrieved.

As in many other areas of information service, the world's major wars have resulted in a spate of documents that

Junior high school student using a video display unit and earphones in his school library, or resource centre, to improve his reading skills.

Ann Hagen Griffiths—Omni-Photo Communications, Inc.

are not published but achieve circulation. They are usually duplicated reports from the laboratories and offices of research organizations and have a more or less restricted readership. Reports on atomic energy research are an outstanding example, the magnitude of the research effort being matched by the magnitude of the document output. The many variations in degree of secrecy have produced a situation of some confusion, and librarians who eagerly sought to attract gifts of these reports in the early days may have lived to regret their enthusiasm. This material, known as grey literature, has established itself and now has a secondary bibliographic literature of its own.

The term "secondary source" applies to documents that contain not original material but catalogs and bibliographies of such. Indexes may apply to the indexes at the ends of books and also may be bibliographies of currently published material, usually of articles in periodicals. The long series, covering many different fields, published by the H.W. Wilson Company of New York City are well known and widely used in other countries, though their coverage is mainly limited to American publications. This has resulted in other national efforts, such as the *Current Technology Index* (British) and the *British Education Index.*

Indexes and abstracts

Abstracts, which are summaries of documents that indicate contents as well as authors and titles, have a respectable history going back at least as far as the *Weekly Memorials for the Ingenious,* published in London in 1682. The *British Librarian* of 1737 published abstracts of well-known and useful books and claimed to cover all the sciences "in a manner never before attempted."

LIBRARY SERVICES

The processes and services found in modern libraries are usually divided into two major categories: technical services, comprising those processes directed at acquiring, arranging, indexing, and storing the stock; and reader services, comprising those processes directed at actively exploiting the stock in satisfying the information needs of the library users.

Technical services. *Acquisition and supply.* The output of published materials, in all forms, is so vast that no single library, not even the largest, can hope to acquire everything; even in relatively specialized fields, some selection has become necessary, and most libraries have an explicit selection policy. The basic principles of selection vary little among different types of libraries, inasmuch as they derive directly from the known interests of the users. Practice is another matter and varies according to the types of user. A national library aims to hold at least one copy of all the publications of its own country and to have a good representation of foreign works, many of which may be obtained through exchange agreements with other national libraries. University, college, and school libraries relate their choice of acquisitions to the programs of teaching and research in their institutions; the academic level of the material naturally varies according to the level of the student population. An elementary school will hold a good selection of books written for children, but a university will tend not to. Many university libraries try to maintain a relatively complete coverage of the reports issued by government and other research establishments. Some universities are designated as repositories for the reports issued by intergovernmental agencies, such as the United Nations, the International Atomic Energy Agency, and the European Economic Community.

Selection of library materials

An important aspect of selection is learning about new publications that would enhance the library. Various surveys have been made of the ways in which specialists gain new information about their fields of work, and the most popular usually turns out to be informal discussions with colleagues. But this is by nature a haphazard process, and most countries now have, or aim to have, a national bibliography based on the acquisitions of the national library. The *British National Bibliography,* begun in 1950 at the British Museum, is a leading example: it is published weekly, with regular cumulations for easy access over long periods. It is a tool for subject inquiry searches as well as for current selection.

The International Federation of Library Associations has established a program to increase the range and number of such bibliographic tools. The program, called Universal Bibliographic Control (UBC), aims to encourage national libraries, or groups of libraries, to institute methods of recording their national publications in a standard format and, wherever possible, of entering them into computer files.

The UBC program is accompanied by a second effort, the Universal Availability of Publications (UAP) program, which aims to provide the necessary follow-up service of document delivery. Both programs have been supported by the British Library Reference and Lending Divisions.

Other aids to the selection of material for acquisition are legion. Many libraries join professional societies and institutions to obtain their publications, which usually contain lists and reviews of new work relevant to their subjects. Leading journals such as *The Times Literary Supplement, The New York Review of Books, Nature,* and *Science* contain reviews by experts, advertisements for new and forthcoming publications, and review articles covering important new books in special fields.

Similarly, the development of electronic means of document delivery is unlikely to supplant the more traditional sources of supply, the publishing and bookselling trades. Some companies combine the two functions. Purchases by libraries have traditionally generated much of the revenue of local bookshops, but firms operating as specialist library suppliers are able to offer many auxiliary services, such as attaching plastic covers and inserting labels of ownership, because they deal in large-scale bulk supply and can afford to maintain machines for such processes. Also, some schools and colleges have set up bookshops of their own, to supply their libraries and also their students with textbooks.

Such is the situation in countries with long-established traditions of reading, research, libraries, and book trade. Far greater difficulties confront the library services in the developing countries, particularly in Africa and Asia. Even in India and China, with their long history of using books, a steady and satisfactory progress is hindered by shortages of finance, materials, and trained staff. Some universities in these nations have large libraries and receive grants that enable them to acquire foreign as well as national publications, but they often meet with delays caused by administrative procedures, shortage of foreign currencies, and problems of language in the postal services. In most African countries, growth of a national literature is hampered by the cost of importing even basic materials such as paper.

Many countries in eastern Europe as well as in the Third World look to exchanges as a means of obtaining materials. Some governments allow libraries to exchange duplicate copies of national publications, as a recognized method of compensation without payment in foreign currency. The practice does present certain administrative problems, but it is a useful means of encouraging the international flow of publications as well as of giving practical help in collection building to libraries in countries with limited resources.

Cataloging. However careful and scholarly the methods used in building a collection, without expert guidance to its access and use the collection remains difficult to approach. Cataloging and classification, well-tried disciplines often combined under the general heading of "indexing," provide the needed guidance. Both techniques have been in use as long as libraries have existed, and their potential value in the "information age" has been enhanced by the sensible use of computers.

Function of the catalog

The function of the catalog is to identify all the items in a collection. All of the great libraries of the ancient world seem to have had lists and inventories, whether kept on clay, stone, papyrus, parchment, palm leaves, or bamboo strips. Examples may be found in museums throughout the world. For many centuries the feature that gave a work its unique identity was the name of the writer, and users of the library were expected to know the names of the authors whose works they wished to consult. Some libraries, scholarly libraries in particular, continue to offer only author catalogs as guides to their collections.

Many factors have contributed to the rise in importance of the subject approach to information. From the earliest times, librarians recognized that readers would be greatly helped if the catalog entries were arranged in groups of related subjects. General classifications of knowledge such as those of Aristotle and Porphyry, scholarly curricula such as the trivium and quadrivium (expounded by Julius Caesar's librarian Marcus Terentius Varro and persisting as a major influence on education), and practical considerations such as the governmental needs of emperors and priests all have formed the basis for the arrangement of subject catalogs. Many Chinese works published under a single title were actually hundreds of volumes, libraries in themselves, with important bibliographical sections cataloging other books now lost. Early in the 7th century the scholar Wei Cheng wrote the bibliographical section of the official *Sui Dynasty History,* dividing the books into four categories: Confucian Classics, historical records, philosophical writings, and miscellaneous works.

Since the late 19th century far more attention has been paid to cataloging the subject contents of books as well as the names of their authors. Most of the impetus for this change came from science and technology, where the practice of working in teams in research institutions largely superseded the practice of single individuals, such as Charles Darwin, working for years to complete their research and then publishing the results as a book. The proliferation of specialist journals that publish short papers charting the progress of teamwork has meant that the names of single authors have become somewhat less important as tools for identifying works in libraries. Catalogs that list the subjects of research are more useful to specialists in related fields around the world, who may not know researchers by name but wish to have access to their work.

Major catalog systems

Despite a steady, if slow, trend toward standardization, various forms of catalog continue to exist. Sets of entries generally are arranged in one of three catalog systems. The first is the dictionary catalog, in which author, title, subject, and any other entries are filed in a single alphabetical sequence. This form is popular in the United States and in public libraries generally and probably presents the least amount of difficulty for the general or casual reader. The second is the divided catalog, still in alphabetical sequence but with subject entries in a separate file. This form has increased in popularity, and many libraries have divided their former dictionary catalogs, recognizing the growing value of the subject approach. The third is the classed, or classified, catalog, which is more popular in Britain and continental Europe and in some developing countries whose librarians trained there. In the classed catalog, as its name suggests, all the entries are filed in the sequence of a classification scheme, that is, in a systematic order of subjects, but separate alphabetical files link names of authors and of subjects to the notation symbols of the classification scheme used in the main file. The chief advantage of a classed catalog is that the entries are related subjects grouped together in the file; thus, a subject search can be made much more simply than in a catalog based on the alphabet. In addition, when different languages are used, the sequence of entries in a classed catalog does not alter, as is the case with the dictionary.

The types of catalog differ on the basis of the information provided in the entries, but the actual physical form may also vary. Originally, catalogs took the same form as the books they listed; being made of the same material, the catalog was an extra item of the collection itself. The earliest catalogs of the great national and scholarly libraries were in book form, with handwritten entries and spaces for new additions. Such manuscript books still exist, for example in older British university libraries such as the Bodleian at Oxford. In the 19th century many libraries began to print a book catalog, of which several copies could be made available. Librarians found the main problem of the book-type catalog to be that of inserting entries for new acquisitions. Most plans, like that of the Bibliothèque Nationale in Paris, made no attempt to add to the printed file and instead placed vain hopes in the prospect of new editions. The British Museum in addition to its published general catalog maintains several copies of a guard book catalog, in many hundreds of volumes, in which entries for new acquisitions are typed on slips and pasted in at the appropriate part of the alphabet. Spaces are left for such inserts, and as these are used up, new pages are added, and the entries are redistributed to make more empty space.

The solution to the problem of adding to a catalog in book form was the catalog on cards, each entry having its own card and each card containing only one entry. In principle, such catalogs can grow in size indefinitely; any new entry can be filed between any two existing entries. Thus the catalog offers the opportunity to have a completely up-to-date file: an entry can be made in the catalog immediately after a book has been purchased.

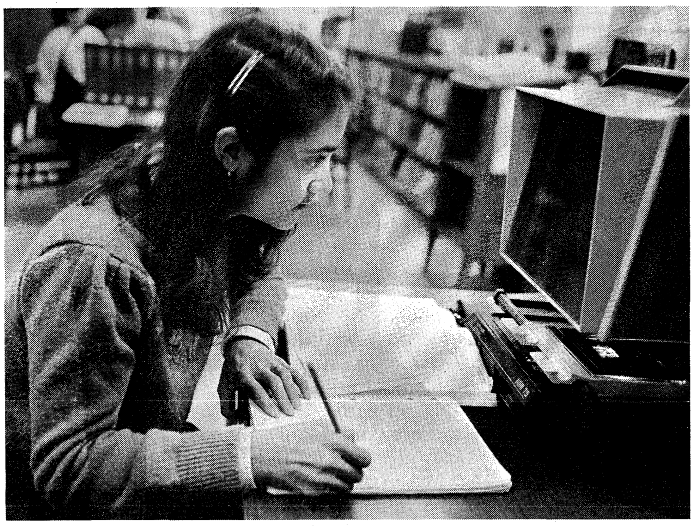
An outstanding innovation was introduced in 1901, when the Library of Congress began offering copies of its own catalog cards for sale to other libraries. For many years this proved of inestimable value, particularly to small libraries unable to afford skilled catalogers. The service was also intended to serve as a central cataloging agency. Many eminent librarians, in conferences and in published papers, had lamented what they argued was wasteful duplication of effort involved in the separate cataloging of the same books in many different libraries. They proposed that a central agency undertake the task and make the results generally available, so that any library could use the central catalog thus produced to complete its own highly professional catalog. In the 1950s the British National Bibliography also began to produce cards from the entries in its weekly lists.

These and similar schemes in other countries in Europe achieved a certain success but for various reasons could not be said to have provided the ultimate solution. The advent of the computerized catalog, however, offers a more practicable approach because the storage capacity and the operating speed of even small machines overcome the main drawbacks to card services: delays in production and the labour of filing the cards when they arrive. The computer makes it possible for the details of any document to be entered into a file at any point and then to be transmitted to a central data file from which other libraries can obtain details on a video display unit. These details can then be transferred to their own files, to be consulted either in printout form or directly on the video display unit if the library does not wish to maintain a printed form of catalog. The process is demonstrated by the revised Machine-Readable Cataloging Project, known since its revision in 1968 as MARC II. Library users find no difficulty in consulting such on-line catalogs and may well prefer them to the more cumbersome, if more familiar, form of cards in drawers.

Use of microform for catalogs

The ever-growing demand for more floor space has led to a widespread use of microform, either in roll film or, more satisfactorily, in the form of the transparent cards known as microfiche. Film and microfiche readers have become common features in libraries of all types, but they have met more resistance from readers than has the video display unit. The advantages they offer, in saving of space and in availability from a central source, have persuaded many librarians to close their card files completely and begin again. Improvements in photo technology make it possible to store hundreds, even thousands, of pages on one microfiche and to provide a perfectly legible enlargement on a reader machine. Because of the amount of time required to produce the microfiche and the mounting costs of cumulations, however, the microfiche is seen by some as only temporarily useful, a step in the evolution from the card file to the video display unit.

Patsy Davidson—The Image Works



High school student taking notes from source material displayed on a microfilm reader in her school library.

The ideal of centralized cataloging led to increased interest in standardized forms of entry. As libraries grew larger after the Renaissance and the invention of printing produced more authors, it became necessary to devise some form of standard to ensure consistency among several catalogs. Perhaps the most famous, the British Museum Rules, was inspired by Sir Anthony Panizzi and has influenced all succeeding codes, though most of them have departed considerably from the original. The Vatican Rules and the Prussian Instructions have both been subject to commissions for revision, but certainly the most influential code is the Anglo-American *Catalog Rules; Author and Title Entries,* first published in 1908 and revised in 1967. A further revision was published in 1978 as *Anglo-American Cataloguing Rules,* second edition; it is commonly referred to as *AACR2.*

Many other discussions, revisions, and simplifications took place after World War II. Short versions of the major codes were published for small libraries, in Czechoslovakia for example; the public instruction ministry in Italy issued new rules; a French commission on cataloging issued standards for anonymous works; and in the Soviet Union proposals were published for standardizing the transcription of Chinese names into Cyrillic script.

All of these codes dealt with the entering of names of authors, including Anonymous in the case of anonymous works, and sometimes, as in *AACR2,* with titles. A separate set of codes for subject cataloging emerged mainly in the United States, where they took the form of lists of subject headings, the three best known being the list compiled by Minnie E. Sears, the *Library of Congress Subject Headings,* and the Medical Subject Headings (MeSH) of the U.S. National Library of Medicine.

Thesaurus. A new use of the term thesaurus, now widespread, dates from the early 1950s in the work of H.P. Luhn, at International Business Machines Corporation, who was searching for a computer process that could create a list of authorized terms for the indexing of scientific literature. The list was to include a structure of cross-references between families of notions, in the manner of P.M. Roget's *Thesaurus of English Words and Phrases* (1852) and similar to the structure of faceted classification schemes. A major thesaurus, and one of the earliest, is the *Thesaurofacet,* a list of engineering terms in great detail designed by Jean Aitchison for the English Electric Company. The thesaurus has proved very useful both for indexing and for searching in machine systems.

Classification. While catalogs aim to identify and list items in a collection, schemes of classification have a more general application in arranging documents in a sequence that will make sense and be helpful to the user. Because they display subjects, and not documents, they can be used in several libraries, and some indeed have found applications in many different countries. Like schemes for grouping entries in catalogs, classifications—whether of knowledge based on philosophical principles, of the subject faculties of universities, or of the pragmatic grouping of books on shelves—have formed the basis of many individual systems.

Dewey Decimal Classification

The best known of all schemes for the classification of documents in libraries is the Dewey Decimal System of Classification, devised by Melvil Dewey in 1873 and published in 1876. Apart from being the first modern classification scheme for libraries, the Dewey system embodies two of Dewey's many contributions to the theory and practice of librarianship. First, he recognized that a systematic arrangement of books on shelves should make sense to the users; his scheme therefore reflected the dominant pattern of current thinking, exemplified by the "classificatory sciences." And second, he used decimals as notation symbols, which illustrated the way in which subjects were divided hierarchically, from main classes to specific topics. An example from the schedule for chemistry shows how numbers are subdivided:

540	chemistry and allied sciences
541	physical and theoretical chemistry
541.2	theoretical chemistry
541.3	physical chemistry
541.34	solutions
541.35	photochemistry
542	laboratories, apparatus, equipment.

Another feature of the Dewey system is the mnemonics used for certain types of subdivisions. Thus, many subjects can be subdivided geographically by the use of the historical-geographical number as decimals:

900	general geography and history
970	history of North America
973	history of the United States.

Combining with the art schedule, the number for history of art in the United States is obtained:

700	the arts
709	history of art
709.73	history of art in the United States.

Universal Decimal Classification

The Universal Decimal Classification, published in 1905, was an immediate offspring of the Dewey system. Paul Otlet and Henri-Marie Lafontaine adapted the Dewey system as the basis for a much more detailed scheme suitable for use in a vast card index of books and periodical articles in classified order—a universal bibliography of recorded knowledge. While retaining the basic generic hierarchies, the Universal Decimal Classification makes far greater use of the technique of synthesis, by providing a series of auxiliary tables for aspects of subjects likely to appear in several parts of the main schedules. These tables are indicated by the use of symbols such as punctuation marks. The colon sign (:) indicates a relationship between any parts and is the most commonly used sign. The numeral 669.1 being the notation for iron and steel and 546.22 for sulfur, the compound subject can be indicated by the notation 669.1:546.22, sulfur in iron and steel.

Like the Dewey Decimal Classification, the Universal Decimal Classification has been translated into many languages, and the International Federation for Documentation (Fédération Internationale de Documentation; FID) has undertaken responsibility for its continual revision. Scientific and technical libraries use the Universal system in preference to the Dewey system.

Library of Congress Classification

At the turn of the 20th century Herbert Putnam, the Librarian of Congress, decided to reclassify the library but rejected the Dewey system. His staff adopted a more pragmatic approach, based entirely on the way in which the books were arranged in their subjects on the shelves. They also rejected the decimal notation, preferring a purely ordinal system combining letters and numbers, leaving blank spaces where they expected new subjects to develop. (Not all of their expectations have proved correct.) American libraries and some scholarly libraries elsewhere have found the scheme attractive for its depth of detail, inasmuch as it is based on a very large library. An additional advantage is that Library of Congress notations appear on the library's catalog cards and on computer tapes produced by the MARC Project. Several American university libraries have undertaken the daunting task of reclassifying their stock from the Dewey system to the Library of Congress system in an effort to reap the maximum advantage of these features.

Other systems

Although not widely used, the bibliographic classification system invented by Henry E. Bliss of the College of the City of New York (published in 1935 as *A System of Bibliographic Classification*) has made important contributions to the theory of classification, particularly in Bliss's acute perception of the role of synthesis and his insistence that a library scheme should reflect the organization of knowledge and the system of the sciences. His systematic auxiliary schedules, designed to achieve what he called composite specification, carry the synthetic principle into every subject area and give a far higher degree of flexibility than does a purely enumerative scheme like the Library of Congress system. The Bliss Classification Association, founded in the United Kingdom in 1967, undertook the production of a new complete edition.

Perhaps the most important advance in classification theory has been made by the Indian librarian S.R. Ranganathan, whose extraordinary output of books and articles has left its mark on the entire range of studies from archival science to information science. He introduced the term facet analysis to denote the technique of dividing a complex subject into its several parts by relating them to a set of five fundamental categories of abstract notions, which he called personality, energy, matter, space, and time. He employed these in his Colon Classification system (1933), which is used in some Indian libraries and has found few followers elsewhere. Nevertheless, the ideas in the scheme, expounded in his *Colon Classification* (1933) and *Prolegomena to Library Classification* (1937), have influenced all later work in classification theory and practice, including subsequent editions of the Dewey, Universal, and Bliss systems.

In both the Soviet Union and the People's Republic of China schemes have been published that depart somewhat from the Anglo-American traditions and claim to reflect the structure of knowledge according to the principles of Marxist philosophy. Both have enumerative structures and may be distinguished by their detail of analysis of, and dependence on, the corpus of Marxist literature—a literature that, in Anglo-American schemes, usually occupies a relatively minor place.

Services to users. *Circulation.* Although many of the libraries in antiquity were open to the literate public, this was almost certainly for reference only. Some monastic libraries, however, are known to have allowed the monks to borrow books for study in their cells; the Rule of St. Benedict explicitly permitted this, and the librarian exacted penance from any monk unable to confirm that he had actually read his book. Some university libraries may have lent books to members of their faculties, but the notion of lending, or circulating, libraries did not become popular until the 18th century.

The rapid development of public libraries in the 19th century led to the extension of the practice and to the introduction of various systems for the recording of loans.

Introduction of circulation systems

All of the early systems depended on the use of one or more cards on which were recorded the name of the book, the name of the borrower, and the date on which the book should be returned. One system that remains popular in small libraries makes use of an author/title card, which is inserted in each book, and a pocketlike ticket, which is issued to the reader. When the reader borrows the book the book card is filed in the reader's ticket, and this "charge" is filed in the order of the return date. This chronologically organized system keeps a check on books that become overdue for return and enables librarians to limit the number of tickets issued to each reader, but the system cannot inform a reader of which books he has already on loan. Where such information may be useful, as in university libraries, a common practice is to require the reader to fill out a form, in duplicate or even triplicate, with the details of the book and the borrower's own name and address. Such a system gives more information but requires an inordinate amount of time and labour for maintenance of files. Variations of these two systems can be found throughout the world.

A computerized circulation system offers the opportunity of access to all of the information needed about loans and borrowers, and even small libraries can find circulation programs using relatively inexpensive hardware, as offered by several manufacturers. Information on each book is recorded on a bar-coded slip inserted into the book, usually adjacent to the date label on which the return date is stamped. Information about the borrower is likewise recorded on a bar code attached to a plastic identity card. Activating the two bar codes together, by a light pen, for example, enters all of the details of a loan into the library's computerized file; activating either bar code alone will then show on the video display unit the details of transactions involving the book or borrower. Computerized circulation methods make borrowing more convenient for the user and reduce staff time and labour required for the task.

Reference and retrieval services. Open access to the shelves and the facility to borrow books mean that much of the use of a modern library is at the free choice of the reader; scholars and scientists continue to emphasize the value of browsing among the shelves of a well-arranged library. "Chance favours only the prepared mind," said Louis Pasteur, and serendipitous discoveries of useful information when searching for some other subject have become a familiar and welcome aspect of using a library or other information service.

In reference service, librarians have traditionally given personal help to readers in making the best use of collections to satisfy their information needs. The publication of printed catalogs and bibliographies and the organizing of interlibrary cooperation have widened the range of the resources available to the individual reader. In scholarly libraries assistance to readers in the past may often have been limited to explaining the layout of the library and the use of the catalog; in universities, members of the faculty would have been expected to know the literature of their subject better than any librarian. But in public libraries, and still more so in special libraries in the fields of science and technology, readers have long been accustomed to seek guidance about information on their subject as well as about the library. This process has been greatly extended by the enormous increases in research worldwide and in the quantity of information and publications available in many languages and by the excellence of the indexes, abstracts, bibliographies, and data bases that help to control the documentation of this massive output.

Retrospective searching and current-awareness service

Reference services can be broadly divided into two main aspects, usually known as retrospective searching (or information retrieval) and current-awareness service (or selective dissemination of information). These terms indicate a specialization that has occurred in this core activity of libraries and that grew mainly out of the expansion of scientific and industrial research during and after World War I. Three factors strongly influenced this process. First, the increase in research and publication affected all types of libraries and brought with it a similar increase in subject specialization. Second, working scientists, accustomed to referring to reports in published papers, were content to leave the organizing of information searches to a colleague who knew and understood their work. And third, the widespread application of scientific research in industry provided an extra stimulus to the division of labour because of the necessity for speedy application of results to gain commercial success in production.

Some information specialists have tried to draw a distinction between their reference work and the more general reference services of librarians, but in most countries there is close cooperation among all engaged in these professional activities. Most acknowledge a mutual interest and influence, while the range of duties allows, indeed requires, different emphases in different institutions.

All agree, however, in acknowledging the duty to assist users to find answers to inquiries and to carry out retrospective searches in existing literature. Such a service requires many qualities, personal and professional: a detailed knowledge of books, periodicals, and all other forms of record; an ability to search efficiently in catalogs, indexes, abstracts, and data bases; and, above all, a sensitive understanding of each user's needs. Matching the terms used by a reader in posing a question to the terms used by authors, indexers, and catalogers may well constitute one of the subtlest of professional skills.

Computer search

The outcome of a search can take many forms, from a short factual statement that gives the needed information to a short list of relevant references or a full-scale bibliography. In a computer search the first request often reveals that the data base contains hundreds or even thousands of "hits," the references relating to the topic requested. The number can be reduced by narrowing the subject, adding more specific details, or persuading the reader to be more precise in defining the information needed. When a reasonable number of hits has been reached the computer can be instructed to display the details of a few references, to show the reader whether or not the search has covered the right subject area. If it has, the set of references or abstracts may then be obtained as printout; if it has not, the search begins again using new terms for the request.

In the specialized information centre a professional researcher can conduct the search and provide a state-of-the-art review of the literature in narrative form instead of as a collection of references. The service represents a peak of efficiency on behalf of the client, who has neither the time nor the resources to make the same review. The value to industry and commerce has encouraged private individuals to set up as information brokers to provide these services as a commercial enterprise.

Current-awareness services. The purpose of a current-awareness service is to inform the users about new acquisitions in their libraries. Public libraries in particular have used display boards and shelves to draw attention to recent additions, and many libraries produce complete or selective lists for circulation to patrons. Lists like these are a general response to the wishes of the users to know about new books and documents that may be of interest to them. Some libraries have adopted a practice of selective dissemination of information, whereby the notices go to specific readers or groups of readers, and their contents are selected with the special interests of those readers in mind. The items listed include books, periodical articles, patent and other specifications, the unpublished but circulated material known as grey literature, and any comment that the librarian or information specialist may identify as being relevant to particular users.

Each reader or group registers a profile of interests with the library staff, who then scan incoming items with the profiles in mind and select items of interest. Where computer equipment is available, profiles of the users can be run periodically against new additions to the public data bases, and the staff can prepare a customized printout showing new additions of interest to the user concerned.

Thus, while retrospective searches are a response to an expressed need on occasion, current-awareness service anticipates needs before they arise and enables the users of the service to rely on the information staff to keep them up-to-date with new work in their subjects. Both these services can be developed to a high degree of effectiveness by the use of computers: the former because of the immense range of data bases available for searching, the latter by means of electronic message transmission on local area networks.

Community information services. The growth of information services in special libraries, followed by college and university libraries, also has influenced public library practice in library extension programs and community information services. Extension programs are usually arranged in cooperation with local educational organizations, university extramural courses, parent–teacher associations, and so on. In developing countries with an infant publishing and book trade, public libraries can offer valuable assistance to local authors, particularly those writing in indigenous languages, by providing facilities for authors to give lectures, hold seminars, and develop their own skills in direct relation with their potential readers. In European, African, and American libraries, poets or writers in residence have appeared as a part of similar action to bring authors and readers together.

Community information service

The community information, or outreach, program has become a recognized feature of a public library service, in rural as well as urban or deprived areas. The Jamaica Library Service, for example, has long made a practice of setting up a stall at farmers' markets to supply up-to-date books and pamphlets on agriculture. Public libraries in China regularly set up special links with local factories for the supply of technical literature and specialist advisory staff.

INTERLIBRARY RELATIONS

Library cooperation. The publication of bibliographies and of library catalogs emphasized that no library could afford to be self-sufficient and stimulated interest in various forms of interlibrary cooperation. This probably originated informally, with readers referring to union catalogs to locate libraries that contained the books they wanted. One of the earliest formal organizations began with the Central Library for Students, founded in London by Albert Mansbridge in 1916. This was transformed in 1930 into the National Central Library, which continued to act as a lending library but also formed the centre of a network of regional library bureaus. The bureaus were located in a major regional library, and, with one exception, built up union catalogs of holdings in the local public libraries to facilitate interlibrary lending. The National Central Library encouraged other university and special libraries to participate. The National Central Library has become part of the British Library Lending Division, which undertakes a major part of interlibrary lending both in the United Kingdom and internationally.

Most countries have some form of organization for interlibrary cooperation. In Europe this has tended to be based on union catalogs compiled in a central national or university library, such as the National Széchényi Library in Hungary; in the Soviet Union the Lenin Library has played a prominent role, and, as in China, regional libraries cooperate extensively with local schools and industries in the provision of books, reading lists, and user education.

British National Book Centre

The progress of interlibrary lending, coupled with the great losses suffered by libraries in Europe and Asia during World War II, led to an interest in cooperative acquisition of new materials. In 1948 the British National Book Centre was set up at the National Central Library in London to gather unwanted duplicates and to distribute them to the libraries that had suffered losses. It proved to be of incalculable value and was soon followed by the United States Book Exchange; both distributed lists of wants and offers to their member libraries.

An ambitious program for cooperative acquisition of foreign materials by American libraries was conceived in the Library of Congress in 1942. This was the Farmington Plan: it involved the recruitment of purchasing agents in many countries, whose task was to buy their countries' current publications and distribute them to American libraries according to a scheme of subject specialization. Many criticisms were leveled at the scheme, and as a blanket operation it inevitably acquired a certain amount of trivia; but many research libraries have benefited by the acquisition of materials that otherwise would have been difficult to obtain.

Before the building boom of the 1960s, pressure on library space spurred librarians to discuss means of cooperative storage. Perhaps the foremost example is the Center for Research Libraries (formerly the Midwest Interlibrary Center) in Chicago, which began in 1952 as a centre for deposit of duplicate and little-used materials from research libraries. With the aid of a special grant, the University of London established a depository Library, at Royal Holloway College away from the centre of London, to which the colleges of the university can send materials for either cooperative or private storage. The British Library Lending Division also acts as a cooperative store; it receives unwanted items from any library and makes them generally available. Both libraries reserve the right to refuse items that they already have in cooperative storage.

International Federation for Documentation

International organizations. The oldest organization in the library and information field is the International Federation for Documentation. It was founded in 1895 in Brussels as the Institut International de Bibliographie by Paul Otlet and Henri-Marie Lafontaine, as part of their plan to create an index of world literature on cards. The institute has many international committees, and some, especially those concerned with classification research and the constant revision of the Universal Decimal Classification, are very active. The Soviet Union's All-Union Institute of Scientific and Technical Information (VINITI) issues *IFID,* the journal of the International Forum on Information and Documentation (IFID). It hosts the secretariat of an IFID committee charged to research the theoretical basis of information. The International Federation of Library Associations and Institutions, IFLA (Fédération Internationale des Associations de Bibliothécaires et des Bibliothèques, FIAB), was founded in 1927 and first met formally in Rome in 1928. The journal *IFLA Communications FIAB,* originally published as part of the international journal *Libri,* is now constituted as a separate *IFLA Journal,* and the organization's annual conference proceedings are published as a separate volume.

The International Council on Archives (ICA) was set up with the help of Unesco in 1948, and the first International Congress of Archivists was held in Paris in 1950. Early and continuing interest has centred on the microfilming, conservation, and preservation of historical records, particularly those of central and local governments. There has also been a considerable increase in interest in the records of business management, including those of what has come to be known as "industrial archaeology." The founding members were mostly from national archives, but universities in some countries have also assumed some responsibility for archive collection and preservation.

All of these associations have received considerable moral and financial support from Unesco, the first General Conference of which took place in 1947. From its inception Unesco has placed great importance on the encouragement of bibliography and libraries, and of public libraries in particular. (Part of its program was inherited from a League of Nations organization called the International Institute of Intellectual Cooperation, a main concern of which was libraries.) Unesco's support has led to seminars on public library development and to pilot public library projects in many nations. Not all have flourished, however, partly because of lack of local support. Other projects have covered education and provision of scholarships, library buildings, microfilming of archives and records, and establishment of scientific and technical documentation centres.

The technical committee of the International Organization for Standardization, another United Nations body, has helped to formulate and promulgate a number of standards on bibliographical formats, particularly those related to computer processing. Some of these have been modeled on national standards, such as those published by the British Standards Institution.

THE INFORMATION PROFESSIONS AND PROFESSIONAL EDUCATION

Throughout the centuries librarians have preserved books and records from the hazards of war, fire, and flood, and it is no idle boast to say that they have played a large part in maintaining the cultural heritage of their countries. Francis Bacon, in sending a copy of his book *The Advancement of Learning* to Sir Thomas Bodley, wrote, "you have built an Ark, to save Learning from Deluge." Although the traditional librarian acted primarily as a keeper of records, the concept of an active service of advice and information eventually appeared as a legitimate extension of the role of custodian.

The rise of scientific and industrial research and the establishment of public libraries in the 19th century led to the greatly increased emphasis on the subject approach and the role of systematic cataloging and classification in addition to the accepted function of building the collection and the consequent need for expert knowledge of bibliography, both systematic and analytical. In the industrial library in particular, the information officer was almost entirely concerned with the contents of documents and was indifferent to their form, so that a scrap of paper recording an important telephone call would have more significance than an incunabulum. The proliferation of different forms of record led to a much wider view of information storage and retrieval methods, used by subject specialists who knew and understood the work of their specialist colleagues.

Professional skills range from those of the archivist, concerned with provenance, preservation, and interpretation of administrative records, to those of the information scientist, concerned with research on the nature of information itself and the process of information flow and transfer between individuals and communities. These branches of the information profession share many objectives, practices, and skills. Each branch works to make the records of human progress readily available, and the contribution of each to society can only suffer if it is not allowed to become integrated with the others.

The personnel requirements of the profession include several categories, based on several kinds of specialist knowledge and skills: a knowledge of the nature of documents and their role in collection building; skills in the organization of knowledge through cataloging and classification; an ability to analyze and survey needs and to disseminate information in response to and in advance of inquiries. Efficient support staff is needed to maintain the machines, hardware and software; and clerical workers, technicians, and stewards are needed to ensure that the foundations on which the professions rely and the buildings in which they work operate in good order.

Programs for professional education have been the subject of debate throughout the world, and representatives of national associations, of Unesco, and of the international professional bodies have attempted to reach a consensus. For the higher levels of professional work, a degree from a university testifies to achievement in a subject field at the level likely to be found among the more demanding of the users of any library.

Library schools

Most of the initiatives for the setting up of courses for the education of professionals have come from librarians or their professional associations. In the United States the first university school for librarians was established in 1887 by Melvil Dewey at Columbia University. The American Library Association pursued a policy of "accreditation" to ensure that library schools offering a professional qualification should all reach a standard laid down by the profession itself. The first British library school was established in University College London in 1919, and until 1946 all other qualifications were gained through public examinations conducted by the Library Association. There are now many other schools, some in universities but most in polytechnic institutes, where the Library Association's own standards still influence the curriculum. The Library Association's successive syllabi have had considerable importance for some Commonwealth countries, such as Ghana, Nigeria, and the Caribbean countries.

In continental Europe most professional education takes place in universities and similar institutions of higher learning. In Hungary the University of Budapest began courses in the Faculty of Philosophy in 1949, and a senior-level course in documentation was later organized jointly by the Chair of Library Science in the university (now the Eötvös Loránd University) and the National Technical Library and Documentation Centre, in 1964. In Czechoslovakia, library and information science courses are given at the Chair of Library Science and Scientific Information, also in the Faculty of Philosophy and Letters of Charles University in Prague and Comenius University in Bratislava. In France the long-established École Nationale des Chartes, which mainly trains archivists, also prepares students for the public, national, and university libraries. The École Nationale Supérieure des Bibliothèques belongs to the Direction des Bibliothèques, and the École de Bibliothécaires-Documentalistes is a private institution of the Institut Catholique de Paris.

The universities of Peking and Wu-han in China have advanced courses and research programs in librarianship, and professional qualifications may also be gained by correspondence. In 1985, with the help of Unesco and the British Council, a master's degree course in information studies was begun at the Institute for Scientific and Technical Information in China.

The curriculum. In the older schools, and in those closely connected with national and scholarly libraries, the emphasis was on historical and bibliographical aspects. The history of libraries, the art of the book, classification and cataloging, literary studies, and the library in society occupied most of the teaching and examining. Subsequent generations of professionals have regarded such a syllabus as too biased toward the humanities to be appropriate for a scientific and technical organization's requirements and have devoted more time to the scientific literature, to indexing and abstracting techniques, and to information technology. Much more research effort is now directed also to the theory of information transfer and the development of mathematical models for this and to other aspects of management in library and information services.

Continuing education. Qualified library and information specialists have many opportunities to enroll in short or part-time refresher or reorientation courses to bring themselves up to date with changing methods or with developments in techniques and attitudes. Opportunities for systematic research, which have been available in the United States since the pioneering days of Louis R. Wilson at the University of Chicago in the 1930s, have now become almost universal. In the United Kingdom the University of Sheffield, the City University in London, and University College, London, have played an important role in offering such opportunities to teachers from the polytechnic library schools. Postgraduate scholarships and grants from Unesco, the Council on Library Resources, the British Council, universities, and academies of science have enabled students from Third World countries to advance their education in the United States, the United Kingdom, the Soviet Union, France, and Australia.

BIBLIOGRAPHY

Encyclopaedias: There are many encyclopaedias on the subject of libraries, especially in the German language. In English, the three most important are THOMAS LANDAU (ed.), *Encyclopaedia of Librarianship,* 3rd rev. ed. (1966); ROBERT WEDGEWORTH (ed.), *ALA World Encyclopedia of Library and Information Services* (1980); and ALLEN KENT *et al., Encyclopedia of Library and Information Science,* 35 vol. (1968–83), continued with supplemental volumes. For terminology see LEONARD MONTAGUE HARROD, *Harrod's Librarians' Glossary of Terms Used in Librarianship, Documentation and the Book Crafts, and Reference Book,* 5th ed., rev. and updated by RAY PRYTHERCH (1984). *The ALA Yearbook of Library and Information Services: A Review of Library Events;* and *The ALA Glossary of Library and Information Science,* ed. by HEARTSILL YOUNG (1983), are useful.

Origins and early history of libraries: A classic, well-written and well-illustrated account of the beginnings of printing and bookmaking is given in DOUGLAS C. MCMURTRIE, *The Book,* 3rd ed. (1943, reprinted 1972). The English library is dealt with in RAYMOND IRWIN, *The Origins of the English Library* (1958, reprinted 1981), and *The Heritage of the English Library* (1964). KARL CHRIST, *The Handbook of Medieval Library History,* rev. by ANTON KERN, trans. from the German and ed. by THEOPHIL M. OTTO (1984), is a part of the comprehensive work *Handbuch der Bibliothekswissenschaft,* 2nd rev. ed., ed. by FRITZ MILKAU and GEORG LEYH, 3 vol. in 4 (1952–61). A shorter survey, also international in scope, is D.N. MARSHALL, *History of Libraries, Ancient and Medieval* (1983).

Kinds of libraries: As a general guide to the field, see *The Bowker Annual of Library and Book Trade Information.* Several monographs aim to relate the library to society: a sociological approach to different types of libraries in different countries can be found in A. ROBERT ROGERS and KATHRYN MCCHESNEY, *The Library in Society* (1984). RONALD CHARLES BENGE, *Cultural Crisis and Libraries in the Third World* (1979), has a wide scope; and D.J. FOSKETT, *Pathways for Communication* (1984), examines the role of books and libraries in the information age.

National libraries are dealt with in MAURICE B. LINE and JOYCE LINE (eds.), *National Libraries* (1979); and in the beautifully illustrated ANTHONY HOBSON, *Great Libraries* (1970). JOHN Y. COLE, *For Congress and the Nation: A Chronological History of the Library of Congress Through 1975* (1979), is an informative survey. All aspects of university libraries are covered in RUTHERFORD D. ROGERS and DAVID C. WEBER, *University Library Administration* (1971); JAMES THOMPSON, *An Introduction to University Library Administration,* 3rd ed. (1979); and KEYES D. METCALF, *Planning Academic and Research Library Buildings* (1965). The general organization of public libraries has been the subject of many treatises: JOSEPH WHEELER and HERBERT GOLDHOR, *Wheeler and Goldhor's Practical Administration of Public Libraries,* rev. ed., ed. by CARLTON ROCHELL (1981); VERNON E. PALMOUR, MARCIA C. BELLASSAI, and NANCY V. DEWATH, *A Planning Process for Public Libraries* (1980); THOMAS KELLY, *A History of Public Libraries in Great Britain, 1845–1975,* 2nd rev. ed. (1977); MIRIAM BRAVERMAN, *Youth, Society, and the Public Library* (1979); DOROTHY M. BRODERICK, *Library Work with Children* (1977); and GENEVIÈVE PATTE and SIGRÚN KLARA HANNESDÓTTIR (eds.), *Library Work for Children and Young Adults in the Developing Countries* (1984). L.J. ANTHONY (ed.), *Handbook of Special Librarianship and Information Work,* 5th ed. (1982), has value for all types of libraries. Close relationship between modern educational methods and libraries is well explored in NORMAN BESWICK, *Resource-Based Learning* (1977); and S.R. RANGANATHAN, *New Education and School Library* (1973). FRANCES LAVERNE CARROLL (ed.), *Recent Advances in School Librarianship* (1981), has an international scope and authorship. Special topics are covered in S. JOHN TEAGUE, *Microform, Video and Electronic Media Librarianship* (1985); and GEORGE MARTIN CUNHA and DOROTHY GRANT CUNHA, *Library and Archives Conservation: 1980s and Beyond,* 2 vol. (1983), the second volume of which contains an extensive bibliography.

Automation in libraries and information services is introduced in C.J. VAN RIJSBERGEN, *Information Retrieval,* 2nd ed. (1979); and a more advanced account is found in F. WILFRID LANCASTER, *Information Retrieval Systems: Characteristics, Testing, and Evaluation,* 2nd ed. (1979). Two simple practical books are IAN LOVECY, *Automating Library Procedures* (1984); and J.E. ROWLEY, *Computers for Libraries,* 2nd ed. (1985).

Technical services: A detailed analysis of all aspects of technical services is offered in B.C. VICKERY, *Information Systems* (1973). ERIK J. HUNTER and K.G.B. BAKEWELL, *Cataloguing,* 2nd rev. ed. (1983), emphasizes automation and the role of networks; some perspectives are outlined in PETER GELLATLY (ed.), *Beyond "1984": The Future of Library Technical Services* (1983). A.C. FOSKETT, *The Subject Approach to Information,* 4th ed. (1982), covers the relation of cataloging and classification. S.R. RANGANATHAN, *Prolegomena to Library Classification,* 3rd ed. (1967), explores the theoretical basis in depth. DAGOBERT SOERGEL, *Indexing Languages and Thesauri* (1974), is an encyclopaedic work with an excellent bibliography.

Services to readers: A useful guide to all types of reference sources is GAVIN HIGGENS (ed.), *Printed Reference Material,* 2nd ed. (1984). Works of more theoretical and philosophical nature include K.J. MCGARRY, *Communication, Knowledge and the Librarian* (1983); and GIRJA KUMAR and KRISHAN KUMAR, *Philosophy of User Education* (1983). BILL KATZ (ed.), *Reference Services in the 1980s* (1982), is an overview of practical issues. J.E. ROWLEY, *Abstracting and Indexing* (1982), treats the subject as an aspect of reader services.

The information profession: The most detailed analysis of all aspects, in very readable form, is found in JESSE H. SHERA, *The Foundations of Education for Librarianship* (1972). See also PATRICIA LAYZELL WARD (ed.), *The Professional Development of the Librarian and Information Worker* (1980). The history of education for the profession is presented in L. HOUSER and ALVIN M. SCHRADER, *The Search for a Scientific Profession: Library Science Education in the U.S. and Canada* (1978); and CAROLYN LEOPOLD MICHAELS, *Library Literacy Means Lifelong Learning* (1985), surveys the role of libraries and librarianship.

(D.J.F.)

Life

The profusion of life on Earth has been studied in great detail, and a number of general principles have been revealed. Foremost among them is the principle of evolution by natural selection—the stepwise adaptation of organisms to their environment with increasing precision by small random mutations, or changes, in their hereditary material—which is the feature that distinguishes living from non-living matter. This article treats first the varieties of definitions of life and then covers, in some detail, the similarities and differences among organisms on Earth. It deals with the problem of the origin of life on Earth and concludes with a consideration of the possibility of life beyond the Earth.

This article is divided into the following sections:

DEFINITIONS OF LIFE

What is known about life

A great deal is known about life. Anatomists and taxonomists have studied the forms and relations of more than a million separate species of plants and animals. Physiologists have investigated the gross functioning of organisms. Biochemists have probed the biological interactions of the organic molecules that make up life on our planet. Molecular biologists have uncovered the very molecules responsible for reproduction and for the passage of hereditary information from generation to generation, a subject that geneticists had previously studied without going to the molecular level. Ecologists have inquired into the relations between organisms and their environments, ethologists the behaviour of animals and plants, embryologists the development of complex organisms from a single cell, evolutionary biologists the emergence of organisms from pre-existing forms over geological time. Yet despite the enormous fund of information that each of these biological specialties has provided, it is a remarkable fact that no general agreement exists on what it is that is being studied. There is no generally accepted definition of life. In fact, there is a certain clearly discernible tendency for each biological specialty to define life in its own terms. The average person also tends to think of life in his own terms. For example, the man in the street, if asked about life on other planets, will often picture life of a distinctly human sort. Many individuals believe that insects are not animals, because by "animals" they mean "mammals." Man tends to define in terms of the familiar. But the fundamental truths may not be familiar. Of the following definitions, the first two are in terms familiar in everyday life; the next three are based on more abstract concepts and theoretical frameworks.

Physiological. For many years a physiological definition of life was popular. Life was defined as any system capable of performing a number of such functions as eating, metabolizing, excreting, breathing, moving, growing, reproducing, and being responsive to external stimuli. But many such properties are either present in machines that nobody is willing to call alive, or absent from organisms that everybody is willing to call alive. An automobile, for example, can be said to eat, metabolize, excrete, breathe, move, and be responsive to external stimuli. And a visitor from another planet, judging from the enormous numbers of automobiles on the Earth and the way in which cities and landscapes have been designed for the special benefit of motorcars, might well believe that automobiles are not only alive but are the dominant life form on the planet. Man, however, professes to know better. On the other hand, some bacteria do not breathe at all but instead live out their days by altering the oxidation state of sulfur.

Metabolic. The metabolic definition is still popular with many biologists. It describes a living system as an object with a definite boundary, continually exchanging some of its materials with its surroundings, but without altering its general properties, at least over some period of time. But again there are exceptions. There are seeds and spores that remain, so far as is known, perfectly dormant and totally without metabolic activity at low temperatures for hundreds, perhaps thousands, of years but that can revive perfectly well upon being subjected to more clement conditions. A flame, such as that of a candle in a closed room, will have a perfectly defined shape with fixed boundary and will be maintained by the combination of its organic waxes with molecular oxygen, producing carbon dioxide and water. A similar chemical reaction, incidentally, is fundamental to most animal life on Earth. Flames also have a well-known capacity for growth.

Biochemical. A biochemical or molecular biological definition sees living organisms as systems that contain reproducible hereditary information coded in nucleic acid molecules and that metabolize by controlling the rate of chemical reactions using proteinaceous catalysts known as enzymes (see BIOCHEMICAL COMPONENTS OF ORGANISMS: *Enzymes*). In many respects, this is more satisfying than the physiological or metabolic definitions of life. There are, however, even here, the hints of counterexamples. There seems to be some evidence that a virus-like agent called scrapie contains no nucleic acids at all, although it has been hypothesized that the nucleic acids of the host animal may nevertheless be involved in the reproduction of scrapie. Furthermore, a definition strictly in chemical terms seems peculiarly vulnerable. It implies that, were a person able to construct a system that had all the functional properties of life, it would still not be alive if it lacked the molecules that earthly biologists are fond of—and made of.

Genetic. All organisms on Earth, from the simplest cell to man himself, are machines of extraordinary powers, effortlessly performing complex transformations of organic molecules, exhibiting elaborate behaviour patterns, and indefinitely constructing from raw materials in the envi-

Instructions for living systems

ronment more or less identical copies of themselves. How could machines of such staggering complexity and such stunning beauty ever arise? The answer, for which today there is excellent scientific evidence, was first discerned by the evolutionist Charles Darwin in the years before the publication in 1859 of his epoch-making work, the *Origin of Species.* A modern rephrasing of his theory of natural selection goes something like this: Hereditary information is carried by large molecules known as genes, composed of nucleic acids. Different genes are responsible for the expression of different characteristics of the organism. During the reproduction of the organism the genes also reproduce, or replicate, passing the instructions for various characteristics on to the next generation. Occasionally, there are imperfections, called mutations, in gene replication. A mutation alters the instructions for a particular characteristic or characteristics. It also breeds true, in the sense that its capability for determining a given characteristic of the organism remains unimpaired for generations until the mutated gene is itself mutated. Some mutations, when expressed, will produce characteristics favourable for the organism; organisms with such favourable genes will reproduce preferentially over those without such genes. Most mutations, however, turn out to be deleterious and often lead to some impairment or to death of the organism. To illustrate, it is unlikely that one can improve the functioning of a finely crafted watch by dropping it from a tall building. The watch may run better, but this is highly improbable. Organisms are so much more finely crafted than the finest watch that any random change is even more likely to be deleterious. The accidental beneficial and inheritable change, however, does on occasion occur; it results in an organism better adapted to its environment. In this way organisms slowly evolve toward better adaptation, and, in most cases, toward greater complexity. This evolution occurs, however, only at enormous cost: man exists today, complex and reasonably well adapted, only because of billions of deaths of organisms slightly less adapted and somewhat less complex. In short, Darwin's theory of natural selection states that complex organisms developed, or evolved, through time because of replication, mutation, and replication of mutations. A genetic definition of life therefore would be: a system capable of evolution by natural selection.

Genetic definition of life

This definition places great emphasis on the importance of replication. Indeed, in any organism enormous biological effort is directed toward replication, although it confers no obvious benefit on the replicating organism. Some organisms, many hybrids for example, do not replicate at all. But their individual cells do. It is also true that life defined in this way does not rule out synthetic duplication. It should be possible to construct a machine that is capable of producing identical copies of itself from preformed building blocks littering the landscape but that arranges its descendants in a slightly different manner if there is a random change in its instructions. Such a machine would, of course, replicate its instructions as well. But the fact that such a machine would satisfy the genetic definition of life is not an argument against such a definition; in fact, if the building blocks were simple enough, such a machine would have the capability of evolving into very complex systems that would probably have all the other properties attributed to living systems. The genetic definition has the additional advantage of being expressed purely in functional terms: it does not depend on any particular choice of constituent molecules. The improbability of contemporary organisms—dealt with more fully below—is so great that these organisms could not possibly have arisen by purely random processes and without historical continuity. Fundamental to the genetic definition of life then is the belief that a certain level of complexity cannot be achieved without natural selection.

Life as an ordered system

Thermodynamic. Thermodynamics distinguishes between open and closed systems. A closed system is isolated from the rest of the environment and exchanges neither light, heat, nor matter with its surroundings. An open system is one in which such exchanges do occur. The second law of thermodynamics states that, in a closed system, no processes can occur that increase the net order (or decrease the net entropy) of the system (see THERMODYNAMICS). Thus the universe taken as a whole is steadily moving toward a state of complete randomness, lacking any order, pattern, or beauty. This fate has been known since the 19th century as the heat death of the universe. Yet living organisms are manifestly ordered and at first sight seem to represent a contradiction to the second law of thermodynamics. Living systems might then be defined as localized regions where there is a continuous increase in order. Living systems, however, are not really in contradiction to the second law. They increase their order at the expense of a larger decrease in order of the universe outside. Living systems are not closed but rather open. Most life on Earth, for example, is dependent on the flow of sunlight, which is utilized by plants to construct complex molecules from simpler ones. But the order that results here on Earth is more than compensated by the decrease in order on the sun, through the thermonuclear processes responsible for the sun's radiation.

Some scientists argue on grounds of quite general open-system thermodynamics that the order of a system increases as energy flows through it, and moreover that this occurs through the development of cycles. A simple biological cycle on the Earth is the carbon cycle. Carbon from atmospheric carbon dioxide is incorporated by plants and converted into carbohydrates through the process of photosynthesis. These carbohydrates are ultimately oxidized by both plants and animals to extract useful energy locked in their chemical bonds. In the oxidation of carbohydrates, carbon dioxide is returned to the atmosphere, completing the cycle. It has been shown that similar cycles develop spontaneously and in the absence of life by the flow of energy through a chemical system. In this view, biological cycles are merely an exploitation by living systems of those thermodynamic cycles that pre-exist in the absence of life. It is not known whether open-system thermodynamic processes in the absence of replication are capable of leading to the sorts of complexity that characterize biological systems. It is clear, however, that the complexity of life on Earth has arisen through replication, although thermodynamically favoured pathways have certainly been used.

The existence of diverse definitions of life surely means that life is something complicated. A fundamental understanding of biological systems has existed since the second half of the 19th century. But the number and diversity of definitions suggest something else as well. As detailed below, all the organisms on the Earth are extremely closely related, despite superficial differences. The fundamental ground pattern, both in form and in matter, of all life on Earth is essentially identical. As will emerge below, this identity probably implies that all organisms on Earth are evolved from a single instance of the origin of life. It is difficult to generalize from a single example, and in this respect the biologist is fundamentally handicapped as compared, say, to the chemist or physicist or geologist or meteorologist, who now can study aspects of his discipline beyond the Earth. If there is truly only one sort of life on Earth, then perspective is lacking in the most fundamental way.

LIFE ON EARTH

Mechanism and vitalism. Human beings are ambulatory collections of some 10^{14} cells. Human cells are in many fundamental respects similar to those that make up all the other animals and plants on the Earth. Each cell typically consists of a central, usually spherical, nucleus and an outer more heterogeneous region, termed the cytoplasm. The substance of nucleus and cytoplasm together has for many decades been called protoplasm. Use of this term implied that there was some special substance underlying living organisms. In the use of the word protoplasm there is occasionally an implication that life cannot be explained solely by physics and chemistry, that some mysterious "vital force" must be invoked. A living cell is a marvel of detailed and complex architecture. Seen through a microscope there is an appearance of almost frenetic activity. On a deeper level it is known that molecules are being synthesized at an enormous rate. Almost any enzyme catalyzes the synthesis of more than

100 other molecules per second. In 10 minutes, a sizable fraction of the total mass of a metabolizing bacterial cell has been synthesized. The information content of a simple cell has been estimated as around 10^{12} bits, comparable to about a hundred million pages of the *Encyclopædia Britannica.* Faced with all this or its equivalent, it is not surprising that early biologists felt despair at ever being able to understand the detailed workings of life.

A Stone Age man, confronted for the first time with a watch, might also deduce that there was some special watch substance in nature, or perhaps even a god of the watch. In ancient times, the most common of biological activities, such as the hatching of an egg or the blooming of a flower, were attributed to the intercession of a deity. After the epochal work of Sir Isaac Newton, when the motion of the planets and comets of the solar system was predictable to some very great precision and understood on the basis of an underlying principle, the idea developed that organisms were also nothing more than a particularly intricate kind of clockwork. But when early investigations failed to unveil the clockwork, a kind of ghostly mainspring was invented—the "vital force." This force was a rebellion from mechanistic biology, an explanation of all that mechanism could not explain or for which mechanism could not be found. It also appealed to those who felt debased by the implication that they were "nothing more" than a collection of atoms, that their urges and apparent free wills arose merely from the interaction of an enormously large number of molecules in a way that, although too complex to use predictably, was in principle determined.

"Vital force"

Not only is there no evidence for a vital force but the idea itself is hardly thought out; it is a sort of catchall concept, covering anything otherwise inexplicable. The alternative approach, that all organisms are made of atoms and nothing else, has proven especially useful and has led to a fundamental new understanding of biological systems. This situation does not imply, of course, that atoms cannot be put together in so complex a way that their collective behaviour is too difficult to understand in terms of the individual atoms; in this sense there may be particular laws of biology not readily derivable from the elementary interaction of atoms. But this is a very different thing from a vital force. Indeed, there is nothing debasing in the thought that a person is made of atoms alone; it means that one is intimately connected with the matter that comprises the inanimate universe. What a wonder that atoms can be put together in so complex a pattern as to produce human beings. Man is a tribute to the subtlety of matter. As the American anthropologist Loren Eiseley has written, ". . . if 'dead' matter has reared up this curious landscape of fiddling crickets, song sparrows, and wondering men, it must be plain even to the most devoted materialist that the matter of which he speaks contains amazing, if not dreadful powers. . . ." (*The Immense Journey,* Random House, New York, 1957.)

Nucleic acids. It is now known that many if not all of the fundamental properties of cells are a function of their nucleic acids, their proteins, and the interactions among these molecules. Within the nuclear regions of cells is a mélange of twisted and interwoven fine threads, the chromosomes. During cell division, in all but the simplest organisms, the chromosomes display an elegantly choreographed movement, separating so that each daughter cell of the original cell receives an equal complement of chromosomal material. This pattern of segregation corresponds in all details to the theoretically predicted pattern of segregation of the genetic material implied by the fundamental genetic laws (see GENETICS AND HEREDITY). The chromosomes are composed of nucleic acids and proteins in a combination called nucleoprotein. The nucleic acid stripped of its protein is known to carry genetic information and to regulate cellular metabolism; the protein in nucleoprotein undoubtedly plays some secondary, probably regulatory, role.

The role of nucleic acids

The specific carrier of the genetic information in higher organisms is a nucleic acid known as DNA, short for deoxyribonucleic acid. DNA is a double helix, two molecular coils wrapped around each other and chemically bound one to another by bonds connecting adjacent bases. Each helix has a backbone that consists of a long sequence of alternating sugars and phosphates. Attached to each sugar is a base. Each sugar-phosphate-base combination is called a nucleotide; a nucleic acid strand can be thought of as a sequence of nucleotides. There is a very significant one-to-one base pairing in the connection of adjacent helices, in the sense that once the sequence of bases along one helix is specified, the sequence along the other is also specified. The specificity of base pairing plays a key role in the replication of the DNA molecule, where each helix makes an identical copy of the other from molecular building blocks in the cell. These nucleic acid replication events are mediated by enzymes, and with the aid of enzymes have been produced in the laboratory.

Ribonucleic acid (RNA) differs from DNA in having a slightly different five-carbon sugar, and in replacing one of the four bases that make up DNA by a slightly different base. RNA does not appear to exist in a double-stranded form. Now DNA, RNA, and the enzymes have a curiously interconnected relation, which appears ubiquitous in all organisms on Earth today.

Commonalities among organisms on Earth. The genetic code was broken in the 1960s. It was found that three consecutive nucleotides code for one amino acid of a protein molecule; *e.g.,* an enzyme. By controlling the synthesis of enzymes, the nucleic acids control the functioning of the cell. Of the four different bases taken three at a time, there are $4^3 = 64$ possible combinations. The meaning of each of these combinations, or codons, is known. Most of them represent a particular amino acid. A few of them represent punctuation marks; for example, instructions to start or stop a synthesis. Some of the code is degenerate in the sense that more than one nucleotide triplet may specify a given amino acid. These interactions among nucleic acids and proteins seem absolutely central to living processes on Earth today. Not only are these processes apparently the same in all organisms on Earth but even the particular dictionary that is used for the transcription of nucleic acid information into protein information seems to be essentially the same in all organisms. Moreover, this code has various chemical advantages over other conceivable codes. The complexity, ubiquity, and advantages of these processes clearly argue that the present interactions among proteins and nucleic acids are themselves the product of a long evolutionary history. At the time of the origin of life this very complex replication and transcription apparatus could of course not have been in operation. A fundamental problem in the origin of life is the question of the origin and early evolution of the genetic code.

There are many other commonalities among organisms on Earth. For example, there is only one class of molecules that store energy for biological processes until the cell has use for it, and these molecules are all nucleotide phosphates. The most common example is ATP (adenosine triphosphate). For this very different function, a molecule identical to the building blocks of the nucleic acids is employed. There are metabolically important molecules—*e.g.,* molecules known as FAD (flavin adenine dinucleotide) and coenzyme A—which include subunits similar to the nucleotide phosphates. Porphyrins represent another category of ubiquitous molecules. Porphyrins are the chemical basis of hemoglobin, which carries oxygen molecules through the bloodstreams of animals; of chlorophyll, which is the fundamental molecule mediating photosynthesis in plants; and of the colours that many animals display. The left- or right-handedness of many biological molecules—discussed more fully below—runs identically through all organisms on Earth. In fact, of the billions of possible organic compounds, less than 1,500 are employed by contemporary life on Earth, and these 1,500 are constructed from less than 50 simple molecular building blocks. Similarly, organisms as diverse as paramecia and human sperm cells have little whiplike appendages called cilia or flagella used to propel themselves through their liquid environments. The cross-sectional structure of the cilia and flagella is almost always nine pairs of peripheral and one pair of internal fibres. There is no immediately obvious selective advantage of the 9 : 1 ratio.

Energy-storing molecules

These commonalities indicate that a few basic chemical and functional patterns are being used over and over again, reflecting the extremely close relations all organisms on Earth have to one another. Many biologists believe that these commonalities—particularly where no obvious selective advantage exists—imply that all organisms on the Earth are descendants of a single common ancestor. But it is possible that there are more subtle selective advantages. The issue may have to await the first detailed study of an extraterrestrial organism.

Chromosome possibilities

The number of possible ways of putting nucleotides together in a chromosome is enormous. The renowned geneticist H.J. Muller estimated that in a human chromosome there are about 4×10^9 base pairs. Each base pair position could be filled by any one of four possible bases; accordingly, the number of possible varieties of human chromosomes is $4^{4 \times 10^9}$ ($10^{2.4 \times 10^9}$), an inconceivably large number. By contrast, the number of elementary particles (electrons and protons) in the entire physical universe is only about 10^{80}. Thus a human being is an extraordinarily improbable object. Most of the $10^{2.4 \times 10^9}$ possible sequences of nucleotides would lead to complete biological malfunction. Our nucleotides work because natural selection, over a 4,000,000,000-year history of life, has destroyed enormous numbers of combinations that did not work. But there still may be combinations that work far better than any now present, and the future holds the promise that man will be able to assemble nucleotides in any desired sequence to produce whatever characteristics of human beings are thought desirable, an awesome and disquieting prospect.

Metabolism. The chemical bonds that make up living organisms have a certain probability of spontaneous breakage. Accordingly, mechanisms must exist to repair this damage, or to replace the broken molecules. In addition, the meticulous control that cells exercise over their internal activities requires the continued synthesis of new molecules. These processes of synthesis and breakdown of the organic molecules of the cell are collectively termed metabolism, and for synthesis to keep ahead of the thermodynamic tendencies toward breakdown, energy must be supplied to the living system. Organisms acquire this energy by two general methods. Some organisms are heterotrophs, acquiring their energy by the controlled breakdown of pre-existing organic molecules (food)—generally those supplied by other organisms. Human beings and most other animals are heterotrophs. Alternatively, organisms may be autotrophs, acquiring their useful free energy from some other source, either from the energy of sunlight, in which case the organism is called a photoautotroph, or from the controlled chemical reaction of inorganic materials, in which case the organism is known as a chemoautotroph. Organisms that use both modes are called photochemoautotrophs.

Methods of acquiring energy

A green plant is a typical example of a photoautotroph. It uses sunlight to break water into oxygen and hydrogen. Hydrogen is then combined with carbon dioxide to produce such energy-rich organic molecules as ATP and carbohydrates, and the oxygen is released back into the atmosphere. Many animals, on the other hand, utilize the atmospheric oxygen to combine chemically with organic materials they have eaten and release carbon dioxide and water as waste products in extracting energy from the organic materials. This is an example of an ecological cycle in which a material (here carbon) is pumped through two different organisms.

More generally, such metabolic cycles—used by the organism to extract useful energy from the environment—can be described in terms of oxidation-reduction reactions. In the case of respiration, molecular oxygen accepts electrons from glucose or other sugars. The oxygen is said to be an electron acceptor (it has a great affinity for electrons), the glucose an electron donor. This is the prototype of oxidation-reduction reactions, but not all such reactions necessarily involve oxygen. Biological electron acceptors other than oxygen include nitrates, sulfates, carbonates, nitrogen, and methanol. Biological electron donors other than sugars include nitrogen, sulfides, methane, ammonia, and methanol. For acceptor-donor transformations to occur over any period of time, biological cycles are necessary. It is possible that, for geologically short periods of time, organisms have lived off a finite supply of material, but for any long-term continuance of life, a dynamic cycling of matter, involving at least two different varieties of organisms, is necessary. If there is life on other planets, a similar cycling must exist. A search for such molecular transformations is one method of detecting extraterrestrial life.

On the Earth, all such useful biological electron transfer reactions lead to the net production of one or more molecules of ATP. Two of the three phosphates of this molecule are held by "energy-rich" bonds sufficiently stable to survive for long periods of time in the cell, but not so strong that the cell cannot tap these bonds for energy when needed. ATP and very similar molecules, all of them having a base, a five-carbon sugar, and three phosphates, are, so far as is known, the general and unique energy currency of living systems on Earth.

Energy from metabolic processes

Metabolic processes do not occur in one step. The ordinary sugar, glucose, is not oxidized to carbon dioxide and water by living cells in the same way that occurs if a flame is applied to glucose in air. The resulting release of energy would be much too sudden, and concentrated in too small a volume, for such a process to be utilized safely by the cell. Instead, the glucose is broken down by a series of successive and coordinated steps, each mediated by a particular and specific enzyme. In almost all organisms that metabolize glucose, the sugar is first broken down in a set of anaerobic steps (that is, in the absence of oxygen). The total number of such steps is about 11. Some organisms are anaerobes; that is, they do not utilize molecular oxygen. In them the anaerobic steps are as far as the glucose metabolism is carried. Other organisms, including man, carry the oxidation of glucose further, gingerly combining glucose breakdown products with molecular oxygen. Such aerobic oxidation of glucose requires about 60 more enzymatically catalyzed steps. Another indication of the relative simplicity of the anaerobic breakdown of sugar is that all the enzymes used are free in solution in the cell; the aerobic steps use enzymes that are localized in specific regions of the cell. The complete aerobic breakdown of sugar to carbon dioxide and water is about 10 times more efficient than the breakdown accomplished by anaerobes; 10 times as many ATP molecules are produced. Similar themes and variations exist for the metabolism of other molecules (see METABOLISM).

Efficiency of aerobic oxidation

The energy made available in this way to ATP is used in a variety of ways by the cell; for example, for motility. When an amoeba extends pseudopods, or a person walks, ATP molecules are being tapped for their energy-rich phosphate bonds. In addition, ATP molecules are used for the synthesis of molecules that the organism needs and does not have available. Among such molecules may be amino acids, the particular five-carbon sugars involved in nucleic acids, the nucleic acid bases, and so on. Each of these synthetic processes is again controlled and enzymatically mediated and may start from a variety of building blocks available to the organism, some simple, some more complex. For example, the amino acid L-leucine is produced from pyruvic acid, itself the product of the anaerobic breakdown of glucose. Synthesis of L-leucine from pyruvic acid involves eight enzyme-mediated steps and the addition of acetic acid and water.

These exquisitely interlocked and controlled metabolic steps are not usually performed in a diffuse manner all over the cell. Instead there is, at least in all higher organisms, a marvellously architectured cellular interior with particular specialized regions where particular chemical reactions are performed. Those oxidation-reduction reactions that involve molecular oxygen occur in an inclusion within the cytoplasm called the mitochondrion. The mitochondrion itself has an intricate substructure, and particular enzymes are thought to reside in particular sites within it; the molecule being metabolized may be passed on from one enzyme to another as through a conveyor belt in a factory. Similarly, photosynthesis occurs in a cytoplasmic inclusion called a chloroplast, which contains the chlorophyll and other pigments that absorb visible light, as well

The work of the mitochondrion

as the detailed enzymatic apparatus for the photosynthetic process. Chloroplasts and mitochondria, as well as other cytoplasmic inclusions at the base of flagella and cilia, all contain DNA. Moreover, this DNA has a somewhat different distribution of bases from that of the nucleus. It has been suggested that the cytoplasmic inclusions are the remnants of once free-living forms that, because of the favourable conditions found there, have taken up residence in the insides of other organisms.

Nucleic acids are known to pass from cell to cell and to perform their replication and coding functions efficiently in the new cell. In fact, viruses are essentially strands of nucleic acid, with a protein coat, that operate in just this way. It is also known that pieces of the genetic material from one cell may migrate into another cell of the same species and produce genetic and permanently heritable changes there. Alternatively, part of the virus nucleic acid may be permanently bound to the nuclear DNA of a host cell. It is likely that a virus is a degenerate form, now highly specialized to live off a specific host, of an organism once free-living and much more generally capable of performing a wide range of metabolic tasks. A virus must use the genetic transcription apparatus of its host cell. Many viruses do this extremely efficiently, turning a bacterium from a factory for making other bacteria into a factory for making viruses. In some cases it takes no more than 10 minutes for a bacterium infected by a single virus to produce a hundred new virus particles, which then burst forth from the host bacterium, destroying it. The line between benign or useful cytoplasmic inclusions and infective agents is not a very sharp one (see VIRUSES).

Differences in nuclear structure

Eucaryotes and procaryotes. In the very simplest one-celled organisms one may distinguish between eucaryotic and procaryotic cells. Many familiar one-celled organisms, such as paramecia and amoebas, as well as the cells of all higher organisms including man, are eucaryotic. Such cells undergo mitosis, a fundamental sequence of events that occurs after DNA replication and that ensures that the DNA is precisely and equally distributed to the daughter cells. Eucaryotic cells have nucleoprotein in their nuclei. There is a membrane that separates the nucleus from the cytoplasm. Mitochondria are generally present in the cytoplasm, as is a very intricately convoluted structure, called the endoplasmic reticulum, that probably serves as the anchoring point for many cytoplasmic enzymes not contained in such inclusions as mitochondria or chloroplasts.

On the other hand there are procaryotic cells, which are most generally typified by the bacteria and the blue-green algae. In these cells nuclear division is nonmitotic, there is no nucleoprotein, and a nuclear membrane is absent. While eucaryotic cells may have more than one chromosome, procaryotic cells have one chromosome only, and that one is dispersed in the cytoplasm. Mitochondria, chloroplasts, and the endoplasmic reticulum are always absent. It is clear from this description that the procaryotes are in many respects more primitive than the eucaryotes. A basic unsolved evolutionary question concerns the evolution of procaryotes into eucaryotes.

The smallest organism

An interesting subject of biological speculation concerns what the smallest and simplest contemporary free-living organism might be. The smallest free-living cells now known are the pleuropneumonia-like organisms (PPLO). While an amoeba has a mass of 5×10^{-7} grams (1 gram = 0.035 ounce), a PPLO weighs 5×10^{-16} grams and is only about $^1/_{10}$ of a micrometre across. It can be seen only in the electron microscope. Such organisms grow very slowly. There may be smaller organisms that grow even more slowly, but they would be extremely difficult to detect. Even an organism of the size of PPLO has room for only about a hundred enzymes. A much smaller organism would have room for many fewer enzymes, and its ability to accomplish the functions that contemporary living systems must accomplish would be severely compromised. Were there, however, an environment in which all the necessary organic building blocks and such energy sources as ATP were provided "free," a functioning organism could be substantially smaller than PPLO. In fact the inside of a cell provides just such an environment; this is why infectious agents, such as viruses, can be substantially smaller than PPLO. But it must be emphasized that such agents are not free-living organisms.

Metazoa, embryology, and sex. The distinction between single-celled and many-celled organisms (in animals, between protozoa and metazoa) is far from a sharp one. An interesting illustration is the slime molds, which undergo an extraordinary sequence of events during their life cycle. The cycle begins with single cells, somewhat like amoebas, which swarm, or combine, into a slimy mass with many nuclei called a plasmodium. The plasmodium in turn forms a sluglike mass that is certainly a multicelled organism. The slug develops into a stalked, fruitlike sporangium, still multicellular. The sporangium produces spores with cellulose cell walls similar to those of plants. The spores in turn germinate into small cells bearing flagella. The flagella are lost and the life cycle is completed with the production of an amoeboid form (for details, see PROTOPHYTES: *Slime molds*).

Biology is replete with life cycles of comparable complexity. The swarming of individual cells to form a plasmodium may in fact be an example of the events that led to the production of metazoa in the early history of the Earth. Such life cycles, while apparently very exotic, are shared by many organisms, including man, where a one-celled, free-swimming sperm stage is part of the life cycle.

The problem of development

The life cycle of slime molds, or men, or any other multicellular organism, brings up a fundamental and still largely unsolved problem. These organisms develop from a single cell that has a single complement of the genetic material. These cells then divide, forming many identical cells. The very early embryology of man goes through stages with 2, 4, 8, 16, etc., cells. Since the genetic information is identical in each cell, how does it ever happen that cells become specialized, forming hair cells, teeth cells, liver cells, blood cells, or bone cells? How can any given cell "know" what sort of specialized cell it must become, since all cells contain identical nucleic acids? Possibly the answer to this question has to do with geometry. After the 16- or 32-cell stage, there is a distinct difference between a cell on the inside of the embryo, which is entirely surrounded by cells, and a cell on the outside of the embryo, which is not entirely surrounded by other cells. One of the earliest major steps in embryonic development is a distinction in function between interior cells (the endoderm) and exterior cells (the ectoderm). There are physical and chemical interactions among adjacent cells. Perhaps any cell then has the capability of becoming any specialized cell, but cells are, as a result of their external cellular environment, called upon to develop in different ways. Occasional embryonic anomalies or cysts occur in which, for example, hair or teeth develop in totally inappropriate portions of the body. Similarly, eyes have been caused to develop on the limbs of frogs. Such incidents demonstrate the capability of the "wrong" cells to produce particular cellular specializations (see GROWTH AND DEVELOPMENT).

Much of the beauty and diversity of contemporary life on Earth is due to sex. A totally asexual organism will be genetically identical to its (single) parent except for occasional mutations. The development of any major new adaptation would then require the acquisition of large numbers of appropriate mutations. Consider, for example, how the ability of an organism to metabolize a given molecule depends on the interaction of many enzymes, each produced by the transcription of the genetic information in hundreds of nucleotides, each nucleotide being the product of a single mutation. Thus, the chance development of any advantageous adaptation in an asexual organism requires the mutations to wait in line for a fortuitous juxtaposition.

Sex: an added dimension

Sex solves this problem in an elegant way. The genetic material of the parents is reassorted so that totally new combinations of genes are produced. In this way mutations acquired by any member of the population are rather quickly distributed to other members, and mutations arising in separate organisms can be combined. The likelihood of producing a useful sequence of mutations is thereby greatly enhanced. The advantages of sexual reproduction are so great that even many simple forms, such as bacteria or protozoa, which largely reproduce asexually, have occasional sexual encounters. While two sexes are clearly adequate for such a random assortment of genetic material, some organisms have developed more sexes: paramecia, for example, have somewhere between five and 10 sexes, defined in terms of the elaborate taboos about which organisms can combine their genetic material. In the process of genetic reassortment, some organisms make a very large number of attempts; for example, frogs lay millions of eggs at a time, and the number of sperm cells in a single human ejaculation is about 3×10^8 (see SEX AND SEXUALITY).

The varieties of organisms and environments. The environment of the Earth is heterogeneous. There are mountains, oceans, and deserts, extremes of temperature and humidity. In addition, there are diverse microenvironments: oxygen-depleted oceanic oozes, ammonia-rich soils, mineral deposits with a high radioactivity content, and so on. The environment of an organism also includes the other organisms in its surroundings. For each of these environmental situations there are corresponding ecological niches, and the variety of ecological niches populated on the Earth is quite remarkable. Furthermore, ecological niches can be filled independently several times. For example, quite analogous to the ordinary mammalian wolf is the marsupial wolf that lives in Australia; the two have striking similarities in physical appearance and in predation behaviour. As another example, the same streamlined shape for high-speed marine motion has evolved independently at least three times: in *Stenopterygius* and other Mesozoic reptiles; in the tuna, which are fish; and in the dolphins, which are mammals. This case of convergent evolution must arise from the fact that hydrodynamics admits a narrow range of solutions to the problem of high-speed marine motion by large animals. Similarly, the eye has independently evolved several times among animals on the Earth; apparently such a structure is the best solution to the problem of visual recording. In those cases where physics or chemistry establishes one most efficient solution to a given ecological problem, natural selection will often tend to reach the solution, but not always. Some adaptations of undoubted utility, such as tractor treads in swampy environments, have never been evolved by natural selection on the Earth.

Life in extreme environments

There is an extraordinarily wide range of ecological niches to which organisms have adapted through the operation of natural selection. The same basic fabric of life has been used to produce very diverse organisms. The alga *Cyanidium caldarium* can grow in concentrated solutions of hot sulfuric acid. Other bacteria, algae, and fungi can live in extremely acidic (pH of 0) or extremely alkaline (pH near 13) environments. Procaryotic bacteria live in pools at Yellowstone National Park at temperatures above 90° C (194° F), almost at the boiling point of water. Sulfate-reducing bacteria are reported to grow and reproduce at 104° C (219° F) under very high pressures. Many organisms employ organic or inorganic antifreezes to lower the freezing point of their internal liquids, so that they can live at several tens of degrees below 0° C (32° F). Some insects use dimethyl sulfoxide as an antifreeze. Other organisms live in briny pools in which dissolved salts lower the freezing point. For example, Don Juan Pond in Antarctica has about one molecule of calcium chloride for every two water molecules and does not freeze until −45° C (−49° F). It contains a possibly unique microflora that continues to metabolize at least down to −23° C (−9° F). Biological activity does not cease at the freezing point of water; in fact some enzymes are actually more active in ice than in water. Many single-celled organisms can be frozen indefinitely to extremely low temperatures—the temperature of liquid air for example—and then be thawed with no decrease in activity. The primary damage that freezing causes is apparently due to the unavailability of liquid water and to the expansion and contraction attendant to freezing and thawing. Some arthropods can be severely dehydrated and then revived simply by adding water. In the dehydrated state they can be brought to any temperature from close to absolute zero to above the boiling point of water without apparent damage. When encysted in response to dehydration, some such organisms seem indistinguishable from a weathered grain of sand.

The great majority of familiar organisms on the Earth, however, are much more sensitive to the temperature of their surroundings. Warm-blooded animals internally regulate their temperatures for this reason. A human being whose body temperature drops below 30° C (86° F) or rises above 40° C (104° F) soon dies. Organisms that inhabit cold climates have special insulating layers of fat and fur. Other organisms adapt to seasonal temperature changes by producing dormant forms, such as spores or eggs, to survive the low temperatures. In all cases dormancy appears to be accompanied by dehydration.

Importance of water

Since organisms are composed largely of water, the availability of water is clearly a limiting factor. Here also, however, remarkable adaptations exist. Certain microorganisms can live on the water adsorbed on a single crystal of salt. Other organisms, such as the kangaroo rat and the flour beetle, obtain no water at all in the liquid state, relying entirely on metabolic water; that is, on water released from chemical bonds through the metabolism of food. A variety of plants, including Spanish moss, live in environments where they have no contact with groundwater—for example, on telephone wires—apparently extracting water directly from the air, although such plants require a relatively high humidity. Plants that live in deserts and other very dry environments have evolved wide-spreading root systems that adsorb subsurface water from a great volume of adjacent soil.

Organisms have been found from the stratosphere to the ocean depths. Bacteria and fungal spores have been discovered near the base of the stratosphere by balloons, and searches for organisms at much greater altitudes (up to 100,000 feet) have been attempted with ambiguous results. Birds have been observed flying at altitudes as great as 27,000 feet, and jumping spiders have been found at 22,000 feet on Mt. Everest. At the opposite extreme, microorganisms, fish, and a variety of other metazoa have been recovered from the ocean depths down to thousands of feet, where the corresponding pressures are hundreds of times that at sea level. At these depths no light can penetrate and the organisms, some of which are quite large and include unique phosphorescent adaptations to the dark, ultimately live off particles of organic matter raining down from the upper reaches of the oceans.

There is a range of adaptations to the radiation environment of the Earth. Some microorganisms are readily killed by the small amount of solar ultraviolet light that filters through the Earth's atmosphere at wavelengths near 3,000 angstrom units (Å; 1 Å = one ten-billionth of a metre). On the other hand, the bacterium *Pseudomonas radiodurans* thrives in the large neutron flux at the cores of swimming-pool reactors, to the continuing annoyance of nuclear physicists. Organisms can avoid radiation by shielding. For example, some algae and some desert plants live under a superficial coating of soil or rocks that are more transparent to visible light than to ultraviolet light. In addition, organisms have active methods of undoing the damage produced by radiation. Some of these repair mechanisms work in the dark; others require visible light. The usual reason for the ultraviolet sensitivity of organisms is that their nucleic acids absorb ultraviolet light very effectively at a wavelength near 2,600 Å. Generally speaking, there is an upper limit to the amount of ionizing radiation (such as gamma rays, X-rays, electrons or protons) that an organism can receive without being killed: in the vicinity of 1,000,000 roentgens. Such a lethal dose applies only to extremely radiation-resistant microorganisms; mammals, for example, are killed by much lower doses because there is more that can go wrong with such complex organisms. A lethal dose of ionizing radiation for human beings is a few hundred roentgens applied to the whole body. A thermonuclear weapon dropped on a populated area may deliver, through direct radiation and fallout, doses of a few hundred roentgens or more to people within a radius of some tens of miles of the target. Much smaller doses can produce a variety of diseases and predominantly deleterious mutations in the hereditary material. Moreover, the effect of small doses is cumulative. But until very recently human beings have not lived in environments characterized by large doses of ionizing radiation (see RADIATION).

Size range of living things

The sizes of organisms on the Earth vary greatly. As discussed above, the smallest free-living organisms on the Earth, PPLO, are about 1,000 Å in diameter; a limitation on the size of the smallest free-living organism is its volume: it must contain all the molecules necessary for metabolism. A variety of influences place an upper limit to the size of organisms. One is the strength of biological materials. Galileo calculated in 1638 that a tree taller than roughly 300 feet (91.4 metres) would, when displaced slightly from the vertical (for example, by a breeze), buckle under its own weight. (Sequoias, some of which exceed 300 feet, are apparently near the upper limit of height for an organism.) Because of the buoyancy of water, large whales are not presented with such stability problems, but other difficulties arise. For a fixed shape, the volume of tissues to be nourished increases as the cube of the characteristic length of the organism, but the surface of the gut, which adsorbs the ingested food, increases only as the square of the length. As the length is increased, a point of diminishing returns is ultimately reached.

The range of organic molecules that organisms on Earth can metabolize is very wide and occasionally includes such foods as formaldehyde or petroleum, which seem unlikely from a human point of view. *Pseudomonas* bacteria are capable of using almost any organic molecule as a source of carbon and of energy, provided only that the molecule is at least slightly soluble in water. Microorganisms cannot metabolize plastics, not because of any fundamental chemical prohibitions but probably because plastics have not been part of the environment of microorganisms for very long. Man tends to think of oxygen as extremely important for life, but there are facultative anaerobes that can take their oxygen or leave it, and obligate anaerobes that are actually poisoned by oxygen. Such organisms use a variety of alternative electron acceptors, as previously discussed.

Chemical constituents

The water content of organisms usually represents between 50 and 90 percent of the live weight. Unless there is a massive mineral skeleton, the dry matter of organisms constitutes about one-half carbon by weight, reflecting the fact that organic molecules are based upon carbon. A wide variety of other chemical elements are used for diverse functions. Amino acids are made of nitrogen and sulfur in addition to carbon, hydrogen, and oxygen. Nucleic acids, as has been seen, employ phosphorus in addition to hydrogen, nitrogen, oxygen, and carbon. Sodium and potassium are used in maintaining the electrolyte balance, and calcium and silicon as structural materials. Iron plays a fundamental role in the transport of molecular oxygen as part of the hemoglobin molecule. In some ascidians (sea squirts), however, vanadium replaces iron. Ascidian blood also contains unusually large amounts of niobium, tantalum, titanium, chromium, manganese, molybdenum, and tungsten. The vanadium and niobium compounds in ascidian blood may be adaptations to low oxygen levels. Occasional organisms use selenium or tellurium as electron acceptors; others may produce the fully saturated gas hydrides of arsenic, phosphorus, or silicon, as metabolic wastes. Still others form compounds of carbon with such halogens as chlorine or iodine. Many of the foregoing elements, plus copper, zinc, cobalt, and possibly gallium, boron, and scandium, perform particular functions in the enzymatic apparatus of cells. Many of these elements, both the uncommon ones and those as common as phosphorus, are very highly concentrated in organisms over their general availability in the environment. This concentration must indicate that such chemicals play unique functional roles where other more abundant elements will not serve.

Behaviour and sensory capabilities. Analogous to the wide range of physiological adaptations and the great variety of elements used by organisms on Earth, there is an enormous range of behaviour patterns and sensory capabilities. Coded into its nucleic acids is the information that allows a bird raised from the egg in the absence of other birds to migrate when migration time arrives, to build a nest characteristic of its species, or to engage in elaborate courtship rituals. Those birds that do not perform acceptably do not leave descendants. Such behavioral information must itself have evolved. Rats that pass through mazes easily can be interbred, as can rats that pass through with difficulty; eventually two populations with inherited characteristics called "maze-smart" and "maze-dumb" will be produced. Fruit fly populations attracted to the light can be separated from those that avoid light. Classical genetic crossing experiments reveal that the two populations differ largely in a small number of genes for phototropism. Similar genetic determinants of behaviour exist in man. Possession of a supernumerary Y-chromosome in males is strikingly correlated with aggressive tendencies—which may, however, have been a selective advantage in more primitive societies. Myopia may have had strong survival value in earlier times: near-sighted males, useless in the hunt, stayed home and painted the walls of the cave. As technology develops, natural selection enters new behavioral arenas; for example, in an age of artificial contraception, the clumsy and forgetful preferentially reproduce.

Response to the electromagnetic spectrum

Human beings use only a small part of the total electromagnetic spectrum, the part called visible light, which extends from about 4,000 to about 7,000 Å in wavelength. While many plants and animals are sensitive to this same range of wavelengths, many of them are sensitive to other wavelengths as well. Most insects are sensitive to ultraviolet light at wavelengths below 4,000 Å, and many flowering plants take advantage of this fact and present patterns visible only in the ultraviolet range. Honeybees use polarized light, which the unaided human eye is quite unable to detect, for direction finding on partly cloudy days. The "pit" of such pit vipers as the rattlesnake is an infrared receptor and direction finder. These reptiles can sense the thermal radiation emitted by warm-blooded prey, radiation to which human beings are completely insensitive.

It is common knowledge that some animals (for example, dogs) are sensitive to sounds that the human ear cannot detect. Bats emit and detect sound waves at ultrahigh frequencies, in the vicinity of 100,000 cycles per second, about five times the highest frequency to which the human ear is sensitive. Bats use these sounds not so much to communicate, however, as to echolocate their prey and were doing this for millions of years before radar and sonar were invented. The audio receptors of many moths that are prey to bats are responsive only to the frequencies emitted by the bats. When the bat sounds are heard, the moths take evasive action. Dolphins have a very wide frequency range and several communication channels, as well as a "click" echolocator. Dolphins and whales use their blowholes rather than their mouths to utter these sounds. Sharks and other marine predators are said to locate their prey by the low-frequency sounds the prey makes when in distress. Some animals develop highly specialized and exotic organs for the detection or transmission of sound—for example, a European grasshopper has a relatively large parabolic antenna on its back that looks very much like a small radio telescope. This antenna is used for producing noises evidently thought attractive by the female of the species.

Many organisms are capable of smell and taste; that is, the detection of specific chemical molecules. According to one theory of smell, there are particular olfactory sensors, each receptive only to a specific chemical group on airborne molecules. The ultimate in olfactory specialization is probably the male silkworm moth: with its feathery antennae it is able to smell essentially nothing except the chemical sex attractant discharged by the female of the species. But it can detect this molecule very well, needing an impact of only 40 molecules per second on its antennae to produce a marked response. One female silkworm moth need release only 10^{-8} grams of sex attractant per second in order to attract every male silkworm moth in a volume hundreds of metres to kilometres on a side.

Extraordinary senses

Besides the senses of sight, hearing, smell, taste, and touch, various animals have a wide variety of other senses (see SENSORY RECEPTION). Man has an inertial orientation system and accelerometer in the cochlear canal of the ear. The water scorpion (*Nepa*) has a fathometer sensitive to hydrostatic pressure gradients. Most higher plants have chemically amplified gravity sensors. Fireflies and squids communicate with their own kind by producing time sequences or patterns of light on their bodies. The African freshwater fish *Gymnarchus niloticus* operates a dipole electrostatic field generator and a sensor to detect the amplitude and frequency of disturbances in the impressed field, an adaptation well suited for its nocturnal activities in turbulent waters. Other organisms have salinity sensors, or humidity sensors. There may be sensors involved in homing instincts of animals that have not yet been discovered. All of these senses confer upon their possessors an awareness of the environment that may be very different from that of such other organisms as man. Man, however, has the remarkable ability to extend his sensory and intellectual capabilities artificially, through the use of instrumentation.

THE ORIGIN OF LIFE

Hypotheses of origins. Perhaps the most fundamental and at the same time the least understood biological problem is the origin of life. It is central to many scientific and philosophical problems and to any consideration of extraterrestrial life. Most of the hypotheses of the origin of life will fall into one of four categories:

1. The origin of life is a result of a supernatural event; that is, one permanently beyond the descriptive powers of physics and chemistry.

2. Life—particularly simple forms—spontaneously and readily arises from nonliving matter in short periods of time, today as in the past.

3. Life is coeternal with matter and has no beginning; life arrived on the Earth at the time of the origin of the earth or shortly thereafter.

4. Life arose on the early Earth by a series of progressive chemical reactions. Such reactions may have been likely or may have required one or more highly improbable chemical events.

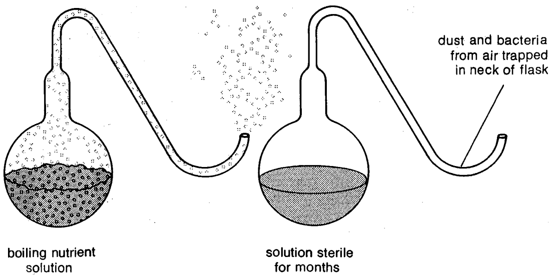
The early theological view

Hypothesis 1, the traditional contention of theology and some philosophy, is in its most general form not inconsistent with contemporary scientific knowledge, although this knowledge is inconsistent with a literal interpretation of the biblical accounts given in chapters 1 and 2 of Genesis and in other religious writings. Hypothesis 2 (not of course inconsistent with 1) was the prevailing opinion for centuries. A typical 17th-century view follows:

> [May one] doubt whether, in cheese and timber, worms are generated, or, if beetles and wasps, in cowdung, or if butterflies, locusts, shellfish, snails, eels, and such life be procreated of putrefied matter, which is to receive the form of that creature to which it is by formative power disposed[?] To question this is to question reason, sense, and experience. If he doubts this, let him go to Egypt, and there he will find the fields swarming with mice begot of the mud of the Nylus [Nile], to the great calamity of the inhabitants.

It was only in the Renaissance, with its burgeoning interest in anatomy, that such transformations were realized to be impossible. A British physiologist, William Harvey, during the mid-17th century, in the course of his studies on the reproduction and development of the king's deer, made the basic discovery that every animal comes from an egg. An Italian biologist, Francesco Redi, in the latter part of the 17th century, established that the maggots in meat came from flies' eggs, deposited on the meat. And an Italian priest, Lazzaro Spallanzani, in the 18th century, showed that spermatozoa were necessary for the reproduction of mammals. But the idea of spontaneous generation died hard. Even though it was proved that the larger animals always came from eggs, there was still hope for the smaller ones, the microorganisms. It seemed obvious that, because of their ubiquity, these microscopic creatures must be generated continually from inorganic matter.

Meat could be kept from going maggoty by covering it with a flyproof net, but grape juice could not be kept from fermenting by putting over it any netting whatever. This was the subject of a great controversy between the famous French bacteriologists Louis Pasteur and F.A. Pouchet in the 1850s, in which Pasteur triumphantly showed that even the minutest creatures came from germs floating in the air, but that they could be guarded against by suitable filtration. Actually, Pouchet was arguing that life must somehow arise from nonliving matter; if not, how had life come about in the first place?


dust and bacteria from air trapped in neck of flask
boiling nutrient solution
solution sterile for months

Pasteur's swan-necked flask experiment (see text).

Life from outer space

Toward the end of the 19th century Hypothesis 3 gained currency, particularly with the suggestion by a Swedish chemist, S.A. Arrhenius, that life on Earth arose from panspermia, microorganisms or spores wafted through space by radiation pressure from planet to planet or solar system to solar system. Such an idea of course avoids rather than solves the problem of the origin of life. In addition, it is extremely unlikely that any microorganism could be transported by radiation pressure to the Earth over interstellar distances without being killed by the combined effects of cold, vacuum, and radiation.

Pasteur's work discouraged many scientists from discussing the origin of life at all. Moreover they were anxious not to offend religious feeling by probing too deeply into the subject. Although Darwin would not commit himself on the origin of life, others subscribed to Hypothesis 4 more resolutely, notably the famous British biologist T.H. Huxley in his *Protoplasm, the Physical Basis of Life* (1869), and the British physicist John Tyndall in his "Belfast Address" of 1874. Although Huxley and Tyndall asserted that life could be generated from inorganic chemicals, they had extremely vague ideas about how this might be accomplished. The very phrase "organic molecule" implies that there exists a special class of chemicals uniquely of biological origin, despite the fact that organic molecules have been routinely produced from inorganic chemicals since 1828. In the following discussion the word organic carries no imputation of biological origin. In fact the problem largely reduces to finding an abiological source of appropriate organic molecules.

The primitive atmosphere. Darwin's attitude was: "It is mere rubbish thinking at present of the origin of life; one might as well think of the origin of matter." The two problems are, in fact, curiously connected, and modern scientists are thinking about the origin of matter. There is convincing evidence that thermonuclear reactions and subsequent explosions in the interiors of stars generate all the chemical elements more massive than hydrogen and helium and then distribute them into the interstellar medium from which subsequent generations of stars and planets form. Because of the commonality of these thermonuclear processes, and because some thermonuclear reactions are more probable than others, there exists a cosmic distribution of the major elements, so far as is

Table 1: Relative Abundances of the Elements (percent)

atom	universe	life (terrestrial vegetation)	earth (crust)
Hydrogen	87	16	3
Helium	12	0*	0
Carbon	0.03	21	0.1
Nitrogen	0.008	3	0.0001
Oxygen	0.06	59	49
Neon	0.02	0	0
Sodium	0.0001	0.01	0.7
Magnesium	0.0003	0.04	8
Aluminum	0.0002	0.001	2
Silicon	0.003	0.1	14
Sulfur	0.002	0.02	0.7
Phosphorus	0.00003	0.03	0.07
Potassium	0.000007	0.1	0.1
Argon	0.0004	0	0
Calcium	0.0001	0.1	2
Iron	0.002	0.005	18

*0% here stands for any quantity less than 10^{-6}%.

known, throughout the universe. Table 1 compares, for some atoms of interest, the relative numerical abundances in the universe as a whole, on the Earth, and in living organisms. There is of course some variation in composition from star to star, from place to place on the Earth, and from organism to organism, but such comparisons are nevertheless very instructive. The composition of life is intermediate between the average composition of the universe and the average composition of the Earth. Ninety-nine percent both of the universe and of life is made of the six atoms, hydrogen (H), helium (He), carbon (C), nitrogen (N), oxygen (O), and neon (Ne). Can it be that life on Earth arose when the chemical composition of the Earth was much closer to the average cosmic composition, and that some subsequent events have changed the gross chemical composition of the Earth?

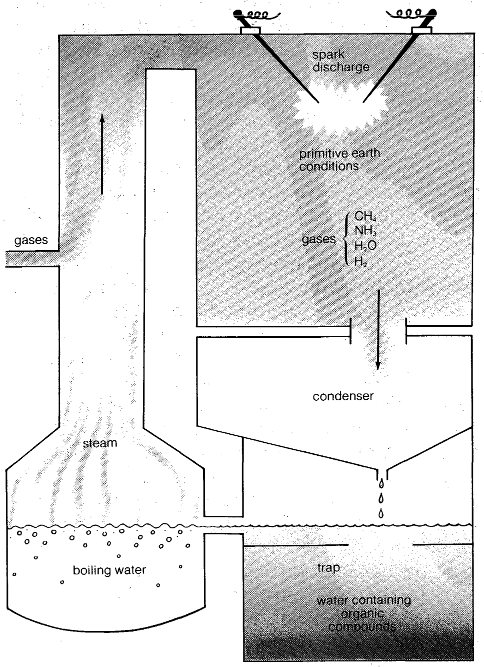
Composition of life

The Jovian planets (Jupiter, Saturn, Uranus, and Neptune) are much closer to cosmic composition than is the Earth. They are largely gaseous, with atmospheres composed principally of hydrogen and helium. Methane (CH_4) and ammonia (NH_3) have been detected in smaller quantities, and neon and water are suspected. This circumstance very strongly suggests that the Jovian planets were formed out of material of typical cosmic composition. They have very large masses, and because they are so far from the sun their upper atmospheres are very cold. Therefore it is impossible for atoms in the upper atmospheres of the Jovian planets to escape from their gravitational fields; escape was probably very difficult even during planetary formation. The Earth and the other planets of the inner solar system, however, are much less massive and most have hotter upper atmospheres. It is possible for hydrogen and helium to escape from the Earth today, and it may well have been possible for much heavier gases to have escaped during the formation of the Earth. It is reasonable to expect that in the very early history of the Earth a much larger abundance of hydrogen prevailed, which has subsequently been lost to space. Thus the atoms carbon, nitrogen, and oxygen were present on the primitive Earth, not as CO_2 (carbon dioxide), N_2, and O_2 as they are today but rather in the form of their fully saturated hydrides, CH_4 (methane), NH_3 (ammonia), and H_2O. In the geological record, the presence of such reduced minerals as uraninite (UO_2) and pyrite (FeS_2) in sediments formed several billions of years ago implies that conditions then were considerably less oxidizing than they are today.

In the 1920s J.B.S. Haldane in Britain and A.I. Oparin in the Soviet Union recognized that the abiological production of organic molecules in the present oxidizing atmosphere of the Earth is highly unlikely; but that, if the Earth once had more reducing (in this context, hydrogen-rich) conditions, the possible abiogenic production of organic molecules would have been much more likely. If large numbers of organic molecules were somehow synthesized on the primitive Earth, there would not necessarily be much trace of them today. In the present oxygen atmosphere, largely produced by green-plant photosynthesis, such molecules would tend, over geological time, to be oxidized to carbon dioxide, nitrogen, and water. In addition, as Darwin recognized, the first microorganisms would consume prebiological organic matter produced prior to the origin of life.

Production of simple organic molecules. The first deliberate experimental simulation of these primitive conditions was carried out in 1953 by a U.S. graduate student, S.L. Miller, under the guidance of the eminent chemist H.C. Urey. A mixture of methane, ammonia, water vapour, and hydrogen was circulated through a liquid water solution and continuously sparked by a corona discharge elsewhere in the apparatus. The discharge may be thought to represent lightning flashes on the early Earth. After several days of exposure to sparking, the solution changed colour. Subsequent analysis indicated that several amino and hydroxy acids, intimately involved in contemporary life, had been produced by this simple procedure. The experiment is in fact so elementary, and the amino acids can so readily be detected by paper chromatography, that the experiment has been repeated many times by high school students. Subsequent experiments have substituted ultraviolet light or heat as the energy source or have altered the initial abundances of gases. In all such experiments amino acids have been formed in large yield. On the early Earth there was much more energy available in ultraviolet light than in lightning discharges. At long ultraviolet wavelengths, in which methane, ammonia, water, and hydrogen are all transparent, but in which the bulk of the solar ultraviolet energy lies, the gas hydrogen sulfide (H_2S) is a likely ultraviolet absorber.

Synthesis of amino acids



Miller–Urey spark-discharge apparatus.

Following such reasoning, a U.S. astrophysicist, Carl Sagan, and his colleagues made amino acids by long wavelength ultraviolet irradiation of a mixture of methane, ammonia, water, and H_2S. The amino acid syntheses, at least in many cases, involve hydrogen cyanide and aldehydes (*e.g.,* formaldehyde) as gaseous intermediaries formed from the initial gases. It is quite remarkable that amino acids, particularly biologically abundant amino acids, can be made so readily under simulated primitive conditions. When laboratory conditions become oxidizing, however, no amino acids are formed, suggesting that reducing conditions were necessary for prebiological organic synthesis.

Under alkaline conditions, and in the presence of inorganic catalysts, formaldehyde spontaneously reacts to form a variety of sugars, including the five-carbon sugars fundamental to the formation of nucleic acids and such six-carbon sugars as glucose and fructose, which are extremely common metabolites and structural building blocks in contemporary organisms. Furthermore, the nucleotide bases as well as porphyrins have been produced in the laboratory under simulated primitive Earth conditions by several investigators. While there is still debate on the generality of the experimental synthetic pathways and on the stability of the molecules produced, most if not all of the essential building blocks of proteins, carbohydrates, and nucleic acids can be readily produced under quite general primitive reducing conditions, plus probably ATP as well.

Production of polymers. The construction of polymers, long-chain molecules made of repeating units of these essential building blocks, however, is a much more difficult experimental problem. Polymerization reactions are generally dehydrations, in which a molecule of water is lost in the formation of a two-unit polymer. Dehydrating agents must be used to initiate polymerization. The polymerization of amino acids to form long protein-like molecules was accomplished through dry heating by a U.S. investigator, S.W. Fox. The polyamino acids that are formed are not random polymers and have some distinct catalytic activities. The geophysical generality of dry heat-

Polymerization as dehydration reactions

ing and return to solution, however, has been questioned. Long polymers of amino acids can also be produced from hydrogen cyanide and anhydrous liquid ammonia. Some evidence exists that nucleotide bases and sugars can be combined in the presence of phosphates or cyanides under ultraviolet irradiation. Some condensing agents such as cyanamide are efficiently made under simulated primitive conditions. Despite the breakdown by water of molecular intermediates, condensing agents are often quite effective in inducing polymerization, and polymers of amino acids, sugars, and nucleotides have all been made this way.

A famous British scientist, J.D. Bernal, suggested that adsorption of molecular intermediates on clays or other minerals may have concentrated these intermediates. Such concentration could offset the tendency for water to break down polymers of biological significance. Of special interest is the possibility that such concentration matrices included phosphates, for this would help explain how phosphorus could have been incorporated preferentially into prebiological organic molecules at a time when biological concentration mechanisms did not yet exist. Mineral catalysis implies that organic synthesis could also occur in deep water where ultraviolet light had been filtered out.

Quite apart from concentration mechanisms, the primitive waters themselves may have been a not very dilute solution of organic molecules. If all the surface carbon on the Earth were present as organic molecules in the contemporary oceans, or if many known ultraviolet synthetic reactions producing organic molecules were permitted to continue for a billion years with products dissolved in the oceans, a 1 percent solution of organic molecules would result. For similar reasons, Haldane suggested that the origin of life occurred in a "hot dilute soup." In addition, concentration mechanisms do exist, such as evaporation or freezing of pools or adsorption on interfaces or the generation of colloidal enclosures called coacervates.

The origin of the code. It has been shown that all the essential building blocks for life and their polymers may have been produced in some fair concentration on the primitive Earth. This possibility is certainly relevant to the origin of life, but it is not the same thing as the origin of life. By the genetic definition of life discussed above in *Definitions of life,* a self-replicating, mutable molecular system, capable of interacting with the environment, is required. In contemporary cells the nucleic acids are the sites of self-replication and mutation. Laboratory experiments have already shown that polynucleotides can be produced from nucleotide phosphates in the presence of a specific enzyme of biological origin and a pre-existing "primer" nucleic acid molecule. If the primer molecule is absent, polynucleotides are still formed, but they of course contain no genetic information. Once such a polynucleotide spontaneously forms, it then acts as primer for subsequent syntheses.

Production of polynucleotides

Imagine a primitive ocean filled with nucleotides and their phosphates and appropriate mineral surfaces serving as catalysts. Even in the absence of the appropriate enzyme it seems likely, although not yet proved, that spontaneous assembly of nucleotide phosphates into polynucleotides occurred. Once the first such polynucleotide was produced, it may have served as a template for its own reproduction, still of course in the absence of enzymes. As time went on there were bound to be errors in replication. These would be inherited. A self-replicating and mutable molecular system of polynucleotides, eventually leading to a diverse population of such molecules, may have arisen in this way. Alternatively, the primitive hereditary material may have involved some other molecule altogether, but no concrete suggestion for such a molecule has ever been proposed.

In any case, a population of replicating polynucleotides cannot quite be considered alive because it does not significantly influence its environment. Eventually, all the nucleotides in the ocean would have been tied in polynucleotides and the entire synthetic process would then have ground to a halt. So far as is known, polynucleotides have no an catalytic properties, and proteins have no reproductive properties. It is only the partnership of the two molecules that makes contemporary life on Earth possible. Accordingly, a critical and unsolved problem in the origin of life is the first functional relation between these two molecules, or, equivalently, the origin of the genetic code. The molecular apparatus ancillary to the operation of the code—the activating enzymes, adapter RNAs, messenger RNAs, ribosomes, and so on—are themselves each the product of a long evolutionary history and are produced according to instructions contained within the code. At the time of the origin of the code such an elaborate molecular apparatus was of course absent.

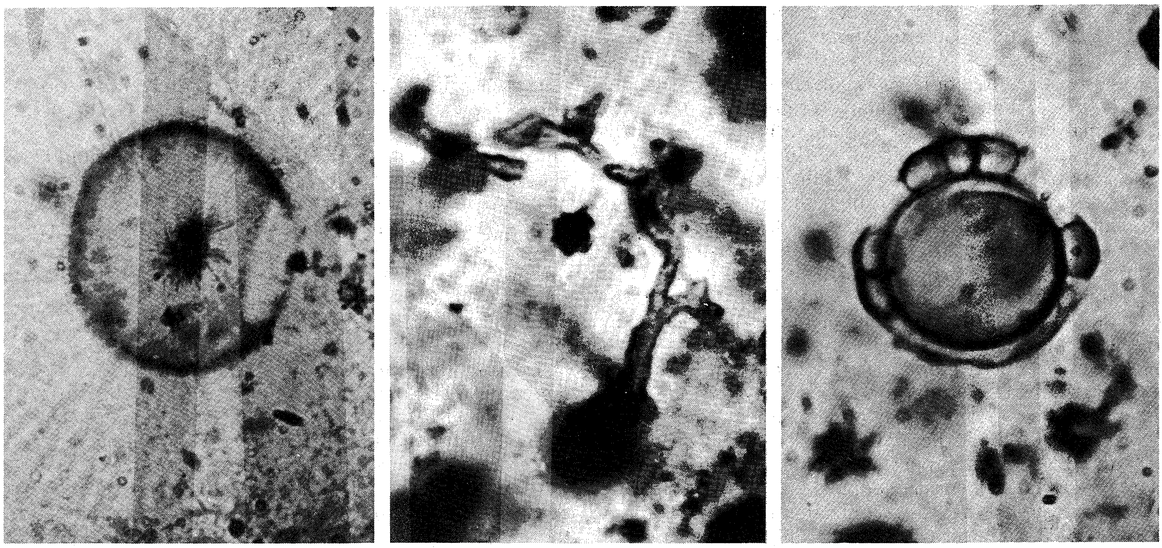
Catalysis and reproduction as necessary for life

It has been proposed that a weak but selective chemical bonding does exist, even in the absence of any of this apparatus, between amino acids and nucleotides. There need not be a very great selectivity; a given nucleotide sequence might in primitive times have coded for many different amino acids or, conversely, the same amino acid may have been coded for by several different nucleotide sequences. All that is required is that a particular linear sequence of nucleotides must code for some nonrandom sequence of amino acids. The active sites largely responsible for the catalytic activity of contemporary enzymes are generally only five or six amino acids long; the remainder of the enzyme is devoted to more sophisticated functions, such as arranging for the enzyme to be turned on and off by the machinery of the cell. With, say, 20 different varieties of amino acids available in the primitive environment, the chance of any given active site being produced by a random sequence of nucleotides is one in 20^5, or one in about 3,000,000. But 3,000,000 combinations to form units five amino acids long is not a very large number for the chemistry and time periods in question. To conclude this speculation, then, if polynucleotides were initially capable of crude, nonenzymatic replication, and if a crude primitive genetic code existed, then any one of a very large number of catalytic properties was available to some self-replicating polynucleotides on the primitive Earth. This situation is all that would be necessary for the origin of life; those polynucleotides that could code for a primitive protein having catalytic properties furthering the replication of the polynucleotide would preferentially replicate. Other polynucleotides coding for less effective proteins would have replicated more slowly. The foregoing is one of several possibilities for the origin of the first living systems. Many separate and rather diverse instances of the origin of life may have occurred on the primitive Earth, but competition eventually eliminated all but one line. Every organism on Earth today would be a descendant of that line.

Optical activity of biological molecules

The earliest living systems. One curious feature of biological organic molecules is their optical activity: they rotate the plane of a beam of plane-polarized light. Organic molecules produced abiologically do not show optical activity. Molecules made of the same units can be put together in complementary ways like a left- and right-handed glove. The same building blocks can be used to produce molecules that are three-dimensional mirror images of each other. This asymmetry is responsible for optical activity. At the time of the origin of life, organic molecules, corresponding both to left- and right-handed forms, were produced. The laboratory simulation experiments always produce both types. But the first living systems could have been made only of one type, for the same reason that carpenters do not use random mixtures of screws with left- and right-handed threads. Whether left- or right-handed activity was adopted was probably purely a matter of chance, but once a particular asymmetry was established in the first living systems, it maintained itself. This belief implies that optical activity should be a feature of life on any planet, and also that the chances should be equal of finding a given terrestrial organic molecule or its mirror image molecule in extraterrestrial life forms.

The first living systems probably resided in a molecular garden of Eden, where all the building blocks that contemporary organisms must work hard at synthesizing were available free. Under such conditions the numbers of organisms must have increased very rapidly. But such increases cannot go on indefinitely. In time the supply of some molecular building block must have become short. Those primitive organisms that had the ability to synthesize the scarce building block, say A, from a more



Photomicrographs of 2,000,000,000-year-old organisms from the geological stratum called the Gunflint Chert. These organisms are about 10 microns across. Organisms more than 1,000,000,000 years older are known.

By courtesy of Elso S. Barghoorn

abundant one, say B, clearly had a competitive advantage over those organisms that could not perform such a synthesis. In time, however, the secondary source of supply, B, would have also become depleted and those organisms that could produce it from a third building block, C, would have preferentially replicated. A U.S. biochemist, N.H. Horowitz, proposed that in this way the enzymatic reaction chains of contemporary organisms—each step catalyzed by a particular enzyme—originally evolved.

Even the evolution of enzymatic reaction chains may have occurred in free nucleic acids before the origin of the cell. The cell may have arisen in response to the need for maintaining a high concentration of scarce building blocks or enzymes, or as protection against the gradually increasing abundance of oxygen on the primitive Earth. Oxygen is a well-known poison to many biological processes, and in contemporary higher organisms the mitochondria that handle molecular oxygen are kept in the cytoplasm, far from contact with the nuclear material. Even today processes are known whereby polyamino acids form small spherical objects, microns to tens of microns across, with some of the properties of cells. These objects, called proteinoid microspheres by Fox, are certainly not cells, but they may indicate processes by which the ancestors of cells arose. Procaryotic cells almost certainly preceded eucaryotic cells, and the evolution of so extremely complex an apparatus as the mitotic spindle (which ensures equal segregation of replicated chromosomes) must have taken very long periods of time to evolve. The development of mitochondria and chloroplasts (each of which contains its own DNA) in the eucaryotic cell may have been the result of a symbiosis, a cooperative arrangement entered into at first tentatively by originally free-living cells.

Proteinoid microspheres

As the competition for building blocks increased among early life forms, and also perhaps as the abiological production of organic molecules dwindled because of the increasing oxygen abundance, the strictly heterotrophic way of life became more and more costly. The utilization of porphyrins, which are also made abiologically, by primitive photoautotrophs would have had great selective advantage. Many of the intermediates and enzymes in photosynthesis and in the anaerobic breakdown of carbon compounds are similar, but there is no generally accepted view of the origin of the photosynthetic process. Photosynthesis in procaryotes is more primitive than in such eucaryotes as green plants. In bacteria, water is not the ultimate source of hydrogen atoms for reducing carbon dioxide, and therefore oxygen is not produced. In addition, when a chlorophyll-containing cell is exposed both to light and to oxygen, it is killed unless it also contains an accessory carotenoid pigment. Thus green-plant photosynthesis had to wait until the appearance of carotenoids while bacterial photosynthesis, which does not produce oxygen, could function without carotenoids.

The antiquity of life. Among the oldest known fossils are those found in the Fig Tree chert from the Transvaal, dated at 3,100,000,000 years old. These organisms have been identified as bacteria and blue-green algae. It is very reasonable that the oldest fossils should be procaryotes rather than eucaryotes. Even procaryotes, however, are exceedingly complicated organisms and very highly evolved. Since the Earth is about 4,500,000,000 years old, this suggests that the origin of life must have occurred within a few hundred million years of that time.

By performing chemical analyses on the oldest sediments, it is possible to say something about the sorts of organic molecules produced, either biologically or abiologically, in primitive times. Thus, amino acids and porphyrins have been identified in the oldest sediments, as have pristane and phytane, typical breakdown products of chlorophyll. There are several indications that these organic molecules, dating from 2,000,000,000 to more than 3,000,000,000 years ago, are of biological origin. For one thing their long-chain hydrocarbons show a preference for a straight chain geometry, whereas known abiological processes tend to produce a much larger proportion of branched chain and cyclic hydrocarbon molecular geometries than have been found in these sediments. Abiological processes tend to produce equal amounts of long-chain carbon compounds with odd and with even numbers of carbon atoms. But the oldest sediments show a distinct preference for odd numbers of carbon atoms per molecule, as do products of undoubted biological origin. Finally, a C^{12} enrichment, for which no abiological process seems able to account, has been discovered in the oldest sediments, evidence that suggests that plantlike life, which concentrates the carbon isotope C^{12} preferentially to C^{13}, was present very early. These departures from thermodynamic equilibrium are often considered to be compelling signs of biological activity. Such evidence again points to the great antiquity of life on Earth.

The fossil record, in any complete sense, goes back only about 600,000,000 years. In the layers of sedimentary rock known by geological methods and by radioactive dating to be that old, most of the major groups of invertebrates appear for the first time. All these organisms appear adapted to life in the water, and there is no sign yet of organisms adapted to the land. For this reason, and because of a rough similarity between the salt contents of blood and of seawater, it is believed that early forms of life developed in oceans or pools. With no evidence for widespread oxygen-producing photosynthesis before this time, and for cosmic

Oceanic origin of life

abundance reasons described above, the oxygen content of the Earth's atmosphere in Precambrian times was very likely less than today. Accordingly, in Precambrian times, solar ultraviolet radiation, especially near the wavelength of 2,600 Å, which is particularly destructive to nucleic acids, may have penetrated to the surface of the Earth, rather than being totally absorbed in the upper atmosphere by ozone as it is today. In the absence of ozone, the ultraviolet solar flux is so high that a lethal dose for most organisms would be delivered in less than an hour. Unless extraordinary defense mechanisms existed in Precambrian times, life near the Earth's surface would have been impossible. Sagan suggested that life at this time was generally restricted to some tens of metres and deeper in the oceans, at which depths all the ultraviolet light would have been absorbed, although visible light would still filter through. As the amount of atmospheric oxygen and ozone increased, due both to plant photosynthesis and to the photodissociation of water vapour and the escape to space of hydrogen from the upper atmosphere, life increasingly close to the Earth's surface would have been possible. It has been suggested that the colonization of the land, about 425,000,000 years ago, was possible only because enough ozone was then produced to shield the surface from ultraviolet light for the first time.

Life then had insinuated itself between the sun and the Earth. It diverted solar energy to its own uses and contrived more and more ways of exploiting more and more environments. Some experiments were faulty and the lines became extinct; others were more successful and the lines filled the Earth. Evolution through natural selection directed the proliferation of a growing array of life forms throughout the biosphere (see EVOLUTION; THE THEORY OF).

EXTRATERRESTRIAL LIFE

It is not known what aspects of living systems are necessary in the sense that living systems everywhere must have them; it is not known what aspects of living systems are contingent in the sense that they are the result of evolutionary accident, so that somewhere else a different sequence of events might have led to different characteristics. In this respect the possession of even a single example of extraterrestrial life, no matter how seemingly elementary in form or substance, would represent a fundamental revolution in biology. It is not known whether there is a vast array of biological themes and counterpoints in the universe, whether there are places that have fugues, compared with which our one tune is a bit thin and reedy. Or it may be that our tune is the only tune around. Accordingly the prospects for life on other planets must be considered in any general discussion of life.

The chemistry of extraterrestrial life. What are the methods and prospects for a search for life beyond the Earth? Each of the definitions of life described in *Definitions of life* (see above) implies a method of searching for life. Particular physiological functions, particular metabolic activities, such specific molecules as proteins and nucleic acids, self-replication and mutation, processes not in closed-system thermodynamic equilibrium—all these might be sought. All the search methods significantly depend upon chemistry.

Life on Earth is structurally based on carbon and utilizes water as an interaction medium. Hydrogen and nitrogen have significant accessory structural roles; phosphorus is important for energy storage and transport, sulfur for three-dimensional configuration of protein molecules, and so on. But must these particular atoms be the atoms of life everywhere, or might there be a wide range of atomic possibilities in extraterrestrial organisms? What are the general physical constraints on extraterrestrial life?

Prerequisites for life

In approaching these questions several criteria can be used. The major atoms should tend to have a high cosmic abundance. A structural molecule for making an organism at the temperature of the planet in question should not be extremely stable, because then no chemical reactions would be possible; but it should not be extremely unstable, because then the organism would fall to pieces. There should be some medium for molecular interaction. Solids are not appropriate because the diffusion times are very long. Such a medium is most likely a liquid (but could possibly be a very dense gas) that is stable in a number of respects. It should have a large temperature range (for a liquid, the temperature difference between freezing point and boiling point should be large). The liquid should be difficult to vaporize and to freeze; in fact, it should be very difficult to change its temperature at all. In addition it should be an excellent solvent. There should also be some gas on the planet in question that could be used in various biologically mediated cycles, as CO_2 is in the carbon cycle on Earth.

The planet, therefore, should have an atmosphere and some near-surface liquid, although not necessarily an ocean. If the intensity of ultraviolet light or charged particles from the sun is intense at the planetary surface, there must be some place, perhaps below the surface, that is shielded from this radiation but that nevertheless permits useful chemical reactions to occur. Since after a certain period of evolution, lives of unabashed heterotrophy lead to malnutrition and death, autotrophs must exist. Chemoautotrophs are, of course, a possibility but the inorganic reactions that they drive usually require a great deal of energy; at some stage in the cycle, this energy must probably be provided by sunlight. Photoautotrophs, therefore, seem required. Organisms that live very far subsurface will be in the dark, making photoautotrophy impossible. Organisms that live slightly subsurface, however, may avoid ultraviolet and charged particle radiation and at the same time acquire sufficient amounts of visible light for photosynthesis.

Thermodynamically, photosynthesis is possible because the plant and the radiation it receives are not in thermodynamic equilibrium; for example, on the Earth a green plant may have a temperature of about 300 K while the sun has a temperature of about 6,000 K. (K = Kelvin temperature scale, in which 0 K is absolute zero; 273 K, the freezing point of water; and 373 K, the boiling point of water at one atmosphere pressure.) Photosynthetic processes are possible in this case because energy is transported from a hotter to a cooler object. Were the source of radiation at the same (or at a colder) temperature as the plant, however, photosynthesis would be impossible. For this reason the idea of a subterranean plant photosynthesizing with the thermal infrared radiation emitted by its surroundings is untenable, as is the idea that a cold star, with a surface temperature similar to that of the Earth, would harbour photosynthetic organisms.

Chemical requirements

It is possible to approach some of the foregoing chemical requirements and see just which atoms are implied. When atoms enter into chemical combination, the energy necessary to separate them is called the bond energy, a measure of how tightly the two atoms are bound to each other. Table 2 gives the bond energies of a number of chemical bonds, mostly involving abundant atoms. The energies are in electron volts (eV; $1 \text{ eV} = 1.6 \times 10^{-12}$ ergs). The symbols are as follows: H, hydrogen; C, carbon; N, nitrogen; O, oxygen; S, sulfur; F, fluorine; Si, silicon; Bi, bismuth (very underabundant, biologically uninteresting, and present only as an illustration of the relatively weak chemical bonds in some metals). Bond energies generally vary between 10 eV and about 0.03 eV; double and triple

Table 2: Energies of Representative Chemical Bonds

(Hydrogen bonds 0.08–0.45 eV; van der Waals bonds 0.04 eV)

bond	energy (eV)	bond	energy (eV)
N≡N	9.8	Si—O	3.8
C≡N	9.4	C—O	3.7
C≡C	8.4	C—C	3.6
C—O	7.4	S—H	3.5
C=C	6.4	Si—H	3.1
H—F	5.4	C—N	3.0
O—H	4.8	Si—Si	1.8
N=N	4.4	N—N	1.7
C—H	4.3	Bi—Bi	1.1
N—H	4.1	O_2N—NO_2	0.57

bonds where two or three electrons are shared between two atoms tend to be more energetic than single bonds, single bonds more energetic than hydrogen bonds where a hydrogen atom is shared between two other atoms, and hydrogen bonds more energetic than the very weak (van der Waals) forces that arise from the attraction of the electrons of one atom for the nucleus of another. At room temperature, atoms, free or bound, move with an average kinetic energy corresponding to about 0.02 eV. Some of the atoms have greater energies, some lesser. At any temperature a few will have energies greater than any given bond energy; hence bonds occasionally will break. The higher the temperature, the more atoms there are moving with sufficient energy to spontaneously break a given bond.

Suppose it is decided arbitrarily (although the decision will not critically affect the conclusions) that for life to exist at any time the fraction of bonds broken by random thermal motions must be no larger than 0.0001 percent. It then turns out that any hypothetical life where the structural bonds are based upon van der Waals forces can only exist where the temperature is below 40 K, for hydrogen bonds below about 400 K, for bonds of 2 eV below 2,000 K, and for bonds of 5 eV below 5,000 K. Now, 2,000 to 5,000 K are typical surface temperatures of stars; 400 K is somewhat above the highest surface temperature found on Earth; and 40 K is about the cloud-top temperature of distant Neptune. Thus, over the entire range of temperatures, from cold stars to cold planets, there seem to exist chemical bonds of appropriate structural stability for life, and it would appear premature to exclude the possibility of life on any planet on grounds of temperature.

Life on Earth lies within a rather narrow range of temperature. Above the normal boiling point of water, much loss of configurational structure or three-dimensional geometry occurs. At these temperatures proteins become denatured, in part because above the boiling point of water the hydrogen bonding and van der Waals forces between water and the protein disappear. Also, similar bonds within the protein molecule tend to break down. Proteins then change their shapes, their ability to participate in lock-and-key enzymatic reactions is gravely compromised, and the organism dies. Similar structural changes, some of them connected with the stacking forces between adjacent nucleotide bases, occur in the heating of nucleic acids. But it is significant that these changes are not fragmentations of the relevant molecules but rather changes in the ways they fold. There appears to be no reason that configurational bonds should not have been evolved that are stable at higher temperatures than terrestrial organisms experience. On planets hotter than the Earth there seems to be no reason that slightly more stable configurational forces should not be operative in the local biochemistry.

Molecular factors

While the bonds that characterize life on Earth are too weak at high temperatures, they are too strong at low temperatures, tending to slow down the rates of chemical reactions generally. There are less stable bonds (*e.g.,* hydrogen bonds, silicon-silicon bonds, and nitrogen-nitrogen bonds), however, that might play structural roles at significantly lower temperatures. At higher temperatures, multiple bonds (*e.g.,* in aromatic, or ring-shaped, hydrocarbons) might be utilized for life. There clearly is a rich variety of little-studied chemical reactions that proceed at reasonable rates either at much lower or at much higher temperatures than those on Earth.

Except for bismuth and fluorine, all the atoms in Table 2 have relatively high cosmic abundances. At terrestrial temperatures, carbon is the unique atom for biological structure. Not only does it have high abundance but it forms a staggering variety of compounds of great stability, it lends itself to compounds that are configured by weaker bonds, and it enters into multiple bonds. These double- and triple-bonded molecules, among other useful properties, absorb long-wavelength ultraviolet light, a process leading to the synthesis of a variety of more complex molecules. A photon of ultraviolet light at a wavelength of 2,000 Å has an energy of 6.2 eV, capable of breaking many bonds, and permitting more complex reactions among the resulting molecular fragments. Photons of blue light have energies of about 3 eV, and of red light about 2 eV.

Silicon compounds do not form double bonds. Silicon-oxygen bonds are slightly more stable than carbon-carbon bonds, but they tend to produce molecules like the silicates, which are crystals of the same unit repeated over and over again, rather than molecules with aperiodic side chains with potential information content. On low-temperature planets, silicon-silicon bonds are more promising than carbon bonds in terms of reaction times, but they do not form double bonds and the carbon abundance is likely to be greater. Nevertheless, silicon compounds may be of limited biological importance both on high-temperature and low-temperature worlds.

Hydrogen bonding confers on liquids the stability properties necessary for life. There seem to be very few reasonable candidates for liquid interaction media. By all odds water is the most suitable. The other candidates, all to some extent hydrogen bonded, are ammonia, hydrogen fluoride, hydrogen cyanide, and mixtures of liquid hydrocarbons. Hydrogen fluoride can be excluded because it is too scarce cosmically. The hydrocarbons are not good solvents of salts, but life elsewhere may not be based on the same acid-base chemistry as life on Earth. The liquid range of water is larger than commonly thought, ranging from about 210 K in saturated salt solutions to 647 K at enormous atmospheric pressures. Water is the biological liquid medium of choice above 200 K, particularly in view of its extremely high cosmic abundance. At lower temperatures ammonia or hydrogen cyanide could serve as a liquid medium.

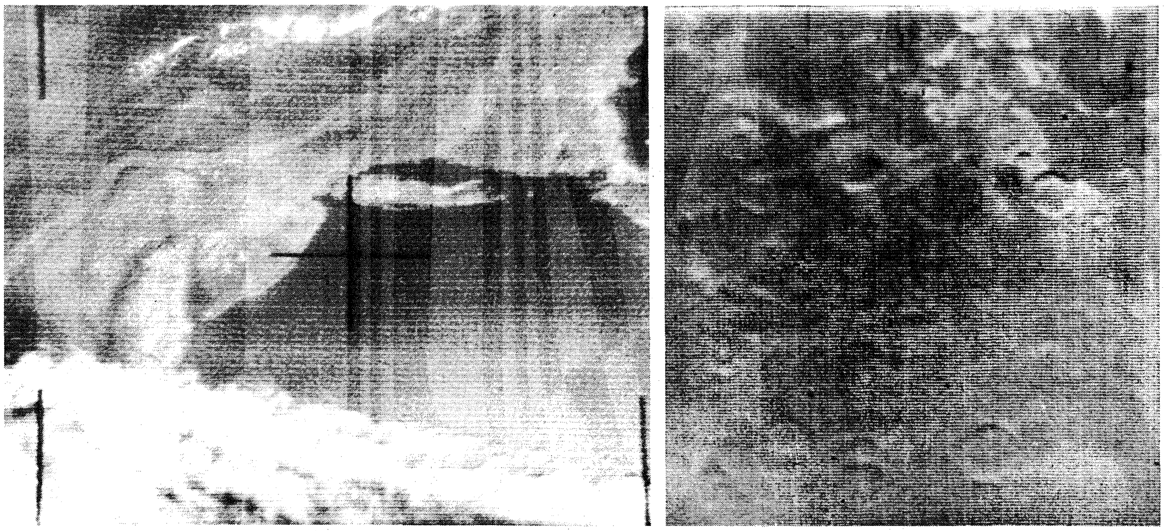
There are functional roles for specific atoms in biology, but except for considerations of structure and a liquid interaction medium they do not seem fundamental. For example, the energy-rich phosphate bonds in ATP are in fact of relatively low energy; they are about as energetic as the hydrogen bonds (see Table 2). The cell must store up large numbers of these bonds to drive a molecular degradation or synthesis. On high-temperature worlds the energy currency may be much more energetic per bond, and on low-temperature worlds much less energetic per bond.

It may be concluded that, in our present state of ignorance, it is premature to exclude life on grounds of temperature on any other planet, particularly when account is taken of the temperature heterogeneity of the other planets. But life does require an interaction medium, an atmosphere, and some protection from ultraviolet light and from charged particles of solar origin.

The conclusion that for the Earth, carbon-based aqueous life is the most appropriate may be slightly suspect, since terrestrial life is manifestly carbon-based and aqueous. In 1913 a U.S. biochemist, L.J. Henderson, published *The Fitness of the Environment* in which the biological advantages of carbon and water were stressed for the first time in terms of comparative chemistry. He was struck by the fact that those very atoms that are needed are just those atoms that are around; it remains a remarkable fact that atoms most useful for life do have very high cosmic abundances.

Exobiology

The search for extraterrestrial life. Exobiology, a term coined by a U.S. biologist, J. Lederberg, for the study of extraterrestrial life, has been called a science without a subject matter. It is certainly true that, as yet, no strong evidence for life beyond the Earth has been adduced. Exobiology, however, has deep significance even if extraterrestrial life is never found. The mere design of exobiological experiments forces man to examine critically the generality of his assumptions about life on Earth. In addition, a lifeless neighbouring planet presents a very interesting quandary: How is it that life has originated and evolved on Earth, but not on the planet in question? There is an entire spectrum of possibilities. A given planet may be lifeless and have no vestiges of primitive organic matter and no fossils of extinct life. It may be lifeless but may have either organic chemical or fossil relics. It may possess life of a simple sort or life of a quite complex biochemistry, physiology, and behaviour. It may possess intelligent life and a technical civilization. Establishment of any one of these five possibilities would be of fundamental biological importance.



(Left) Eastern seaboard of the United States photographed by a TIROS weather satellite; Cape Cod (right), Long Island (centre), and Delaware Bay (centre left) can be seen. (Right) Surface of Mars photographed by Mariner 4 spacecraft. Both photographs at one-kilometre resolution. No sign of life, intelligent or otherwise, can be discerned on either planet.

By courtesy of National Aeronautics and Space Administration

The difficulties and opportunities inherent in exobiological exploration, in determining which of these five possibilities applies to a given planet, is most clearly grasped by imagining the situation reversed, with man on some neighbouring planet, say Mars, examining the Earth for life with the full armoury of contemporary scientific instrumentation and knowledge. First a distinction must be made between remote and *in situ* testing. In remote testing light of any wavelength reflected from or emitted by the target planet can be examined, but with *in situ* studies samples of the planet must be acquired by visiting them or by sending instruments that land on the planet, perform experiments, and radio back their findings. Since biological exploration involves the detailed characterization of any life found, rather than its mere detection, *in situ* experiments are necessary.

Sensing methods

The bulk of the remote sensing methods are directed toward finding some thermodynamic disequilibrium on the planet. This may be a chemical disequilibrium, a mechanical disequilibrium, or a spectral disequilibrium. For example, it would be quite easy to determine spectroscopically from Mars that the Earth's atmosphere contains large amounts of molecular oxygen and about one part per million (10^6) of methane. It would also be possible to calculate that, at thermodynamic equilibrium, the abundance of methane should be less than one part in 10^{35}. This huge discrepancy implies the existence of some process continuously generating methane on the Earth so rapidly that methane increases to a very large steady-state abundance before it can be oxidized by oxygen. Now such a methane-production mechanism need not be biological. It is conceivable that relatively stable aromatic hydrocarbons were produced abiologically in the early history of the Earth and that their slow thermal degradation leads to a continuous loss of methane from the planetary subsurface. But this and similar nonbiological explanations of the observed disequilibrium are unlikely. From Mars this thermodynamic discrepancy would be considered not as proof of life on Earth but as a significant hint of life on Earth. In fact the methane abundance on the Earth is produced by bacteria that, in the course of the reduction of a more oxidized form of carbon, release methane. Some methane bacteria live in swamps (hence, the term marsh gas for methane), and others—a significant fraction—live in the intestinal tracts of cows and other ruminants. The methane abundance over India is probably larger than over most other areas of the world, and if an extraterrestrial observer knew how to interpret the methane disequilibrium accurately (which is unlikely) it would be possible for him to deduce cows on Earth by spectrochemical analysis. The existence of relatively large quantities of methane in the presence of an excess of oxygen would remain a tantalizing but enigmatic hint of life on Earth. Similarly, the large amount of oxygen might itself be a sign of life if one could reliably exclude the possibility that the photodissociation of water and the escape to space of hydrogen were the source of oxygen. Also such relatively complex reduced organic molecules as terpenes, a hydrocarbon given off by plants, might conceivably be detected spectroscopically, perhaps by a spectrometer in orbit about the Earth. Not only would the chemical disequilibrium of terpenes in an excess of oxygen be suggestive of life, but equally suggestive would be the fact that terpenes are much more abundant over forested areas than over deserts.

Photographic observation

Photographic observations of the daytime Earth from Mars would give equivocal results. Even with a resolution of 100 metres (that is, an ability to discriminate fine detail at high contrast only if its components are more than 100 metres apart), it would be extremely difficult to discern cities, canals, bridges, the Great Wall of China, highways, and other large-scale accoutrements of the Earth's technical civilization. In satellite photographs with 100-metres (one metre = 1.0936 yards) resolution only about one in a thousand random photographs of the Earth yields features even suggestive of life. As the ground resolution is progressively improved, it becomes increasingly easy to make out the regular geometrical patterns of cultivated fields, highways, airports, and so on. But these are only the products of a civilization recently developed on Earth, and even photographs of the Earth with a ground resolution of 10 metres, but taken 100,000 years ago, would still have shown no clear sign of life. The lights of the largest cities might be just marginally detectable from Mars at night. Seasonal changes in the colour or darkness of plants would be detectable from Mars, but such cycles might easily have nonbiological explanations.

To detect individual animals a ground resolution of a few metres is required, and even here a low sun and long shadows are generally necessary. This detection could be accomplished with a large telescope in Earth orbit. It would then be possible to determine, for example, that objects with the general shape of cows are frequent on the Earth. But suppose that members of the civilization examining the Earth thus remotely are not even approximately quadrupedal and do not immediately associate the shape of cows with life. They would nevertheless be able to deduce life. They would observe that certain locales on Earth have a quantity of raised lumps connected to the ground by four stilts. It would be possible to calculate that wind and water erosion would cause the lumps to topple to the ground in geologically short periods of time. Such stilted lumps are mechanically unstable; they are not in equilibrium; if pushed hard, they fall. Accordingly, there must be a process for generating stilted lumps on the Earth, and in short periods of time. It would be very difficult to avoid the implication that this generation process is biological.

Radio emission

A third detection technique arises upon scanning the radio spectrum of the Earth. Because of domestic television transmission, the high-frequency end of the AM broadcast band, and the radar defense networks of the United States and of the Soviet Union, the amount or energy put out by the Earth to space at certain radio frequencies is enormous. At some frequencies, if this radiation were interpreted as ordinary thermal emission, the temperature of the Earth would have to be hundreds of millions of degrees, according to an estimate made by a Soviet astrophysicist, I.S. Shklovskii. Moreover, it would be possible to determine that this radio "brightness temperature" of the Earth had been steadily increasing with time over the last several decades. Finally, it would be possible to analyze the frequency and time variation of these signals and deduce that they were not purely random noise.

Now imagine *in situ* studies by vehicles that enter the Earth's atmosphere and land at some predetermined locale. There are many places on the Earth (the ocean surface, the Gobi Desert, Antarctica) where large organisms are infrequent and a life-detection attempt based solely on television searches for large life forms would be a risky investment. On the other hand, if such an experiment were successful (the camera records a dolphin cavorting, a camel chewing its cud, a penguin waddling) it would provide quite convincing evidence of life.

Although the oceans, the Gobi Desert, and Antarctica are relatively devoid of large life forms, they are in many places replete with minute life forms. Therefore, microorganism detectors would be a good investment. A television camera coupled to a microscope (optical or electron) would be a promising life detector if the sample acquisition problem could be solved: the early Dutch microscopist Antonie van Leeuwenhoek had no difficulty at all in identifying as alive the little "animalcules" that he found in a drop of water, although nothing similar had previously been seen in human history.

In addition to morphological criteria for the detection of microorganisms, there are metabolic and chemical criteria. For example, a sample of terrestrial soil, or seawater, say, might be acquired and introduced into a chamber containing food the investigators guess the earthlings might find tasty. Such food might be an abundant product of prebiological organic synthetic experiments. It could then be determined whether any characteristic molecules, such as carbon dioxide or ethanol, are produced metabolically, or whether the medium containing food and terrestrial sample changes its acidity or becomes cloudy because of the growth of micro-organisms, or it might be investigated whether there is heat given off in the chamber containing sample and food. Alternatively, photosynthesis could be tested by measuring the fixation of some gas, say carbon dioxide, as a function of illumination provided artificially to the sample by the instrument. Along chemical lines a direct test of terrestrial soil or seawater for optical activity might be made. Organic molecules could certainly be searched for with a combined gas chromatograph and mass spectrometer, or by a remote analytic chemistry laboratory. The detection of any organic matter would of course be interesting and relevant, whether or not it was biological in origin. Such criteria as have been used in the analysis of Precambrian sediments (described in *The antiquity of life,* above) might be used to test for biological origin.

Ambiguities of tests for life

It is remarkable, however, that many of these tests are ambiguous. It would be possible, for example, for the Martian investigator to guess wrong about what terrestrial organisms eat and to make incorrect assumptions about their structural chemistry or their interaction medium. If forms of regular geometry that do not move were detected microscopically, there might be serious questions of biological versus mineralogical origin. Chemical criteria (such as the expectation that if odd-numbered carbon chains are more prominent than even-numbered carbon chains, then life is detected) might not be valid unless it was certain which processes actually occurred in the prebiological organic chemistry of the planet in question. In addition, there might be the galling problem of contamination. The Martians' spacecraft might carry living organisms from their own planet and report them as detected on the planet Earth. For this reason great care would have to be taken that spacecraft were rigorously sterilized.

In fact, many of these problems have already arisen in an analysis of a variety of meteorite called carbonaceous chondrites. These meteorites, which fall on Earth probably from the asteroid belt, contain about 1 percent organic matter by mass, far too much to be largely the result of terrestrial contamination. The most abundant organic molecules, however, are not clearly of biological origin, and some of the biologically more interesting molecules may be contaminants. Reports of optical activity have been contested and might alternatively be due to contamination. Geometrically interesting microscopic inclusions have been detected in these bodies. The most abundant inclusions, however, are probably mineralogical in origin, while the most highly structured and lifelike are very rare and, at least in some cases, are obviously due to contamination (in one case by ragweed pollen). Finally, claims have been made of the extraction of viable microorganisms from the interiors of carbonaceous chondrites. These meteorites are porous, however, and "breathe" air in and out during their entry into the atmosphere. There also have been significant opportunities for their contamination after arrival on the Earth. Moreover, one of the organisms extracted was a facultative aerobe. Since, as yet, no planet in the solar system besides the Earth is known to contain significant quantities of molecular oxygen, it seems quite curious that the complex electron transfer apparatus required for oxygen metabolism would be evolved out on the asteroid belt in expectation of ultimate arrival on the Earth. Here, again, contamination has proved a serious hazard. The large amounts of organic matter found in carbonaceous chondrites, however, suggest that the production of organic molecules occurred with very great efficiency in the early history of the solar system.

From such a hypothetical exercise as the instrumental detection of life on Earth by an extraterrestrial observer, and from the actual experience acquired in the analysis of carbonaceous chondrites, the following conclusions can be drawn: There is no single and unambiguous "life detector." There are instruments of great generality that make few ambiguous assumptions about the nature of extraterrestrial organisms, particularly their chemistry. These systems, however, require a fair degree of luck (an animal must walk by during the operating lifetime of the instrument), or they require the solution of difficult instrumental problems (such as the acquisition and preparation of samples for remote microscopic examination). Other instruments, such as metabolism detectors, have great sensitivity and are directed at the more abundant microorganisms. They are quite specific, however, and are critically dependent upon certain assumptions (for example, that extraterrestrial organisms eat sugars) that are no better than informed guesses. Therefore, an array of instruments, both very general and very specific, seems required. Stringent sterilization of such spacecraft appears necessary, both to avoid confusion of the life-detection experiments, and to prevent interaction of contaminants with the indigenous ecology. Many of the instruments and strategies discussed in the preceding paragraphs continue to be adapted by the United States and the Soviet Union in attempts to search for life on the Moon and the nearby planets (see EXPLORATION: *Space exploration*).

Biological prospects of the Moon

An exobiological survey of the solar system. A brief survey of the physical environments and biological prospects of the moons and planets of the solar system, so far as is known, follows. The Moon's surface seems inhospitable to life of any sort. The diurnal temperatures range from about 100 to about 400 K. In the absence of any significant atmosphere or magnetic field, ultraviolet light and charged particles from the Sun penetrate unimpeded to the lunar surface, delivering in less than an hour a dose lethal to the most radiation-resistant microorganism known. For other reasons already mentioned, the absence of an atmosphere and of any liquid medium on the surface also argues against life. The subsurface environment of the Moon is not nearly so inclement. About a metre or so subsurface there is no penetration of ultraviolet light or solar protons, and the temperature is maintained at a relatively constant value about 230 K. Even there, however, the absence of an atmosphere and the probable absence of abundant liquids make the biological prospects rather dim.

It is not out of the question, however, that prebiological organic matter, produced in the early history of the Moon, might be found sequestered beneath the lunar surface. Such organic matter may have been produced either in an original lunar atmosphere that has subsequently been lost to space, or in a secondary lunar atmosphere produced by release of gases after the formation of the Moon, and also subsequently lost to space. The depth at which such organic matter may be found depends upon the unknown history of the early lunar atmosphere, if any, and upon whether the Moon has, on the whole, gained or lost matter due to meteoritic impact. An apparent gaseous emission near the lunar crater Alphonsus was recorded in 1958 and a spectral identification was made of the molecule C_2, a likely organic fragment, but this identification subsequently has been disputed.

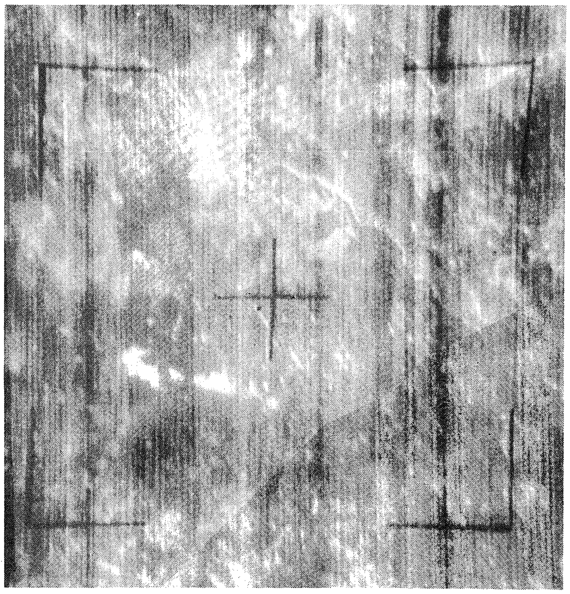
Because of contamination by unmanned spacecraft, the lunar surface had accumulated a microbial load estimated by the late 1960s at some 100,000,000 microorganisms. Since such organisms will be immediately killed unless shielded from radiation, and since the likelihood of their growth seems remote, such contamination may not be a serious problem in subsequent microbial analysis of returned lunar samples. A much more serious contamination problem occurs during the acquisition of such samples by astronauts. Samples obtained during the historic Apollo 11 Moon landing in July 1969 were tested for possible organic molecules, but results were inconclusive. Such a finding might shed significant light on the early history of organic molecules in the solar system.

The environment of Mercury is rather like that of the Moon. Its surface temperatures range from about 100 to about 620 K, but about a metre subsurface the temperature is constant, very roughly at comfortable room temperature on Earth. But the absence of any significant atmosphere, the unlikelihood of bodies of liquid, and the intense solar radiation make life unlikely.

Speculations about life on Mars

Direct evidence for life on Mars has been claimed for many decades. The first such argument was posed by a French astronomer, E.L. Trouvelot, in 1884: "Judging from the changes that I have seen to occur from year to year in these spots, one could believe that these changing grayish areas are due to Martian vegetation undergoing seasonal changes." The seasonal changes on Mars have been reliably observed, not only visually but also photometrically. There is a conspicuous springtime increase in the contrast between the bright and dark areas of Mars. Accompanying colour changes have been reported, but their reality has been disputed. While such changes have been attributed to the growth of vegetation, seasonally variable dust storms are an equally convincing possibility.

The most famous case, historically, for life on Mars is the discovery of the "canals," a set of apparent thin straight lines that cross the Martian bright areas and extend for hundreds and sometimes thousands of kilometres. They change seasonally as do the Martian dark areas. These lines, first systematically observed by an Italian astronomer, G.V. Schiaparelli, in 1877, were further cataloged and popularized by a U.S. astronomer, Percival Lowell, around the turn of the century. Lowell argued from the unerring straightness of the lines that they could not be of geological origin but must instead be the artificial constructs of a race of intelligent Martians. He suggested that they might be channels carrying water from the melting polar caps to the parched equatorial cities of Mars. While considerable skepticism has been expressed about these straight lines, there is no doubt that approximately rectilinear features do exist on the Martian surface. More probable explanations, however, include crater chains, terrain contour boundaries, faults, mountain chains, and ridges analogous to the suboceanic ridge systems that are features of the Earth.



Photograph, from a TIROS weather satellite, of a region near Cochrane, Ontario. The crisscross pattern in white (top left) shows logging swaths, a sign of intelligent life on Earth.

By courtesy of National Aeronautics and Space Administration

In July and August 1976, two U.S. probes bearing equipment designed to detect the presence or remains of organic material made successful landings on Mars. Analyses of atmospheric and soil samples met with procedural difficulties and yielded initially ambiguous and inconclusive results, although the data were later generally interpreted as negative, at least for the vicinity of the probe.

Features of Venus

According to both ground-based and space-borne observations, the average surface temperatures of Venus are around 750 K. It does not seem likely, either at the poles or on the tops of the highest Venus mountains, that the surface temperature will be below 400 K, and noontime temperatures are probably significantly hotter than 700 K. Thus, quite apart from the other surface conditions, the temperatures on Venus seem too hot for terrestrial life. It is still not possible to exclude a Venus surface life with a rather different chemistry, although hydrogen bonding would be much less suitable for the geometrical configuration of polymers on Venus than it is on Earth. The clouds of Venus, however, are another matter. There, carbon dioxide, sunlight, and (according to the results of the Venera space vehicles) water are to be found. These are the prerequisites for photosynthesis. Some molecular nitrogen also is expected at the cloud level, and some supply of minerals can be expected from dust convectively raised from the surface. The cloud pressures are about the same as on the surface of the Earth, and the temperatures in the lower clouds also are quite Earthlike. Despite the fact that there is little oxygen, the lower clouds of Venus are the most Earthlike extraterrestrial environment known. While there are no recorded cases of organisms on Earth that lead a completely airborne existence throughout their life cycle, it is not impossible that such organisms could exist in the vicinity of the Venus clouds, perhaps buoyed, as is a fish by its swim bladder, to avoid downdrafts carrying them to the hotter lower atmosphere.

A similar speculation can be entertained with regard to the lower clouds of Jupiter. On Jupiter the atmosphere is composed of hydrogen, helium, methane, ammonia, and probably neon and water vapour. But these are exactly those gases used in primitive-Earth simulation experiments directed toward the origin of life. Laboratory and computer experiments have been performed on the application of energy to simulated Jovian atmospheres. In addition to the immediate gas-phase products, such as hydrogen cyanide and acetylene, more complex organic molecules, including aromatic hydrocarbons, are formed in lower yield. The visible clouds of Jupiter are vividly coloured, and it is possible that their hue is attributable to such coloured organic compounds. There is also an apparent absorption feature near 2,600 Å, in the ultraviolet spectrum of Jupiter, which has been attributed both to aromatic hydrocarbons and to nucleotide bases. In any event it is likely that organic molecules are being produced in significant yield on Jupiter; it is possible that Jupiter is a vast planetary laboratory that has been operating for 5,000,000,000 years on prebiological organic chemistry.

Satellites and comets of the solar system

The other Jovian planets, Saturn, Uranus, and Neptune, are similar in many respects to Jupiter, although much less is known about them. Their cloud-top temperatures progressively decrease with distance from the Sun. In the case of Saturn, microwave studies have indicated that the atmospheric temperature increases with depth below the clouds; similar situations are expected on Jupiter, Uranus, and Neptune. Thus, it is by no means clear that the low temperatures of the upper clouds of the Jovian planets apply to the lower clouds, or to the underlying atmosphere. The environment of Pluto is almost completely unknown. In addition to these planets, the solar system contains 32 natural satellites, some of which, such as Titan, a satellite of Saturn, and Io, a satellite of Jupiter, appear to have atmospheres. There are also tens of thousands of comets, which, judging from their spectra, contain organic molecules, as well as some thousands of asteroids and asteroidal fragments revolving about the Sun between the orbits of Mars and Jupiter. These are the presumed sources of the carbonaceous chondrites, which contain organic matter.

In short, there is a wide range of environments of biological interest within the solar system. There is no direct evidence for extraterrestrial life on these planets, but, on the other hand, there is no strong evidence against life on many of these worlds. Beyond this is the near certainty that biologically interesting organic molecules will be found throughout the solar system.

Intelligent life beyond the solar system. For thousands of years man has wondered whether he is alone in the universe or whether there might be other worlds populated by creatures more or less like himself. The common view, both in early times and through the Middle Ages, was that the Earth was the only "world" in the universe. Nevertheless, many mythologies populated the sky with divine beings, certainly a kind of extraterrestrial life. Many early philosophers held that life was not unique to the Earth. Metrodorus, an Epicurean philosopher in the 3rd and 4th centuries BC, argued that "to consider the Earth the only populated world in infinite space is as absurd as to assert that in an entire field sown with millet, only one grain will grow." Since the Renaissance there have been several fluctuations in the fashion of belief. In the late 18th century, for example, practically all informed opinion held that each of the planets was populated by more or less intelligent beings; in the early 20th century, by contrast, the prevailing informed opinion (except for the Lowellians) held that the chances for extraterrestrial intelligent life were insignificant. In fact the subject of intelligent extraterrestrial life is for many people a touchstone of their beliefs and desires, some individuals very urgently wanting there to be extraterrestrial intelligence, and others wanting equally fervently for there to be no such life. For this reason it is important to approach the subject in as unbiased a frame of mind as possible. A respectable modern scientific examination of extraterrestrial intelligence is no older than the 1950s. The probability of advanced technical civilizations in our galaxy depends on many controversial issues.

The Green Bank formula

A simple way of approaching the problem, which illuminates the parameters and uncertainties involved, has been devised by a U.S. astrophysicist, F.D. Drake. The number N of extant technical civilizations in the galaxy can be expressed by the following equation (the so-called Green Bank formula):

$$N = R_* f_p n_e f_l f_i f_c L$$

where R_* is the average rate of star formation over the lifetime of the galaxy; f_p is the fraction of stars with planetary systems; n_e is the mean number of planets per star that are ecologically suitable for the origin and evolution of life; f_l is the fraction of such planets on which life in fact arises; f_i is the fraction of such planets on which intelligent life evolves; f_c is the fraction of such planets on which a technical civilization develops; and L is the mean lifetime of a technical civilization. What follows is a brief consideration of the factors involved in choosing numerical values for each of these parameters, and an indication of some currently popular choices. In several cases these estimates are no better than informed guesses and no very great reliability should be pretended for them.

There are about 2×10^{11} stars in the galaxy. The age of the galaxy is about 10^{10} years. A value of $R_* = 10$ stars per year is probably fairly reliable. While most contemporary theories of star formation imply that the origin of planets is a usual accompaniment of the origin of stars, such theories are not well enough developed to merit much confidence. Through the painstaking measurement of slight gravitational perturbations in the proper motions of stars, it has been found that about half of the very nearest stars have dark companions with masses ranging from about the mass of Jupiter to about 30 times the mass of Jupiter. The nearest of these dark companions orbit Barnard's star, which is only six light-years from the sun and is the second nearest star system. The most direct indication that planetary formation is a general process throughout the universe is the existence of satellite systems of the major planets of our own solar system. Jupiter, with 16 satellites, Saturn with 20 or more, and Uranus with five each closely resemble miniature solar systems. It is not known what the distribution of distances of planets from their central star are in other solar systems and whether they tend to vary systematically with the luminosity of the parent star. But considering the wide range of temperatures that seem to be compatible with life, it can be tentatively concluded that $f_p n_e$ is about one.

Likelihood of origin of life

Because of the apparent rapidity of the origin of life on Earth, as implied by the fossil record, and because of the ease with which relevant organic molecules are produced in primitive-Earth simulation experiments, the likelihood of the origin of life over a period of billions of years seems high, and some scientists believe that the appropriate value of f_l is also about one. For the quantities of f_i and f_c the parameters are even more uncertain. The vagaries of the evolutionary path leading to the mammals, and the unlikelihood of such a path ever being repeated has already been mentioned. On the other hand, intelligence need not necessarily be restricted to the same evolutionary path that occurred on the Earth; intelligence clearly has great selective advantage, both for predators and for prey.

Similar arguments can be made for the adaptive value of technical civilizations. Intelligence and technical civilization, however, are clearly not the same thing. For example, dolphins appear to be very intelligent, but the lack of manipulative organs on their bodies has apparently limited their technological advance. Both intelligence and technical civilization have evolved about halfway through the relevant lifetime of the Earth and Sun. Some, but by no means all, evolutionary biologists would conclude that the product $f_i f_c$ taken as 10^{-2} is a fairly conservative estimate.

Still more uncertain is the value of the final parameter, L, the lifetime of a technical civilization. Here, fortunately for man, but unfortunate for the discussion, there is not even one example. Contemporary world events do not provide a very convincing counterargument to the contention that technical civilizations tend, through the use of weapons of mass destruction, to destroy themselves shortly after they come into being. If we define a technical civilization as one capable of interstellar radio communication, our technical civilization is only a few decades old. If then L is about 10 years, multiplication of all of the factors assumed above leads to the conclusion that there is in the second half of the 20th century only about one technical civilization in the galaxy—our own. But if technical civilizations tend to control the use of such weapons and avoid self-annihilation, then the lifetimes of technical civilizations may be very long, comparable to geological or stellar evolutionary time scales; the number of technical civilizations in the galaxy would then be immense. If it is believed that about 1 percent of developing civilizations make peace with themselves in this way, then there are about 1,000,000 technical civilizations extant in the galaxy. If they are randomly distributed in space, the distance from the Earth to the nearest such civilization will be several hundred light-years. These conclusions are, of course, very uncertain.

Communications with extraterrestrial civilizations

How is it possible to enter into communication with another technical civilization? Independent of the value of *L*, the above formulation implies that there is about one technical civilization arising every decade in the galaxy. Accordingly, it will be extraordinarily unlikely for man soon to find a technical civilization as backward as his. From the rate of technical advance that has occurred on the Earth in the last few hundred years, it seems clear that man is in no position to project what future scientific and technical advances will be made even on Earth in the next few hundred years. Very advanced civilizations will have techniques and sciences totally unknown to 20th-century man. Nevertheless man already has a technique capable of communication over large interstellar distances. This technique, already encountered in the discussion of life on Earth, is radio transmission. Imagine that we employ the largest radio telescope available on Earth, the 1,000-foot-diameter dish of Cornell University, the Arecibo Observatory in Puerto Rico, and existing receivers, and that the identical equipment is employed on some transmitting planet. How distant could the transmitting and receiving planets be for intelligible signals to be transmitted and received? The answer is a rather astonishing 1,000 light-years. Within a volume centred on the Earth, with a radius of 1,000 light-years, there are over 10,000,000 stars.

There would of course be problems in establishing such radio communication. The choices of frequency, of target star, of time constant, and of the character of the message would all have to be selected by the transmitting planet so that the receiving planet would, without too much effort, be able to deduce the choices. But none of these problems seem insuperable. It has been suggested that there are certain natural radio frequencies (such as the 1,420-megacycle line of neutral hydrogen) that might be tuned to; the first choice might be to listen to stars of approximately solar spectral type; in the absence of a common language there nevertheless are messages whose intelligent origin and intellectual content could be made very clear without making many anthropocentric assumptions.

Because of the expectation that the Earth is relatively very backward, it does not make very much sense to transmit messages to hypothetical planets of other stars. But it may very well make sense to listen for radio transmissions from planets of other stars. Project Ozma, a very brief program of this sort, oriented to two nearby stars, Epsilon Eridani and Tau Ceti, was organized in 1960 by Drake. On the basis of the Green Bank formula, it would be very unlikely that success would greet an effort aimed at two stars only 12 light-years away, and Project Ozma was unsuccessful. It remains, however, the first pioneering attempt at interstellar communication. Related programs were organized on a larger scale and with great enthusiasm in the 1960s in the U.S.S.R., where a state scientific commission devoted to such an effort was organized. Other communication techniques including laser transmission and interstellar spaceflight have been discussed seriously and may not be infeasible, but if the measure of effectiveness is the amount of information communicated per unit cost, then radio is the method of choice.

The search for extraterrestrial intelligence is an extraordinary pursuit, in part because of the enormous significance of possible success, but in part because of the unity it brings to a wide range of disciplines: studies of the origins of stars, planets, and life; of the evolution of intelligence and of technical civilizations; and of the political problem of avoiding man's self-annihilation. But at least one point is clear. In the words of Loren Eiseley (also from *The Immense Journey*),

> Lights come and go in the night sky. Men, troubled at last by the things they build, may toss in their sleep and dream bad dreams, or lie awake while the meteors whisper greenly overhead. But nowhere in all space or on a thousand worlds will there be men to share our loneliness. There may be wisdom; there may be power; somewhere across space great instruments, handled by strange, manipulative organs, may stare vainly at our floating cloud wrack, their owners yearning as we yearn. Nevertheless, in the nature of life and in principles of evolution we have had our answer. Of men [as are known on earth] elsewhere, and beyond, there will be none forever.

BIBLIOGRAPHY

The definition of life and life on Earth: H.J. MOROWITZ, *Energy Flow in Biology* (1968); A.L. LEHNINGER, *Bioenergetics: The Molecular Basis of Biological Energy Transformations* (1965); P. HANDLER (ed.), *Biology and the Future of Man* (1970); G. and M. BEADLE, *The Language of Life: An Introduction to the Science of Genetics* (1966); V.M. INGRAM, *The Biosynthesis of Macromolecules* (1965); G.N. COHEN, *Biosynthesis of Small Molecules* (1967); R.B. SETLOW and E.C. POLLARD, *Molecular Biophysics* (1962); "Explanation in Biology," *J. Hist. Biol.*, vol. 2, no. 1 (1969); L.J. HENDERSON, *The Fitness of the Environment: An Inquiry into the Biological Significance of the Properties of Matter* (1958); J. KEOSIAN, *The Origin of Life*, 2nd ed. (1968).

Origin of life and extraterrestrial life: I.S. SHKLOVSKII and C. SAGAN, *Intelligent Life in the Universe* (1966); C.S. PITTENDRIGH, W. VISHNIAC, and J.P.T. PEARMAN (eds.), *Biology and the Exploration of Mars* (1966); E.A. SHNEOUR and E.A. OTTESEN (comps.), *Extraterrestrial Life: An Anthology and Bibliography* (1966); S. GLASSTONE, *The Book of Mars* (1968); S.W. FOX (ed.), *The Origins of Prebiological Systems and of Their Molecular Matrices* (1965); A.G.W. CAMERON (ed.), *Interstellar Communication* (1963); W. SULLIVAN, *We Are Not Alone: The Search for Intelligent Life on Other Worlds*, rev. ed. (1966); C. SAGAN, *The Cosmic Connection* (1973).

Evolution of life: G.G. SIMPSON, *The Meaning of Evolution* (1949) and *This View of Life* (1964); L. EISELEY, *The Immense Journey* (1957); G.J. HARDIN, *Nature and Man's Fate* (1959).

(C.Sn./Ed.)